# — Multi-Mosaikbilder —

# Ein Ansatz zur ikonischen Repräsentation von Bilddaten aktiver Kameras

Birgit Möller

**urn:nbn:de:gbv:3-000008784**
[http://nbn-resolving.de/urn/resolver.pl?urn=nbn%3Ade%3Agbv%3A3-000008784]

Dipl.-Inform. Birgit Möller
AG Mustererkennung & Bioinformatik
Institut für Informatik
Martin-Luther-Universität Halle-Wittenberg

Dienstanschrift:

Institut für Informatik
Von-Seckendorff-Platz 1
06099 Halle (Saale)

Dissertation,
der Mathematisch-Naturwissenschaftlich-Technischen Fakultät
der Martin-Luther-Universität Halle-Wittenberg
vorgelegt am 10. Mai 2005,
verteidigt am 8. Juli 2005.

Gutachter:

1. Prof. Dr.-Ing. Stefan Posch,
   Martin-Luther-Universität Halle-Wittenberg
2. Prof. Dr.-Ing. Gerhard Sagerer,
   Universität Bielefeld

# — Multi-Mosaikbilder —

# Ein Ansatz zur ikonischen Repräsentation von Bilddaten aktiver Kameras

**Dissertation**

zur Erlangung des akademischen Grades

**Doktor der Naturwissenschaften (Dr. rer. nat.)**

der Mathematisch-Naturwissenschaftlich-Technischen Fakultät
(mathematisch-naturwissenschaftlicher Bereich)
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

**Birgit Möller**

geb. am 20. September 1976 in Bielefeld

Halle (Saale), im Mai 2005

**urn:nbn:de:gbv:3-000008784**
[http://nbn-resolving.de/urn/resolver.pl?urn=nbn%3Ade%3Agbv%3A3-000008784]

# Danksagung

Die Anfertigung dieser Dissertation wäre ohne die permanente Unterstützung vieler Personen über den zurückliegenden Bearbeitungszeitraum hinweg kaum möglich gewesen. Mein besonderer Dank gilt dabei Prof. Dr. Stefan Posch, der mich von Beginn an engagiert betreut hat. Obgleich sich insbesondere in der ersten Phase der Promotion durch die räumliche Distanz zwischen Bielefeld und Halle eine enge Zusammenarbeit nicht immer einfach realisieren ließ, war er stets erreichbar und fand Zeit für wertvolle Diskussionen und einen fruchtbringenden Gedankenaustausch. Er wusste jedoch nicht nur die Arbeit durch viele Anregungen und neue Fragestellungen voranzubringen, sondern hat es auch verstanden, mir durch seinen Rückhalt die notwendige Sicherheit bei der Bearbeitung des Promotionsthemas zu vermitteln. Darüber hinaus hatte er auch für außerfachliche Probleme stets ein offenes Ohr, das ich zu schätzen gelernt habe.

Mein weiterer Dank gilt Prof. Dr. Gerhard Sagerer, der mir in seiner Arbeitsgruppe an der Universität Bielefeld den Einstieg in die Promotion ermöglichte und sich auch bereit erklärte, das Zweitgutachten für die vorliegende Arbeit zu verfassen. Die angenehme und produktive Arbeitsatmosphäre in der AG Angewandte Informatik war Ansporn und Motivation zugleich. Auf dieser Basis hat sich auch die Zusammenarbeit mit der AG entwickelt, die nach meinem Wechsel nach Halle bestehen blieb. In diesem Zusammenhang möchte ich mich insbesondere bei Axel Haasch, Marcus Kleinehagenbrock und Jannik Fritsch bedanken, die die Integration der Mosaikbilder in die Architektur ihres Roboters BIRON engagiert vorantrieben und damit ein interessantes Anwendungsfeld schufen.

Auch Thomas Plötz verdient besonderen Dank. Trotz der zeitgleichen Bearbeitung und Fertigstellung seiner eigenen Dissertation fand er Zeit und Interesse, diese Arbeit mit konstruktiver Kritik und wertvollen Anregungen zu bereichern. Jenseits der rein fachlichen Diskussionen standen mir er und seine Frau Alexandra Schubert aber auch als sehr gute Freunde zur Seite, die mir oftmals den nötigen Rückhalt gaben und mir gelegentlich halfen, verlorene Motivation und fehlenden Optimismus wiederzufinden.

An der Martin-Luther-Universität in Halle hat die gute Arbeitsatmosphäre innerhalb der Arbeitsgruppe Mustererkennung und Bioinformatik, und auch innerhalb des Instituts für Informatik als Ganzem, ebenfalls einen entscheidenden Teil zum Gelingen der vorliegenden Arbeit beigetragen. Dabei gilt mein Dank insbesondere Denis Williams für die gute und enge Zusammenarbeit bei der Implementierung der Software und die vielen fruchtbaren Diskussionen, sowie für sein Engagement, um der Arbeitsgruppe eine stabile und komfortable Rechnerumgebung bereitzustellen. Darüber hinaus haben in Halle wie in Bielefeld zahlreiche weitere Personen einen mehr oder minder großen Einfluss auf diese Arbeit ausgeübt, sei es als Bürokollegen, die wertvolle Tipps und Tee bereithielten sowie den Unialltag aufzulockern wussten, oder als Korrekturleser, die diese Arbeit mit vielen Anregungen verfeinerten. Insbesondere sind in diesem Zusammenhang Markus Wienecke, Alf Gerisch, Marc Hanheide und Steffen Neumann zu nennen.

Abschließend gebührt meinen Eltern und meiner Schwester Claudia großer Dank. Sie haben mir Zeit meines Studiums und insbesondere auch während der Promotion stets einen starken emotionalen Rückhalt gegeben, der viele Probleme und Sorgen aus dem Weg zu räumen half. Zahlreiche, oftmals schwierige Entscheidungen im Verlauf des Studiums und danach wären ohne diesen Rückhalt sehr viel komplizierter zu treffen gewesen und hätten die Fertigstellung der nun vorliegenden Arbeit bedeutend erschwert.

# Inhaltsverzeichnis

# 1 Einleitung

Die stetigen technischen Fortschritte im Bereich der Informationstechnologie treiben die zunehmende Integration von Computern in die menschliche Alltagswelt voran. Insbesondere das fruchtbare Wechselspiel in der Entwicklung leistungsfähiger elektronischer Bauteile einerseits und effizienter Algorithmen andererseits hat zur Erschließung zahlreicher neuer Anwendungsfelder beigetragen. Der Einsatz von Computern im Alltag bringt dabei in erster Linie eine Abkehr von strikten, exakt definierten Laborbedingungen mit sich. Alltagstaugliche Systeme müssen eine hohe Flexibilität aufweisen, um auch mit unvorhersehbaren Ereignissen, die in Laboren oft faktisch ausgeschlossen sind, angemessen umgehen zu können. Sie werden mit einer größeren Menge an vielfältigeren Daten konfrontiert, so dass der zu analysierende Datenraum oft sehr komplex und hochdimensional ist. Um dennoch Lösungswege finden zu können, sind effiziente und robuste Mechanismen zur Informationsselektion und -analyse notwendig, ohne die ein System in dynamischen Umgebungen nicht eingesetzt werden kann. Darüber hinaus ist auch eine reibungsfreie Einbindung der Systeme in den Lebensbereich des Menschen unabdingbar. Je enger die Verbindung zwischen Mensch und Maschine ist und je näher ihre Aktionsradien beieinander liegen, desto wichtiger ist die Bereitstellung geeigneter Kommunikations- und Interaktionsschnittstellen, die dem Menschen einen intuitiven Umgang mit der Maschine erlauben.

In der Informatik werden in der Entwicklung von flexiblen Computersystemen und benutzerfreundlichen Mensch-Maschine-Schnittstellen verschiedenste Ansätze verfolgt. Nicht selten dient dabei ein natürliches System als Vorbild, in dem die aufgezeigten Probleme einer hohen Flexibilität und vielfältiger Kommunikations- und Interaktionsmöglichkeiten nahezu perfekt gelöst sind: der Mensch. Durch seine multimodale Sensorik, verknüpft mit komplexen Verarbeitungspfaden zur Analyse der aufgenommenen Informationen, ist er optimal ausgestattet, um sich in seiner Umwelt orientieren und damit letztendlich existieren zu können.

Betrachtet man das Verhalten eines Menschen in Alltagssituationen, fällt zunächst die stetige Interaktion zwischen ihm und seiner Umwelt auf. Basierend auf spezifischen Handlungs- und Verhaltensmustern werden fortwährend Informationen ausgetauscht. Die ausgeführten Handlungen lassen sich dabei nicht allein durch passive Reaktionen auf wahrgenommene Reize erklären. Sie entspringen vielmehr einem iterativen Prozess, der detaillierte Analysen akquirierter Daten und auch eine aktive Neuausrichtung der Sensorik, d.h. eine gezielte Informationsselektion, einschließen kann. Letztere wird entscheidend durch Vorwissen und Erwartungen geprägt. Der Mensch muss damit über ein komplexes Verarbeitungssystem verfügen, in dem die Wahrnehmung nur *ein* Bestand-

teil unter vielen ist, und in dessen Rahmen erst das Zusammenspiel verschiedenster Komponenten die Gewinnung schlüssiger Interpretationen als Grundlage für das eigene Verhalten erlaubt.

In [Gol02] wird dieses System durch einen Kreisprozess charakterisiert (Abb. 1.1), der die verschiedenen Komponenten mit ihren wechselseitigen Verknüpfungen veranschaulicht. Die Wahrnehmungskomponente stellt darin gemeinsam mit der Sensumotorik[1] die Schnittstelle zur Außenwelt dar, wobei auch Aufmerksamkeit (vgl. Kap. 7) zur gezielten Fokussierung einzelner Informationsquellen und ein Kurz- bzw. Arbeitsgedächtnis dazugezählt werden. Die Sensumotorik selbst gliedert sich in *ausführende Motorik*, die direkte Manipulationen der Umwelt bewirkt, *kommunikative Motorik*, etwa Mimik und Gestik, und *explorative Motorik*. Hierunter fallen motorische Bewegungen der Sinnesorgane, wie Augen- und Kopfbewegungen, durch die eine aktive Selektion von Informationen aus der Umwelt möglich wird. Mit Wahrnehmung und Sensumotorik verknüpft sind Komponenten, die im Wesentlichen Verhalten, Handeln und Zielbildung, sowie kognitive und reflexive Prozesse umfassen. Reflexion ist dabei der Schlüssel zur Fähigkeit des Menschen, bewusst in die Wahrnehmung einzugreifen. Der Kreisprozess hat keinen festen Einstiegspunkt, sondern kann vielmehr von verschiedenen Komponenten aus initiiert werden. Auch hierin spiegelt sich die enge Verknüpfung der Wahrnehmung mit aus kognitiven Prozessen resultierenden Motiven und Zielen wider.

Die vorliegende Arbeit stellt ein Konzept zur Realisierung eines visuellen „Gedächtnisses“ in technischen Systemen vor, das in das obige Schema als Teil der Wahrnehmungskomponente eingeordnet werden kann, dabei aber auch enge Verbindungen zur Sensumotorik beinhaltet. Daher wird nachfolgend zunächst die Wahrnehmung visueller Informationen und ihre interne Repräsentation beim Menschen genauer betrachtet.

Allgemein erfüllt die Wahrnehmung nach [Gol02] verschiedenste Aufgaben, angefangen bei der Festlegung von Basisbezugssystemen für die grundsätzliche Orientierung, über die Erkennung von Gegenständen, Orten und Ereignissen, bis hin zur Steuerung und Kontrolle der Motorik. Jeder Mensch hat dabei seine individuelle, subjektive Wahrnehmung, die durch seinen Wissens- und Erfahrungsschatz geprägt und eng mit den aktuellen Erwartungen und Handlungsabsichten verknüpft ist. Das visuelle System nimmt dabei einen besonderen Stellenwert ein [Mat92], es ermöglicht dem Menschen das Erkennen von Farben, Bewegungen und Strukturen in der Umwelt. Im Gegensatz zu den anderen Sinnesorganen, die weniger Flexibilität bieten, kann die Informationsaufnahme dabei durch Kombinationen von Kopf- und Augenbewegungen besonders selektiv erfolgen. So ist es möglich, jeweils relevante Bereiche zu fokussieren, aber auch für die aktuelle Zielsetzung nicht benötigte Informationen explizit auszublenden. Letzteres spielt unter dem Aspekt beschränkter Kapazitäten zur weiteren Verarbeitung eine wichtige Rolle.

[1] auch Sensomotorik, Bezeichnung für Nervenprozesse, bei denen sowohl sensorische wie motorische Fasern in Tätigkeit sind, so wie für die Nervenstruktur, die der Träger dieses Prozesses ist; in entsprechendem Sinne auch Bezeichnung für Prozesse, in denen ein unmittelbarer Zusammenhang zwischen Wahrnehmungen und Verhalten besteht, z.B. Hand-Auge-Koordination (aus [Dor04], S. 856ff.)

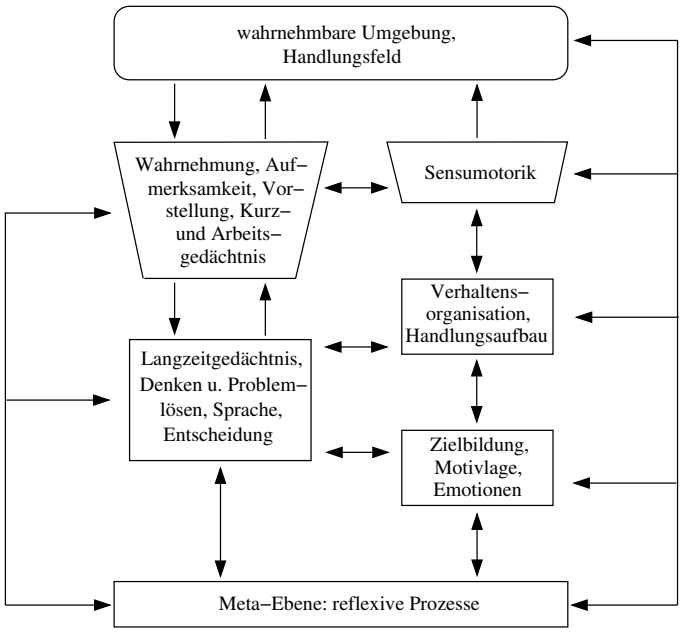



**Abbildung 1.1:** Wahrnehmungskreislauf des Menschen (nach [Gol02]).

In der Psychologie werden im Hinblick auf ein Verständnis der Funktionsweise und Organisation der visuellen Informationsverarbeitung unter anderem Prozesse der Aufmerksamkeitssteuerung und die interne Repräsentation visueller Daten erforscht. Dabei stehen insbesondere Fragen der Struktur und Kapazität interner Speicher im Mittelpunkt (z.B. [Sch95, Sch98a]), wobei zahlreiche konkurrierende Modelle zum Aufbau, sowie zu Art und Umfang der gespeicherten Daten existieren. Einerseits wird eine eher semantische Natur der Daten propagiert [Irw96], während andererseits auch das Vorhandensein präziser räumlicher Informationen als Grundlage zur Motoriksteuerung hypothetisiert wird [Hay02]. Pashler [Pas95] schließlich schlägt zusätzlich eine explizite Unterscheidung zwischen den initial (möglicherweise nur implizit) gespeicherten Daten und den tatsächlich abrufbaren und damit nutzbaren Informationen vor.

Aus der Analyse des menschlichen Wahrnehmungssystems lassen sich für die Entwicklung flexibler, interaktiver technischer Systeme verschiedene wichtige Schlussfolgerungen ziehen. Vorrangig wird deutlich, dass eine passive Sensorik nur in seltenen Fällen eine hinreichende Grundlage zur Lösung komplexer Aufgaben darstellt. Eine aktive Sensorik, optional mit geeigneten Aktoren (etwa Roboterarmen, Greifern, etc.) kombiniert, verspricht größere Spielräume für eine zielgerichtete Informationsakquisition. Sie ermöglicht damit eine mitunter hilfreiche, oftmals aber auch zwingend notwendige Einschränkung des zu durchsuchenden Lösungsraumes. Im Hinblick auf die Verarbeitung visueller Daten hat nicht zuletzt auf Basis dieser Erkenntnisse in den letzten Jahren das Forschungsgebiet der *Active Vision* zunehmend an Bedeutung gewonnen. Nach Swain und Stricker [Swa93] umfasst Active Vision grundsätzlich alle Mechanismen, die eine zielgerichtete, oft auf sich anschließende Handlungen ausgelegte Selektion visueller Daten in Raum, Zeit und Auflösung erlauben. Somit sind darunter einerseits aktive Kameras zu verstehen,

deren intrinsische (Brenn- bzw. Bildweite, Fokus, Blende, etc.) und extrinsische Parameter (Rotationswinkel, Position im Weltkoordinatensystem) modifizierbar sind. Gleichermaßen fallen aber auch algorithmische Ansätze zur gezielten Datenauswahl in diesen Bereich. Active Vision ermöglicht damit eine flexiblere, stärker verhaltens- und zielorientierte Informationsauswahl bei der Lösung gestellter Aufgaben (z.B. [Den02, Roy04]).

Ein zweiter, für technische Systeme wichtiger Aspekt menschlicher Wahrnehmung folgt aus der Betrachtung des Wahrnehmungskreislaufs als Ganzem. Wie zuvor bereits deutlich wurde, sollte eine aktive Sensorik nicht als allein stehendes Konzept angesehen werden. Erst ihre Kopplung mit weiteren Komponenten zur zielgerichteten Analyse und Planung führt letztendlich in Richtung der beim Menschen zu beobachtenden Leistungsfähigkeit. Analog dazu ist in interaktiven Systemen oftmals eine Interpretation von Daten auch auf Basis unabhängigen Welt- oder Modellwissens unerlässlich. Selbst verhältnismäßig einfache Analysen erfordern bereits ein Mindestmaß an intern gespeicherten Referenzdaten früherer Zeitpunkte. Erst sie ermöglichen die Etablierung von Korrespondenzen und bilden somit die Grundlage für eine über die Zeit stabile Wahrnehmung der Umgebung, wie sie auch beim Menschen zu beobachten ist [Irw90]. Schlussendlich ergibt sich daraus die Notwendigkeit, interaktive Systeme sowohl mit einer aktiven Sensorik als auch mit vielfältigen Repräsentationsstrukturen für Zusatzinformationen und dazu passenden Abfrage- und Analysemechanismen auszustatten.

Zur internen Repräsentation *visueller* Daten gibt es in Abhängigkeit vom jeweiligen Anwendungsgebiet verschiedenste Ansätze. Ihre Komplexität reicht von der Speicherung kaum oder gar nicht vorverarbeiteter Bildsignale [Mat96], über merkmalsbasierte Repräsentationen [Ish96, Ras96] bis hin zu weitgehend vom Eingangssignal abstrahierenden Wissensbasen [Sag97]. Der konkrete Abstraktionsgrad in einem spezifischen Anwendungskontext hängt im Wesentlichen von der mit der Datenrepräsentation verbundenen Aufgabenstellung ab. Signalnahe Ansätze zielen vorrangig auf eine kompakte Speicherung der Daten, wobei zunächst keine Analyse der Bildinhalte erfolgt. Sie wird erst zu einem späteren Zeitpunkt unter einer detaillierteren Zielvorgabe nachgeholt. Im Gegensatz dazu geht stärker abstrahierenden Repräsentationsansätzen eine solche Analyse zumeist voraus. Die Signale werden bereits im Hinblick auf den späteren Anwendungskontext interpretiert und in eine dafür geeignete Darstellung überführt (z.B. 3D-Szenenmodelle bei Navigations- und Lokalisationsaufgaben in der Robotik oder eine Extraktion spezifischer Merkmale für eine Objekterkennung). Dies reduziert zumeist den Speicherbedarf und vereinfacht die Behandlung konkreter Fragestellungen, beschränkt aber auch die Flexibilität in der Verwendung der Datenstrukturen bei späteren Modifikationen des ursprünglich implizierten Anwendungsziels.

Die vorliegende Arbeit stellt ein Konzept zur Repräsentation von visuellen Daten in einem *ikonischen Speicher* vor, der auf *Mosaikbildern* basiert. Dabei werden überwiegend wenig bis nicht vorverarbeitete Daten abgelegt, die interaktiven Systemen eine effiziente, weitgehend verlustfreie Speicherung der über einen längeren Zeitraum hinweg akquirierten Bildinformationen ermöglichen. Durch die zusätzliche Bereitstellung einer Schnittstelle zur direkten Anwendung gängiger Bildverarbeitungsansätze lässt sich außerdem

eine hohe Flexibilität im Hinblick auf eine spätere Auswertung der Daten im Rahmen tiefer gehender Analyseschritte erzielen. Mosaikbilder bilden für einen solchen Ansatz insbesondere bei der Verwendung aktiver Kameras eine gute Grundlage. Im Folgenden wird zunächst eine kurze Einführung in die Grundlagen der Mosaikbilder gegeben, gefolgt von einem inhaltlichen Überblick der vorliegenden Arbeit sowie ihrer Einordnung und Zielsetzung.

## 1.1 Mosaikbilder

Der Sichtbereich heutiger Kameras ist in aller Regel beschränkt. In Abhängigkeit von der gewählten Zoomstufe lassen sich jeweils nur mehr oder weniger große Teile einer Szene fokussieren. Großwinklige Aufnahmen sind im Allgemeinen nur mit speziellen Kameras bzw. Linsen möglich. Insbesondere im Bereich der Computergrafik und ihrer Anwendung in der *Virtual Reality*, aber auch im Umfeld der *Computer Vision* sind in den letzten Jahren jedoch zunehmend Anwendungsfelder entstanden, die eine künstliche Erweiterung des Sichtfeldes einer Kamera interessant erscheinen lassen.

Die Zielsetzung im Forschungsfeld der Virtual Reality besteht in der Generierung möglichst realitätsnaher, künstlicher Umgebungen, die der Mensch beispielsweise immersiv mit so genannten „Head Mounted Displays“ oder auch unter Verwendung von 3D-Brillen und Datenhandschuhen erkunden kann [Vin95]. Um in diesen Welten eine ausreichende Detailtreue und damit einhergehend einen hohen Realitätsgrad erreichen zu können, wird eine große Rechenleistung zum Rendern der notwendigen Ansichten aus 3D-Modellen benötigt. Dieser Aufwand lässt sich reduzieren, wenn darzustellende Objekte nicht mehr aus Modelldaten gerendert, sondern stattdessen durch die Projektion von realen, objektspezifischen Texturen auf relativ einfache Oberflächen generiert werden (*Image Based Rendering*). Dabei motiviert das vorrangige Ziel, in den künstlichen Welten möglichst natürliche Bewegungen zuzulassen und dem Besucher keine unnötigen Beschränkungen des Sichtfeldes aufzuerlegen, die Verwendung von Bilddaten, die diesen Einschränkungen ebenfalls nicht unterliegen.

In der Computer Vision, deren Ziel nach [Bal82] in der Konstruktion expliziter, bedeutsamer Beschreibungen von physikalischen Objekten aus Bildern liegt, ist unter anderem mit dem Forschungsfeld der Active Vision die Notwendigkeit erwachsen, Bildfolgen effizient verarbeiten zu können. Dazu sind einerseits Verfahren zur ressourcenschonenden Speicherung der Folgen erforderlich, sowie andererseits Techniken, die es ermöglichen, die Daten einzelner Bilder zueinander in Beziehung zu setzen und damit das Sichtfeld der Kamera in Raum und Zeit zu erweitern.

Mosaikbilder stellen hierfür einen Lösungsansatz dar. Die Grundidee basiert auf der Annahme, dass in den Bildern einer Folge einzelne Szenenbereiche zumeist wiederholt auftreten und damit redundante Information in der Bildsequenz enthalten ist. Unter Ausnutzung und gleichzeitiger Eliminierung dieser Redundanz lässt sich die Bildfolge zu einem, das Sichtfeld der gesamten Folge umfassenden Mosaikbild zusammensetzen (Abb. 1.2). Mosaik-Techniken finden, außer zur effizienten Repräsen-

**Abbildung 1.2:** Ein exemplarisches Mosaikbild, berechnet aus 46 Bildern, die mit einer von Hand geführten, handelsüblichen Digitalkamera aufgenommen wurden.

tation von Bildfolgen und als Grundlage bei der Generierung von virtuellen Welten [Che95, Tel98, Shu99, Lao00] auch Anwendung im Bereich der Videokompression [Ira95, Ira96], sowie in der Hintergrundstabilisierung und Bewegungsdetektion in Bildfolgen bewegter Kameras [Bha00, Mit00]. Die (implizite) Detektion und Verknüpfung der mehrfach auftretenden Szenenteile eröffnet des Weiteren Möglichkeiten, aus mehreren gering aufgelösten Bildern ein hochauflösendes Bild zu rekonstruieren (*Super-resolution*, [Zom00, Cap04]). Im Bereich der Robotik werden Mosaikbilder bislang zumeist als Grundlage visueller Umgebungskarten in der Lokalisation und Navigation eingesetzt. Dabei wird etwa der Standort eines Roboters relativ zu seiner Umgebung durch einen Vergleich aktueller Bilddaten mit einer zuvor erstellten, visuellen Referenzrepräsentation rekonstruiert [Tsu93, Ish94].

Ein Mosaikbild lässt sich aus einer Folge von Einzelbildern mit Hilfe eines zweistufigen Verfahrens konstruieren. Die grundsätzliche Vorgehensweise mit den beiden Hauptphasen der *Registrierung* und *Integration* ist schematisch in Abbildung 1.3 dargestellt. Im ersten Schritt der Registrierung werden die Koordinatensysteme der Einzelbilder mit Hilfe eines mathematischen Modells zueinander und zu einem Referenzkoordinatensystem in Beziehung gesetzt. Die Wahl des Referenzsystems hängt einerseits von den Freiheitsgraden der eingesetzten aktiven Kamera und andererseits von der verfolgten Zielsetzung ab. Im einfachsten Fall dient das lokale, euklidische Koordinatensystem eines Einzelbildes der Folge als Grundlage für das Referenzkoordinatensystem (wie z.B. bei dem Mosaikbild in Abb. 1.2). Bei stationären Kameras, die um ihr optisches Zentrum rotieren, werden alternativ häufig zylindrische [Bis95] oder sphärische Koordinaten [Coo00] gewählt. Grundsätzlich kann das zur Bildregistierung verwendete Koordinatensystem auch von dem des späteren Mosaikbildes verschieden sein, in der vorliegenden Arbeit wird zwischen beiden Systemen jedoch nicht weiter unterschieden bzw. bei Bedarf an den entsprechenden Stellen explizit darauf verwiesen.

Die Registrierung der einzelnen Bilder erfolgt auf Basis eines parametrisierten Bewegungsmodells für die angenommene Kamerabewegung. Das Modell beschreibt mathematisch die zwischen den Einzelbildern der Folge beobachteten, durch die Kamera induzierten Veränderungen und bildet die Grundlage für eine Transformation der Bilder

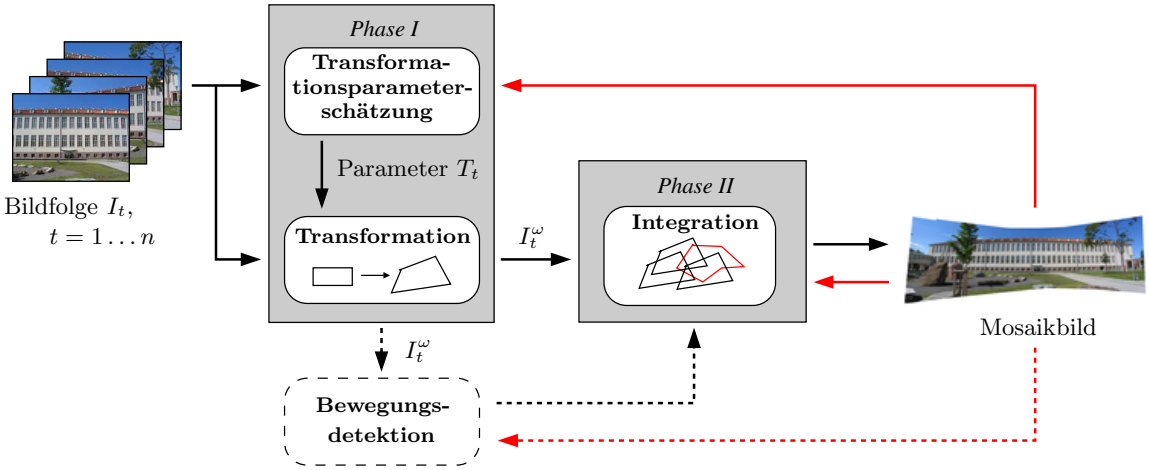



**Abbildung 1.3:** Die Konstruktion von Mosaikbildern durchläuft die beiden Phasen der Registrierung (I) und Integration (II). In der ersten Phase werden Transformationsparameter $T_t$ für die einzelnen Bilder einer Folge $I_t, t = 1 \ldots n$, geschätzt, mit denen diese in das gemeinsame Koordinatensystem transformiert werden können. Anschließend folgt in der zweiten Phase die Fusion der transformierten Bilder $I_t^\omega$ zu einem Mosaikbild. Im Gegensatz zu einer Offline-Verarbeitung von Bildfolgen, bei denen alle Bilder gleichzeitig registriert und integriert werden, erfolgt im Online-Modus (durch die roten Elemente in der Grafik angedeutet) eine schrittweise Registrierung der einzelnen Bilder und eine sequenzielle Integration in das stetig wachsende Mosaikbild. Zur Berücksichtigung unabhängiger Bewegungen ist weiterhin vielfach eine Komponente zur Bewegungsdetektion in die Verfahren eingebettet.

in das gemeinsame Referenzkoordinatensystem. Die notwendige Komplexität des Modells leitet sich dabei aus den Freiheitsgraden der Kamera sowie der Struktur der Szene ab. Nach der Transformation der Bilder mit Hilfe des Bewegungsmodells erfolgt in der zweiten Phase ihre Integration ins Mosaikbild. Der Farbwert eines einzelnen Pixels im Mosaikbild ergibt sich dabei aus den Farbinformationen der korrespondierenden Pixel in den Einzelbildern, die durch die Transformation auf das Mosaikpixel abgebildet werden.

Durch die Umrechnung der Bilder in das gemeinsame Koordinatensystem werden Veränderungen zwischen einzelnen Bildern der Folge kompensiert, die aus der Kamerabewegung resultieren. Weitere Unterschiede, die vorrangig in Bildfolgen dynamischer Szenen auftreten und z.B. durch unabhängig bewegte Objekte hervorgerufen werden, folgen dem globalen Bewegungsmodell nicht und bleiben somit in der Regel auch nach der Registrierung erhalten. Da hierdurch einerseits Integrationsfehler im Mosaikbild verursacht werden können, andererseits aber auch interessante Hinweise für eine Interpretation der visuellen Daten gegeben sind, beinhalten viele Algorithmen zur Mosaikbildgenerierung optional Verfahren zur Detektion unabhängig bewegter Szenenteile.

Für die Generierung von Mosaikbildern aus einer Bildfolge gibt es grundsätzlich zwei verschiedene Herangehensweisen: *online* oder *offline*[2]. Beide Ansätze arbeiten in dem zuvor skizzierten 2-Phasen-Schema, sie unterscheiden sich jedoch in den dabei zu Grunde gelegten Bildinformationen. Offline-Algorithmen nutzen zur Parameterschätzung alle Bilder der Folge gleichzeitig und versuchen damit global optimale Parameter zu ermitteln, mit denen die Bilder simultan in das Mosaikbild integriert werden können. Sie

[2]In der Literatur werden Offline-Ansätze häufig auch als *globale* Techniken bezeichnet.

durchlaufen infolgedessen beide Phasen der Registrierung und Integration in der Regel nur einmal, wobei sie jeweils ein Vorhandensein der kompletten Bildfolge voraussetzen.

Im Gegensatz dazu arbeiten Online-Verfahren inkrementell (vgl. auch die roten Pfeile und Elemente in Abb. 1.3). Für jedes Bild der Folge werden dabei separat Parameter ermittelt, mit denen es anschließend direkt in das sich sukzessive aufbauende Mosaikbild integriert wird. Somit steht zu jedem Zeitpunkt im Prozess der Generierung – und nicht erst nach Aufnahme und Verarbeitung der kompletten Folge – ein aktuelles Mosaikbild zur Verfügung, das auch als zusätzliche Informationsquelle in der Parameterschätzung und bei der Bewegungsdetektion genutzt werden kann. Die geschätzten Parameter sind in diesem Fall aufgrund der Vorgehensweise höchstens optimal bezüglich der bis zum aktuellen Zeitpunkt verarbeiteten Bilddaten und des gerade vorliegenden Mosaikbildes. Aus diesem Grund sind die mit Online-Verfahren generierten Mosaikbilder im Allgemeinen von geringerer Qualität als bei Offline-Ansätzen. Letztere empfehlen sich daher insbesondere zur Erstellung hochqualitativer Bilder, z.B. als Grundlage für virtuelle Welten. Für den Einsatz in interaktiven Systemen, wo stetig neue Bilddaten akquiriert werden, sind demgegenüber Online-Verfahren notwendig, die jederzeit eine ikonische Repräsentation der bis dato aufgenommenen Bilddaten bereitstellen können.

## 1.2 Zielsetzung und Gliederung der Arbeit

Der fortwährende Einzug des Computers in viele Bereiche des alltäglichen Lebens motiviert das Bestreben in der Informatik, Systeme sowohl interaktiv zu gestalten als sie auch mit einer hohen Flexibilität auszustatten. Auf diese Weise wird eine schnelle Anpassung an unbekannte Situationen und Umgebungen möglich. In Anlehnung an das zu Beginn dieser Einleitung skizzierte, komplexe Interaktions- und Kommunikationssystem des Menschen werden viele Systeme dazu inzwischen einerseits mit einer aktiven, multimodalen Sensorik, und andererseits mit vielfältigen internen Repräsentationsstrukturen und effizienten Zugriffsmechanismen für ergänzendes Wissen ausgestattet. Vor diesem Hintergrund präsentiert die vorliegende Dissertation ein neues Konzept zur effizienten Repräsentation visueller, ikonischer Daten auf Basis von Mosaikbildern. Das Ziel besteht dabei weniger in der exakten Umsetzung psychologischer Wahrnehmungs- und Gedächtnismodelle. Vielmehr soll interaktiven Systemen eine ergänzende, interne Repräsentationsstruktur zur Verfügung gestellt werden, die die Verknüpfung visueller Daten über die Zeit unterstützt. Damit wird eine stabile Abbildung der wahrgenommenen Umgebung möglich, die einen Beitrag zur Nachbildung der Funktionalität menschlicher Wahrnehmungs- und Handlungsprozesse in interaktiven Systemen bilden kann.

Aus technischer Sicht soll der vorgestellte ikonische Speicher im Wesentlichen zwei Anforderungen erfüllen. Einerseits wird angestrebt, die Interaktivität der Systeme so wenig wie möglich einzuschränken und den Systemen einen flexiblen Umgang mit den akquirierten Bilddaten zu ermöglichen. Dies impliziert, dass zu jedem Zeitpunkt Zugriff auf die im Mosaikbild gespeicherten Daten gewährt werden soll, so dass eine Online-Verarbeitung der Daten unverzichtbar ist. Eine solche Vorgehensweise als notwendige Voraussetzung

zum Einsatz des visuellen Speichers in interaktiven Systemen wird auch dadurch bekräftigt, dass insbesondere mobile, interaktive Systeme oftmals nur beschränkte Kapazitäten zur Verarbeitung und Speicherung großer Datenmengen zur Verfügung stellen. Damit ist eine Offline-Berechnung von Mosaikbildern auf Basis vollständiger Bildsequenzen zumeist schon allein aufgrund der technischen Ausstattung der Systeme nicht durchführbar, ein Aspekt, der bislang nur in wenigen Arbeiten zur Berechnung von Mosaikbildern explizit thematisiert wird (vgl. auch Unterkap. 3.3).

Als zweite wichtige Anforderung soll die Speicherstruktur darüber hinaus eine größtmögliche Flexibilität im Hinblick auf potenzielle Anwendungsfelder gewährleisten und keine vorzeitige Beschränkung ihres Einsatzbereiches auf einige wenige ausgewählte Zielapplikationen bedingen. Dies bedeutet insbesondere, dass die direkte Anwendung gängiger Verfahren der Bildanalyse unterstützt werden soll, so dass sich die Extraktion zusätzlicher Informationen aus der Szene allein auf Basis des Speichers und ohne erneuten Hardwarezugriff realisieren lässt.

Bei einer Repräsentation von visuellen Daten in Mosaikbildern spielt die Wahl des Referenzkoordinatensystems in Abhängigkeit von der Kamerabewegung, der angestrebten Verwendung der Mosaikbilder und des gewählten Verarbeitungsmodus eine entscheidende Rolle. In der vorliegenden Arbeit wird eine stationäre Kamera zu Grunde gelegt, die Rotationen um das optische Zentrum durchführen und zoomen kann. Diese Wahl gründet einerseits auf der exakten, mathematischen Modellierbarkeit der Bewegungen und Abbildungseigenschaften einer derartigen Kamera, wie sie etwa bei translatorischen Kamerabewegungen nicht mehr gegeben ist [Pel97, Pel00]. Darüber hinaus bietet eine stationäre, rotierende Kamera aber auch für zahlreiche Anwendungsgebiete bereits eine hinreichende Flexibilität bei der Akquisition visueller Daten, so dass eine Berücksichtigung zusätzlicher Freiheitsgrade insbesondere hinsichtlich des damit verbundenen Mehraufwandes im Allgemeinen nicht notwendig erscheint (vgl. auch Kap. 8).

Zur Repräsentation von Mosaikbildern, die mit einer stationären, rotierenden Kamera aufgenommen wurden, stellen sphärische Koordinatensysteme, d.h. Kugeloberflächen, im Hinblick auf eine Minimierung von Verzerrungen die optimale Projektionsbasis dar (s. auch Kap. 6). Allerdings erfordert die Zielsetzung, eine Anwendung konventioneller Bildverarbeitungsalgorithmen direkt auf den gespeicherten ikonischen Daten zu unterstützen, euklidische Koordinaten, die auf einer Kugeloberfläche nicht gegeben sind. Darüber hinaus ist auch eine Online-Integration neuer Bilddaten auf Basis sphärischer Koordinatensysteme oftmals schwierig, so dass in dieser Arbeit alternativ das neue Konzept der *Multi-Mosaikbilder* vorgeschlagen wird. Das Referenzkoordinatensystem dieser Bilder ist durch eine Menge von Ebenen definiert, die gleichmäßig um das optische Zentrum der Kamera angeordnet werden. Ihre globale Anordnung leitet sich dabei von Polyedern ab, die je nach Anzahl vorhandener Teilflächen eine Kugeloberfläche mehr oder minder exakt approximieren und damit den vollständigen Sichtbereich einer stationären, rotierenden Kamera erfassen. Die stückweise Planarität des Referenzkoordinatensystems trägt der geforderten Unterstützung konventioneller Bildverarbeitungsalgorithmen Rechnung. Um darüber hinaus auch Bilddaten in verschiedenen Auflösungen möglichst verlustfrei und

effizient speichern zu können, wird das visuelle Gedächtnis in einer *Auflösungshierarchie* organisiert. Ein Multi-Mosaikbild besteht somit aus mehreren, ineinander geschachtelten Mengen polyedrisch angeordneter Teilflächen, die je nach aktueller Bildweite als Projektionsziele ausgewählt werden.

Beim Menschen kann die interne Repräsentation aufgenommener Sensordaten und ihre vorherige, aktive Selektion und Akquisition durch Prozesse der Aufmerksamkeitssteuerung nicht getrennt voneinander betrachtet werden. Damit wird auch in interaktiven technischen Systemen eine Kopplung von aktiver Sensorik und zielgerichteter Datenakquisition auf der einen Seite, und der internen Repräsentation auf der anderen Seite nahe gelegt. Mosaikbilder können diese Verknüpfung, d.h. insbesondere die Auswahl neuer Aufmerksamkeitspunkte aus zuvor aufgenommenen Daten, durch ihr in Raum und Zeit erweitertes Sichtfeld optimal unterstützen. Um dies zu illustrieren, wird im Rahmen dieser Arbeit ergänzend ein auf dem Konzept der Multi-Mosaiks aufbauender, datengetriebener Ansatz zur *aktiven Szenenexploration* vorgestellt, der die Verknüpfung zwischen Datenrepräsentation und -akquisition in interaktiven Systemen komplettiert.

Die Gliederung der Arbeit orientiert sich zunächst an dem zuvor skizzierten Verfahren zur Konstruktion von Mosaikbildern. Nach einer einführenden Beschreibung des zu Grunde liegenden Kameramodells in Kapitel 2, verbunden mit der Herleitung des Bewegungsmodells, werden in den Kapiteln 3 und 4 Algorithmen zur Parameterschätzung und Bildtransformation, sowie zur Integration vorgestellt. Dabei stehen Methoden im Vordergrund, die eine robuste Berechnung des Mosaiks im Online-Modus ermöglichen. Nur so ist der Einsatz des Speichers in interaktiven Systemen realisierbar, die eine effiziente, dynamische Aktualisierung der gespeicherten Daten erfordern. Da die Repräsentation weiterhin nicht auf statische Szenen beschränkt bleiben soll, enthält Kapitel 5 Ansätze zur Detektion und Repräsentation unabhängiger Bewegungen in den Bildfolgen.

Basierend auf den in diesen ersten Abschnitten vorgestellten Methoden wird in Kapitel 6 schließlich das Konzept der hierarchischen Multi-Mosaikbilder als Grundlage einer auch große Winkel und verschiedene Auflösungsstufen umfassenden, ikonischen Szenenrepräsentation eingeführt und diskutiert. Die enge Verknüpfung von aktiver Sensorik und interner Repräsentation bildet den Ausgangspunkt für Kapitel 7. Darin wird ein prototypischer Ansatz zur aktiven Szenenexploration auf Basis hierarchischer Multi-Mosaikbilder vorgestellt. Er demonstriert die Verwendung der Mosaiks als Grundlage zur Ansteuerung aktiver, visueller Sensorik in interaktiven Systemen.

Durch die signalnahe Repräsentation der Daten eröffnen sich zahlreiche Perspektiven für einen praktischen Einsatz des entwickelten Konzepts der Multi-Mosaikbilder. Kapitel 8 zeigt dies anhand eines konkreten Szenarios auf, das in Zusammenarbeit mit Kollegen der Arbeitsgruppe „Angewandte Informatik“ der Universität Bielefeld entworfen wurde. Der Kern der Kooperation besteht in der Integration des visuellen Speichers in die Architektur des *Bielefeld Robot Companion (BIRON)*, einem multimodal interagierenden, mobilen Roboter. Das Kapitel beschreibt den grundlegenden Ansatz und erste Resultate der bisherigen Zusammenarbeit. Die Dissertation schließt mit einer zusammenfassenden Diskussion und einem Ausblick auf verbleibende Fragestellungen in Kapitel 9.

# 2 Das Kameramodell

Ein grundlegender Schritt bei der Generierung eines Mosaikbildes besteht in der Registrierung der Einzelbilder einer gegebenen Folge. Dabei werden die zwischen den einzelnen Bildern beobachteten Veränderungen mit Hilfe eines geeigneten Bewegungsmodells $T_{\vec{p}}$ mathematisch beschrieben. Insbesondere wird in Abhängigkeit von den Daten ein passender Parametersatz $\vec{p}$ für das Modell geschätzt. Die Bilder können dann durch Anwendung von $T_{\vec{p}}$ in das gemeinsame Referenzkoordinatensystem transformiert und durch Fusion ihrer Bilddaten zum Mosaikbild zusammengefügt werden. Das Bewegungsmodell selbst leitet sich bei einer bewegten Kamera in der Regel direkt aus der Kamerabewegung ab, wobei seine Komplexität im Wesentlichen mit den Freiheitsgraden der eingesetzten Kamera skaliert. Hieraus resultiert die Notwendigkeit, der Wahl eines geeigneten Bewegungsmodells zunächst eine Modellierung der Kamera voranzustellen.

In diesem Kapitel wird das der Arbeit zu Grunde liegende Modell einer *endlichen, projektiven Kamera* skizziert, aus dem sich direkt das gewählte Bewegungsmodell einer *projektiven Transformation* motiviert. Dabei reichen zur Schätzung von Parametern für dieses Modell grundsätzlich allein die Bilddaten aus, weitere Informationen, d.h. insbesondere Kalibrierungsdaten der Kamera, sind zumeist nicht notwendig. Allerdings erfordert eine Repräsentation der visuellen Daten auf einem polyedrischen Grundkörper, wie sie in der Einleitung als Kernidee der Multi-Mosaikbilder beschrieben wurde, zur korrekten Festlegung der Skalierung des Körpers zumindest die Kenntnis der Bildweite. Daher werden im Unterkapitel 2.4 auch Verfahren zur Online- und Offline-Kalibrierung von endlichen, projektiven Kameras diskutiert. Die Darstellung in diesem Kapitel orientiert sich im Wesentlichen an [Har00], wobei zur Formulierung mathematischer Sachverhalte unter anderem homogene Koordinaten aus der projektiven Geometrie verwendet werden. Aus diesem Grund findet sich im nachfolgenden Abschnitt zunächst eine kurze Einführung in die projektive Geometrie, einschließlich der verwendeten Notation. Weiterführende Details sind in [Moh96] oder [Har00] zu finden.

## 2.1 Grundzüge der projektiven Geometrie

Die projektive Geometrie hat nach [Edg02] ihren Ursprung in der Malerei der italienischen Renaissance. Dort erwuchs das Bestreben, Eindrücke aus der 3D-Welt, insbesondere bezogen auf die Architektur, korrekt in Gemälden widerzuspiegeln und in diesem Zusammenhang mathematisch zu formalisieren. Die gängige Geometrie der euklidischen Räume $\mathbb{R}^n$ stieß dabei an ihre Grenzen. Insbesondere projektive Effekte im Unendlichen, wie beispielsweise der Schnitt zweier parallel verlaufender Geraden im imaginären

Fluchtpunkt unter projektiver Abbildung auf eine Ebene, erforderten neue Formalismen, die in der Einführung projektiver Räume $\mathbb{P}^n$ gipfelten. Diese stellen im Wesentlichen eine Erweiterung der euklidischen Räume $\mathbb{R}^n$ um *ideale* Punkte, Geraden und Ebenen dar, die im Unendlichen liegen und damit den Ausgangspunkt zur mathematischen Beschreibung entsprechender Effekte bilden.

Die Grundlage projektiver Räume ist durch *homogene Koordinaten* gegeben, deren Haupteigenschaft die Invarianz gegenüber Skalierungen ist. Punkte im $\mathbb{P}^n$ werden durch $(n+1)$-dimensionale Vektoren beschrieben, wobei Nullvektoren $\vec{0}$ ausgeschlossen sind. Die Skalierungsinvarianz lässt sich durch eine Äquivalenzrelation ausdrücken, die nachfolgend exemplarisch für Punkte $u$ des $\mathbb{P}^2$ angegeben ist:

$$u = (x, y, z)^T \sim \lambda\,(x, y, z)^T \quad , \quad \lambda \neq 0 \ , \ \lambda \in \mathbb{R}.$$

Ein aus dieser Eigenschaft leicht abzuleitendes Modell für den $\mathbb{P}^2$, das zum besseren Verständnis projektiver Räume hier vorgestellt werden soll, beschreibt die projektiven Punkte des Raumes als Strahlen, die durch den Ursprung eines 3D-Koordinatensystems verlaufen (Abb. 2.1). Alle auf einem Strahl liegenden 3D-Punkte sind durch Multiplikation mit einem geeigneten Skalierungsfaktor ineinander überführbar und somit in diesem Modell äquivalent. Werden die Strahlen des $\mathbb{P}^2$ mit einer beliebigen Ebene des Raumes geschnitten, so lassen sich die resultierenden Schnittpunkte als eindeutige Repräsentanten der Strahlen auffassen. Dabei existieren für parallel zur jeweils gewählten Ebene verlaufende Strahlen keine Repräsentanten. Im Sonderfall der parallel zur xy-Ebene gelegenen Ebene $E : z = 1$ ergeben sich alle Repräsentanten aus den geschnittenen Strahlen durch Skalierung der dritten Vektorkomponenten auf 1. Vernachlässigt man darüber hinaus diese dritten Komponenten der Repräsentanten vollständig, resultiert gerade die Menge der zweidimensionalen Vektoren des $\mathbb{R}^2$. Die Ebene $E : z = 1$ lässt sich somit als Einbettung des $\mathbb{R}^2$ in den $\mathbb{P}^2$ auffassen, wobei der $\mathbb{P}^2$ gegenüber dem $\mathbb{R}^2$ noch um die Punkte ergänzt wird, deren Strahlen die Ebene nicht schneiden. Letztere verlaufen parallel zu $E$ in der xy-Ebene des 3D-Koordinatensystems und weisen in ihrer dritten Komponente eine Null auf. Sie liegen im Prinzip im Unendlichen und werden als *ideale* Punkte bezeichnet. Formal lässt sich zu jedem nicht-idealen Punkt $u \in \mathbb{P}^2$ ein eindeutiger 2D-Repräsentant $\tilde{u} = (x, y)^T \in \mathbb{R}^2$ durch folgende Funktion bestimmen:

$$f_{PR} : u = (x, y, z)^T \rightarrow \tilde{u} = (x/z, y/z)^T \text{ mit } (x, y, z)^T \sim (x/z, y/z, 1)^T. \tag{2.1}$$

Umgekehrt kann für jeden Punkt des $\mathbb{R}^2$ ein Vertreter der repräsentierten Äquivalenzklasse im $\mathbb{P}^2$ durch Ergänzung einer 1 in der dritten Vektorkomponente ermittelt werden:

$$f_{RP} : \tilde{u} = (x, y)^T \rightarrow u = (x, y, 1)^T. \tag{2.2}$$

Im Rahmen dieser Arbeit werden zur Bezeichnung von Punkten des $\mathbb{R}^2$ bzw. $\mathbb{P}^2$ Kleinbuchstaben verwendet, und für Punkte aus $\mathbb{R}^3$ und $\mathbb{P}^3$ Großbuchstaben. Punkte aus den euklidischen Räumen werden zur Unterscheidung zusätzlich mit einer Tilde versehen.

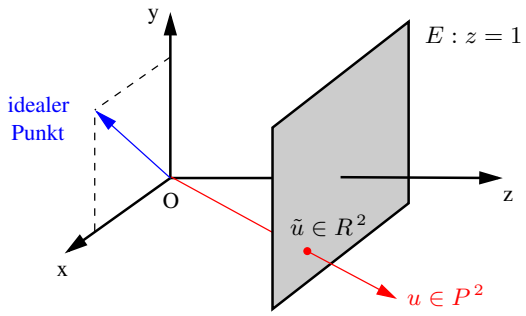



**Abbildung 2.1:** Visualisierung des Strahlenmodells des $\mathbb{P}^2$, wobei die eingezeichnete Ebene $E : z = 1$ die Einbettung des $\mathbb{R}^2$ in den $\mathbb{P}^2$ veranschaulicht.

Neben dem mathematischen Formalismus zur Beschreibung projektiver Effekte bringen homogene Koordinaten einen weiteren Vorteil auch im Hinblick auf andere Anwendungsfelder mit sich. Viele mathematische Konzepte, insbesondere Abbildungen, lassen sich unter Verwendung von homogenen Koordinaten durch leicht zu handhabende Matrix-Vektor-Operationen realisieren. Das nachfolgende Beispiel veranschaulicht dies. Eine affine Transformation $A$ im $\mathbb{R}^2$ setzt sich aus einer Rotationsmatrix $R$ und einem Translationsvektor $\vec{t}$ zusammen:

$$A(\tilde{u}) = R \cdot \tilde{u} + \vec{t} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \cdot \tilde{u} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}.$$

In homogenen Koordinaten lässt sich die obige Gleichung durch eine einfache Matrixmultiplikation ausdrücken:

$$A(u) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) & t_x \\ \sin(\alpha) & \cos(\alpha) & t_y \\ 0 & 0 & 1 \end{bmatrix} \cdot u = \begin{bmatrix} R & \vec{t} \\ \vec{0}^T & 1 \end{bmatrix} \cdot u.$$

Die Darstellung der Matrix im hinteren Teil der Gleichung, die Teilmengen der Matrixkomponenten in *Blöcken* zusammenfasst, wird als *Block-Schreibweise* bezeichnet und bietet sich an, wenn die einzelnen Komponenten der Matrix nicht direkt von Interesse sind. Dabei können die einzelnen Blöcke zur besseren Lesbarkeit durch senkrechte Striche voneinander abgegrenzt werden (vgl. etwa Gl. 2.3).

## 2.2 Endliche, projektive Kameras

Der im Rahmen dieser Arbeit entwickelte, visuelle Speicher dient der Repräsentation von Bilddaten, die mit stationären, jedoch um ihr optisches Zentrum rotierenden und zoomenden Kameras aufgenommen werden. Für den Aufbau der Kameras wird dabei das *Lochkameramodell* zu Grunde gelegt (Abb. 2.2). Das Projektionszentrum der Kamera $O_{cam}$ (*optisches Zentrum*) liegt im Ursprung eines 3D-Kamerakoordinatensystems, die optische Achse fällt mit der z-Achse des Koordinatensystems zusammen. Im Abstand der

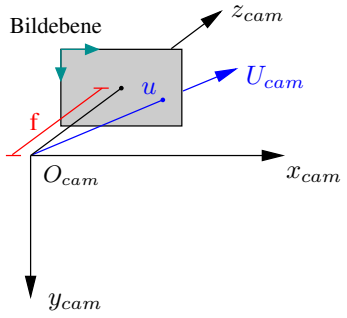



**Abbildung 2.2:** Das Lochkameramodell.

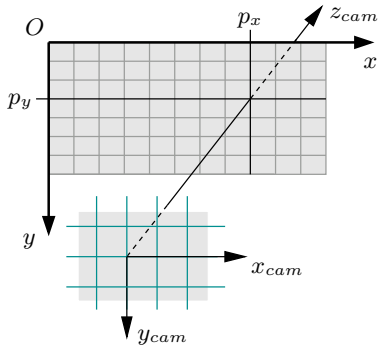



**Abbildung 2.3:** Bildkoordinatensystem.

*Bildweite (focal length)* $f > 0$ vom optischen Zentrum befindet sich die Bildebene, auf die die betrachtete 3D-Welt zweidimensional abgebildet wird. Die Bildebene ist dabei stets parallel zur xy-Ebene des 3D-Koordinatensystems ausgerichtet. Sie besitzt ein eigenes lokales 2D-Koordinatensystem (Abb. 2.3), dessen Ursprung in der Regel in die linke obere Ecke der Ebene gelegt wird (*upper left*).

Die Position eines Punktes auf dieser Ebene wird in homogenen Koordinaten angegeben, die sich aus der Anwendung der Funktion $f_{RP}$ aus Gleichung 2.2 auf die lokalen Bildkoordinaten ergeben. Der *Hauptpunkt* der Kamera, an dem die optische Achse die Bildebene durchstößt, liegt entsprechend an Position $(p_x, p_y, 1)^T$ auf der Ebene.

Die Projektion $u$ eines beliebigen 3D-Punktes $U_{cam}$ des Kamerakoordinatensystems auf die Bildebene ergibt sich in diesem Modell als Schnittpunkt des Richtungsstrahls vom optischen Kamerazentrum zum Punkt $U_{cam}$ mit der Bildebene. Er kann unter Anwendung des Strahlensatzes in homogenen Koordinaten wie folgt berechnet werden:

$$u \;=\; K\,[\,I\,|\,\vec{0}\,]\cdot U_{cam} \quad \text{mit} \quad K = \begin{bmatrix} f & & p_x \\ & f & p_y \\ & & 1 \end{bmatrix}, \; I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.3}$$

Dabei ist noch unberücksichtigt geblieben, dass die beiden Koordinatenachsen der Bildebene in der Regel jeweils unterschiedlich und darüber hinaus unabhängig vom 3D-Koordinatensystem der Kamera skaliert sind. Auf diese Weise wird der Tatsache Rechnung getragen, dass die einzelnen Zellen der in den Kameras üblicherweise zur Bildaufnahme eingesetzten CCD-Arrays[1] nicht immer quadratisch sind. Um die Bildkoordinaten in Pixeleinheiten korrekt errechnen zu können, müssen somit bei der Abbildung in x- und y-Richtung jeweils spezifische Skalierungsfaktoren berücksichtigt werden. Wenn $m_x$ und $m_y$ die Anzahlen an CCD-Elementen pro Einheitslänge im Kamerakoordinatenssystem in beiden Achsenrichtungen angeben, so sind die 3D-Koordinaten bei der Abbildung mit den Faktoren $\alpha_x = f \cdot m_x$ und $\alpha_y = f \cdot m_y$ zu multiplizieren, gegeben die Bildweite $f$ in Pixeln. Ergänzend ist auch der Hauptpunkt umzurechnen.

[1] CCD-Arrays sind hier und im weiteren Verlauf der Arbeit als Synonym für pixelbasierte Aufnahmetechniken zu verstehen, zu denen beispielsweise auch CMOS-Verfahren gehören.

Zusätzlich wird zur Quantifizierung von Abweichungen der Achsen des Bildkoordinatensystems von der üblicherweise angenommenen Orthogonalität noch ein weiterer Parameter $s$ (*Skew*) eingeführt. Für normale Kameras und Abbildungen wird er aber im Allgemeinen zu Null angenommen. Die vollständige *Kalibrierungsmatrix* $K$ aus Gleichung 2.3 ergibt sich folglich unter Berücksichtigung dieser Aspekte zu

$$K = \begin{bmatrix} \alpha_x & s & x_0 \\ & \alpha_y & y_0 \\ & & 1 \end{bmatrix} \quad \text{mit} \quad x_0 = p_x \cdot m_x \;\;,\;\; y_0 = p_y \cdot m_y. \tag{2.4}$$

Insgesamt werden somit bislang zur Beschreibung des Kameramodells und damit auch der projektiven Abbildung die fünf *intrinsischen* Parameter $x_0, y_0, \alpha_x, \alpha_y$ und $s$ benötigt. Die Kamera ist jedoch in ein 3D-Weltkoordinatensystem eingebettet, das ihre globale Position in der Szene festlegt. In diesem Fall lassen sich 3D-Punkte in der Welt nur dann direkt unter Anwendung von Gleichung 2.3 auf die Bildebene projizieren, wenn Kamera- und Weltkoordinatensystem übereinstimmen. Andernfalls muss die Transformation zwischen den beiden Systemen explizit in die Abbildung einbezogen werden.

Zwei gleichartig orientierte[2] 3D-Koordinatensysteme lassen sich durch Rotation und Translation ineinander überführen. Für einen Punkt $U$ im 3D-Weltkoordinatensystem bedeutet dies, dass er vor der eigentlichen Abbildung in 3D-Kamerakoordinaten $U_{cam}$ umgerechnet werden muss:

$$U_{cam} = \begin{bmatrix} R & -R\tilde{C} \\ \vec{0}^T & 1 \end{bmatrix} \cdot U.$$

In der vorstehenden Formel bezeichnet $R$ die $3 \times 3$-dimensionale Rotationsmatrix zwischen den beiden Koordinatensystemen, die durch drei Drehwinkel um die einzelnen Koordinatenachsen parametrisiert ist. Der 3D-Vektor $\tilde{C}$ gibt das Zentrum der Kamera in euklidischen Weltkoordinaten an und kodiert damit die durchzuführende Translation. $R$ und $\tilde{C}$ werden unter dem Oberbegriff der *extrinsischen* Kameraparameter zusammengefasst. Für das Gesamtmodell der Kamera ergeben sich unter Berücksichtigung der intrinsischen und extrinsischen Parameter insgesamt elf Freiheitsgrade, die in der *Kamera-* oder *Abbildungsmatrix* $P$ zusammengefasst werden:

$$P \;=\; K\,[\,I\,|\,\vec{0}\,] \cdot \begin{bmatrix} R & -R\tilde{C} \\ \vec{0}^T & 1 \end{bmatrix} \;=\; K\,R\,[\,I\,|\,-\tilde{C}\,] \;=\; [\,M\,|\,-M\tilde{C}\,]. \tag{2.5}$$

Ein solches Kameramodell wird als *endliche, projektive Kamera* bezeichnet, wobei die Nicht-Singularität der Matrix $M = KR$ entscheidend ist. Die Abbildung eines Punktes $U$ auf die Bildebene ist abschließend durch folgende Formel gegeben:

$$u \;=\; K\,[\,I\,|\,\vec{0}\,] \cdot U_{cam} \;=\; P \cdot U \;=\; K\,R\,[\,I\,|\,-\tilde{C}\,] \cdot U. \tag{2.6}$$

[2] 3D-Koordinatensysteme können entweder rechts- oder linkshändig orientiert sein.

## 2.3 Projektive Transformationen

Bei der Generierung eines Mosaikbildes müssen einzelne Bilder einer Folge zueinander registriert werden. Dies bedeutet insbesondere, dass Zusammenhänge zwischen den projektiven Abbildungen einer Szene auf verschiedene Bildebenen herzustellen sind. Während dies im allgemeinen Fall beliebiger Kameras und Konstellationen sehr komplex sein kann, vereinfacht sich die Situation, wenn endliche, projektive Kameras zu Grunde gelegt werden, die ortsgebunden sind und lediglich um ihre optischen Zentren rotieren (Abb. 2.4). In diesem Fall kann über die Gleichungen 2.5 und 2.6 direkt ein Zusammenhang zwischen den Abbildungsmatrizen hergeleitet werden. Sei $P = KR[\,I\,|\,-\tilde{C}\,]$ die Kameramatrix der ersten Kamera, und $P' = K'R'[\,I\,|\,-\tilde{C}\,]$ die der zweiten. Dann resultiert aus einer Gleichsetzung der identischen Teilmatrizen $[\,I\,|\,-\tilde{C}\,]$ direkt der folgende Zusammenhang:

$$P \;= (KR) \cdot (K'R')^{-1} \cdot P'.$$

Für die Abbildungen $u$ und $u'$ eines Punktes $U$ aus der Szene auf die zwei verschiedenen Bildebenen gilt damit:

$$u \;=\; P \cdot U \;=\; (KR) \cdot (K'R')^{-1} \cdot P' \cdot U \;=\; (KR) \cdot (K'R')^{-1} \cdot u'. \tag{2.7}$$

Basierend auf der bei endlichen, projektiven Kameras gegebenen Nicht-Singularität der Matrizen $KR$ und $K'R'$ lässt sich das Matrixprodukt aus der vorstehenden Gleichung 2.7 schließlich durch eine einzelne, nicht-singuläre $3 \times 3$-Matrix $H$ ersetzen, die eine Abbildung vom $\mathbb{P}^2$ in den $\mathbb{P}^2$ realisiert:

$$u \;=\; (KR) \cdot (K'R')^{-1} \cdot u' = H \cdot u' \quad , \quad H : \mathbb{P}^2 \rightarrow \mathbb{P}^2. \tag{2.8}$$

Nicht-singuläre $3 \times 3$ -Matrizen mit dieser Eigenschaft werden allgemein als *projektive Transformationen* oder *Homographien* bezeichnet. Sie sind aufgrund der Skalierungsinvarianz homogener Koordinaten ebenfalls skalierungsinvariant, d.h. sie lassen sich durch acht Parameter exakt spezifizieren. In der praktischen Anwendung von Homographien werden die neun Einträge der Homographiematrix daher in der Regel konsistent skaliert, wobei beispielsweise der Eintrag rechts unten in der Matrix auf 1 normiert werden kann. Bei der Abbildung von Punktemengen aus dem $\mathbb{P}^2$ mit Homographien bleiben gerade Linien grundsätzlich als solche erhalten, Längenverhältnisse und Winkel können sich jedoch verändern. Die Menge der Homographien, die der Menge aller reellen, regulären Matrizen entspricht, bildet bezüglich der Matrixmultiplikation eine Gruppe. Daraus folgt, dass sich jede projektive Transformation invertieren lässt. Außerdem wird dadurch auch die Konkatenation '$\circ$' von Homographien eindeutig definiert, wobei sie auf die Multiplikation der Homographiematrizen zurückgeführt wird.

Aus den vorstehenden Ausführungen folgt, dass sich Bilder, die mit einer ortsfesten, rotierenden Kamera bei variierenden intrinsischen Parametern aufgenommen wurden, durch die Berechnung einer Homographie zueinander registrieren lassen. Dieses Modell bildet damit die Grundlage zur Bildregistrierung in der vorliegenden Arbeit. Dabei sei

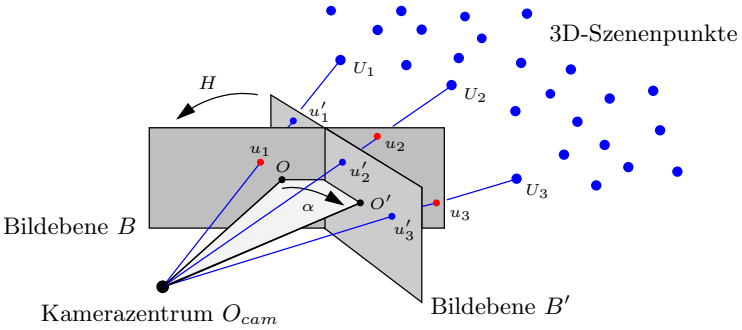



**Abbildung 2.4:** Modell einer stationären, rotierenden Kamera: Zwischen den Abbildungen von 3D-Punkten auf die beiden Bildebenen $B$ und $B'$ besteht ein projektiver Zusammenhang $H$.

insbesondere darauf hingewiesen, dass zur reinen Verknüpfung der Bilder weder intrinsische noch extrinsische Kameraparameter bekannt sein müssen. Diese werden erst für eine metrische Rekonstruktion der betrachteten Szene und im Kontext dieser Arbeit vorrangig zur Auswahl geeigneter Projektionskörper für die Multi-Mosaikbilder benötigt. Aus diesem Grund werden im nachfolgenden Abschnitt Methoden zur Bestimmung der Kameraparameter *(Kalibrierung)* vorgestellt und diskutiert.

## 2.4 Kamerakalibrierung

Das Ziel einer *Kalibrierung* besteht in der Rekonstruktion der extrinsischen und intrinsischen Kameraparameter, mit denen vorliegende Bilder aufgenommen wurden. Projektive Zusammenhänge zwischen Bildern im Rahmen einer Berechnung von Mosaikbildern lassen sich auch ohne Kenntnis der konkreten Parameterwerte schätzen. Für eine Extraktion von Tiefen- oder 3D-Informationen einer betrachteten Szene sind diese Daten jedoch oftmals unerlässlich. Obgleich die vorliegende Arbeit nicht auf eine solche 3D-Szenenrekonstruktion zielt, kann aber auch hier nicht auf eine Kalibrierung der Kameras verzichtet werden. Dies resultiert aus der Tatsache, dass eine möglichst verzerrungsfreie Projektion von Bilddaten auf Polyeder, wie sie das in der Einleitung skizzierte Konzept des visuellen Speichers vorsieht, eine korrekte Skalierung der polyedrischen Grundkörper voraussetzt, und diese direkt von der Bildweite $f$ der Kameras abhängt.

In der Literatur existieren verschiedene Ansätze zur Bestimmung der intrinsischen und extrinsischen Kameraparameter $K$, $R$ und $\tilde{C}$. Im Allgemeinen erfolgt dabei keine explizite Rekonstruktion der Bildweite $f$, sondern die Skalierungsfaktoren $\alpha_x$ und $\alpha_y$ aus Gleichung 2.4 werden direkt als achsenspezifische Bildweiten interpretiert. $f$ lässt sich zwar mit Hilfe der konkreten Werte für $m_x$ bzw. $m_y$, die der technischen Spezifikation einer Kamera entnommen werden können, auch exakt berechnen, $\alpha_x$ und $\alpha_y$ spezifizieren die gesuchte Abbildung jedoch zumeist bereits hinreichend eindeutig, so dass üblicherweise auf eine explizite Bestimmung von $f$ verzichtet wird.

Die Grundidee vieler Verfahren besteht darin, die Parameter aus Paaren von 3D-Szenenpunkten mit bekannten Weltkoordinaten und ihren 2D-Abbildungen in aufgenomme-

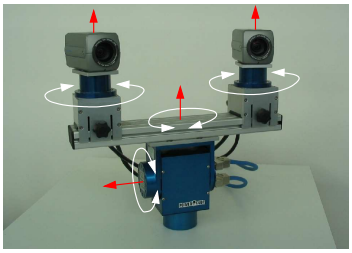


**Abbildung 2.5:** Eingesetzte Kameratypen der Firma Sony$^{TM}$: links die zwei Kameras des Typs „DFW-VL500“, montiert auf dem Stereokamerakopf des Typs „PowerCube“ der Firma Amtec$^{TM}$, rechts die Kamera des Typs „EVI-D31“. Die Pfeile kennzeichnen jeweils verfügbare Rotationsachsen.

nen Bildern zu rekonstruieren. Die Festlegung des 3D-Weltkoordinatensystems erfolgt dabei zumeist anhand eines geeigneten Kalibriermusters (z.B. Abb. 2.6(a)), so dass diese Art der Kalibrierung *offline* durchgeführt werden muss. Als Alternative, die auch *online* angewandt werden kann, haben sich in der Vergangenheit *Autokalibrierverfahren* etabliert. Insbesondere im Rahmen von Homographieschätzungen zwischen Bildern wurden verschiedene Algorithmen publiziert, die eine Rekonstruktion der intrinsischen Parameter $K$ erlauben. In den beiden folgenden Abschnitten werden ausgewählte Ansätze zur Offline- und Online-Kalibrierung vorgestellt, während Abschnitt 2.4.1 Ergebnisse einer Kalibrierung der in dieser Arbeit eingesetzten Kameras (Abb. 2.5) enthält.

## Offline-Kalibrierung

Im Allgemeinen stellt eine Menge von Paaren dreidimensionaler Szenenpunkte und ihrer korrespondierenden zweidimensionalen Abbildungen in aufgenommenen Bildern den Ausgangspunkt einer Offline-Kalibrierung zur Schätzung der intrinsischen und extrinsischen Parameter einer Kamera dar. Das mithin gängigste Verfahren auf dieser Basis zur direkten Ermittlung der intrinsischen und extrinsischen Parameter, sowie zusätzlich von Korrekturfaktoren für eine eventuell vorliegende Linsenverzerrung, wurde bereits 1986 von Tsai veröffentlicht [Tsa86]. Der Algorithmus arbeitet in zwei Schritten, wobei zunächst für möglichst viele Parameter initiale Schätzungen über einen effizienten, linearen Optimierungsansatz ermittelt werden. Anschließend erfolgt im Rahmen einer iterativen, nichtlinearen Optimierung eine Verfeinerung dieser initialen Werte, wobei auch Nebenbedingungen für die einzelnen Parameter in die Berechnungen einfließen.

Eine alternative Methode, die von Hartley [Har00] publiziert wurde, basiert im Gegensatz dazu auf einer expliziten Berechnung der Kameramatrix $P$ (Gl. 2.5). Sie lässt sich bestimmen, indem die gegebenen 2D/3D-Punktkorrespondenzen in die Abbildungsgleichung projektiver Kameras (Gl. 2.6) eingesetzt werden. Die resultierenden Bedingungsgleichungen können zu einem (in der Regel überbestimmten) linearen Gleichungssystem zusammengefasst werden, das beispielsweise über eine Singulärwertzerlegung lösbar ist. Dabei sind für eine vollständige Schätzung aller elf Parameter mindestens sechs Punktkorrespondenzen bei zwei Bedingungsgleichungen pro Korrespondenz erforderlich. Aus der auf diese Weise ermittelten Matrix $P$ können abschließend die gesuchten Vektoren

und Matrizen $\tilde{C}$, $R$ und $K$ hergeleitet werden ([Har00], S. 150ff.). $\tilde{C}$ resultiert aus einer Singulärwertzerlegung von $P$, während $R$ und $K$ mit Hilfe einer $RQ$-Zerlegung aus $M$ abgeleitet werden können (s. Gl. 2.5). Die $RQ$-Zerlegung von $M$ errechnet sich dabei, indem $M^T$ zunächst mit einer Hilfsmatrix $T$ multipliziert wird. Das Ergebnis dieser Operation kann dann mit einer herkömmlichen $QR$-Zerlegung ([Pre92], S. 98ff.) in zwei Matrizen $\bar{Q}$ und $\bar{R}$ aufgespalten werden, aus denen sich schließlich die gesuchten Matrizen $R$ und $Q$ der $RQ$-Zerlegung von $M$ bestimmen lassen:

$$M^T \cdot T = M^T \cdot \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \bar{Q}\bar{R} \ , \ \text{so dass} \quad R = T\bar{R}^T T \quad \text{und} \quad Q = T\bar{Q}^T.$$

Die Qualität der aus einer Offline-Kalibrierung resultierenden Kameraparameter hängt, unabhängig von dem eingesetzten Kalibrierverfahren, direkt mit der Genauigkeit der 3D-Welt- und der 2D-Bildkoordinaten korrespondierender Punkte zusammen. Folglich sollte ein möglichst exaktes Muster zur Kalibrierung verwendet werden, das eine einfache und vor allem auch exakte Lokalisation ausgewählter Punkte in den Bildern bis auf Subpixelgenauigkeit unterstützt. Da die Qualität der Kalibrierung weiterhin mit der Anzahl zu Grunde gelegter Punktepaare zunimmt, ist eine automatische Punktdetektion wünschenswert, um eine mit einem sehr hohen Aufwand verbundene und zudem relativ ungenaue manuelle Auswahl von Punkten vermeiden zu können.

## Online-Autokalibrierung

Eine Kamerakalibrierung mit Hilfe von Kalibriermustern wird im Allgemeinen vor Beginn des eigentlich implizierten Einsatzzieles der Kameras durchgeführt. Dabei sind oftmals für verschiedene Einstellungen einer Kamera voneinander unabhängige Kalibrierungen erforderlich. Eine solche Vorgehensweise ist somit aufwändig, beschränkt die Flexibilität beim Einsatz von Kameras in Anwendungsfeldern, in denen eine Kalibrierung unumgänglich ist, und erzwingt die Bereitstellung eines möglichst exakten Kalibriermusters. *Autokalibrierverfahren* umgehen diese Beschränkungen und erlauben eine automatische, schritthaltende Kamerakalibrierung, die beispielsweise parallel zu einer Online-Generierung von Mosaikbildern erfolgen kann. Die Grundidee der Verfahren beruht auf einer Rekonstruktion der Parameter aus Korrespondenzen zwischen *verschiedenen* Bildern, wobei *keine* bekannten 3D-Weltkoordinaten mehr vorausgesetzt werden. In Abhängigkeit von den bei der Bildaufnahme ausgeführten Kamerabewegungen lassen sich dabei verschiedene Teilmengen der Parameter ermitteln [Fau92, Har99, Pol99]. Im Kontext der vorliegenden Arbeit sind vorrangig die intrinsischen Parameter in $K$ von Interesse, die unmittelbar auf Basis geschätzter Homographien bestimmt werden können.

Shum und Szeliski propagieren dazu in [Shu00] eine direkte Berechnung der Bildweite $f$ aus den Einträgen einer Homographiematrix (vgl. S. 31):

$$f^2 = \frac{h_{23}^2 - h_{13}^2}{h_{11}^2 + h_{12}^2 - h_{21}^2 - h_{22}^2} \quad \text{oder alternativ} \quad f^2 = -\frac{h_{13} \cdot h_{23}}{h_{11} \cdot h_{21} + h_{12} \cdot h_{22}}.$$

Dabei werden quadratische CCD-Elemente ($\alpha_x = \alpha_y$) und eine Positionierung des Hauptpunktes im Bildzentrum (d.h. $x_0 = y_0 = 0$, für ein Bildkoordinatensystem, dessen Ursprung im Hauptpunkt liegt) vorausgesetzt.

Ein zweites Verfahren, das eine Kombination verschiedener Bedingungen für die zu schätzenden Kameraparameter erlaubt, stammt von de Agapito und Kollegen [dA99]. Der lineare Ansatz beruht auf den Eigenschaften des *absolute Conic* $\Omega_\infty$ des $\mathbb{P}^3$, dessen Punktemenge durch die folgende Bedingungsgleichung definiert wird ([Har00], S. 63ff.):

$$\Omega_\infty : \; x_1^2 + x_2^2 + x_3^2 = 0 \; \wedge \; x_4 = 0.$$

Es ist direkt ersichtlich, dass der Conic ausschließlich aus idealen Punkten besteht. Seine Abbildung ${}^i\omega$ durch eine endliche, projektive Kamera $P_i$ hängt allein von deren intrinsischen Parametern $K_i$ ab (unter der Bedingung $s = 0$):

$${}^i\omega = K_i^{-T} \cdot K_i^{-1} = \begin{bmatrix} 1/\alpha_x^2 & 0 & -x_0/\alpha_x^2 \\ 0 & 1/\alpha_y^2 & -y_0/\alpha_y^2 \\ -x_0/\alpha_x^2 & -y_0/\alpha_y^2 & 1 + x_0/\alpha_x^2 + y_0/\alpha_y^2 \end{bmatrix} = \begin{bmatrix} {}^i\omega_{11} & {}^i\omega_{12} & {}^i\omega_{13} \\ {}^i\omega_{12} & {}^i\omega_{22} & {}^i\omega_{23} \\ {}^i\omega_{13} & {}^i\omega_{23} & {}^i\omega_{33} \end{bmatrix}.$$

Mit Hilfe dieser Definition und Gleichung 2.7 lässt sich ein Zusammenhang zwischen den Abbildungen ${}^i\omega$ und ${}^j\omega$ des Conic auf zwei verschiedene Bildebenen $B_i$ und $B_j$ und der zwischen den Ebenen gegebenen, projektiven Abbildung $H_{ij}$ herstellen:

$$H_{ij} = K_j R_j R_i^{-1} K_i^{-1} \iff H_{ij} = K_j R_{ij} K_i^{-1} \iff R_{ij} = K_j^{-1} H_{ij} K_i.$$

Da $R_{ij}$ eine Rotationsmatrix ist, gilt $R_{ij} R_{ij}^T = I$. Setzt man die Definitionen aus den beiden vorstehenden Gleichungen in diese Bedingung ein, folgt daraus durch Umformung:

$$H_{ij}^{-T} K_i^{-T} K_i^{-1} H_{ij}^{-1} = K_j^{-T} K_j^{-1} \iff H_{ij}^{-T} \, {}^i\omega \, H_{ij}^{-1} = {}^j\omega.$$

Für die einzelnen Komponenten der Abbildung ${}^j\omega$ von $\Omega_\infty$ auf die Bildebene $B_j$ können nun Bedingungen formuliert werden, die sich aus den Annahmen über die Kameraparameter ableiten. Aus der Voraussetzung eines verschwindenden Skews folgt direkt ${}^j\omega_{12} = 0$. Liegen darüber hinaus quadratische CCD-Elemente vor, gilt $\alpha_x = \alpha_y$ und damit ${}^j\omega_{11} = {}^j\omega_{22}$. Schlussendlich führt ein im Mittelpunkt der Bilder gelegener Hauptpunkt mit $x_0 = y_0 = 0$ auf die beiden weiteren Bedingungen ${}^j\omega_{13} = 0$ und ${}^j\omega_{23} = 0$. Alle Bedingungsgleichungen können zu einem linearen Gleichungssystem in den unbekannten Einträgen aus ${}^i\omega$ kombiniert werden, dessen Lösung schließlich auch die Bestimmung der intrinsischen Parameter $K_i$ erlaubt. In Abhängigkeit von den zu Grunde gelegten Bedingungen sind dabei zwei, drei oder fünf Homographien erforderlich, um die notwendige Mindestanzahl von fünf Bedingungsgleichungen zur Berechnung der fünf gesuchten Parameter aus $K_i$ formulieren zu können.

Grundsätzlich ist bei diesem Verfahren zu berücksichtigen, dass nicht alle Kamerabewegungen einer rotierenden und zoomenden Kamera eine stabile Kalibrierung erlauben und die Qualität der Ergebnisse darüber hinaus eng mit den verwendeten Annahmen verzahnt ist (siehe z.B. [Stu97, dA00, Wan04] bzw. den nachfolgenden Abschnitt).

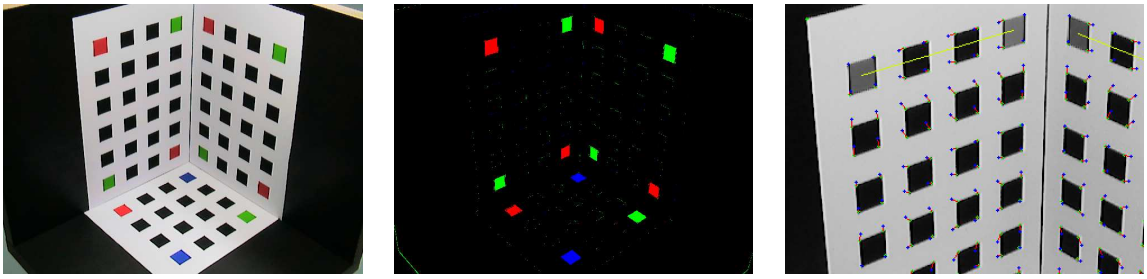
### 2.4.1 Ergebnisse & Auswertung

Im Rahmen dieser Arbeit kommen zwei Kameratypen der Firma Sony$^{TM}$ zum Einsatz (Abb. 2.5). Während die Kamera des Typs „EVI-D31“ direkt angesteuert und in horizontaler ($\approx \pm 100°$) und vertikaler ($\approx \pm 25°$) Richtung um das optische Zentrum rotiert und gezoomt werden kann, erlauben die beiden Kameras des Typs „DFW-VL500“ lediglich eine aktive Änderung des Zooms und keine direkte Rotation. Sie lassen sich jedoch auf einem ansteuerbaren Stereo-Kamerakopf der Firma Amtec$^{TM}$ montieren, womit eine Bewegung der Kameras ermöglicht wird. Der Kopf weist dabei insgesamt vier Dreh- bzw. Kippgelenke auf, von denen allerdings nur die oberen beiden eine Drehung der Kameras um die optischen Zentren gewährleisten (s. Abb. 2.5). Somit sind lediglich horizontale Rotationen durchführbar, die jedoch einen Drehwinkel von 360° überschreiten können. Die Kamera „EVI-D31“ verfügt zudem über einen Autofokusmechanismus, während sich die beiden anderen Kameras ausschließlich manuell fokussieren lassen.

Beide Kameratypen wurden zunächst unter Anwendung des von Hartley publizierten Verfahrens offline kalibriert, um einen funktionalen Zusammenhang zwischen den Hardwareeinstellungen der Kameras und den jeweils korrespondierenden, intrinsischen Parametern herzustellen. Mit Hilfe dieser Funktionen lassen sich anschließend im Verlauf der Mosaikbildberechnung die jeweils aktuellen Hardware-Zoomeinstellungen direkt auf die korrespondierenden Bildweiten abbilden. Dem System stehen somit bei dieser Vorgehensweise insbesondere auch über längere Zeiträume hinweg konsistente, in ihrer Genauigkeit nicht variierende Werte für die Bildweite zur Verfügung.

Der Zoomparameter der Kamera des Typs „EVI-D31“ weist einen Einstellungsbereich von 0 bis 1023 auf, während die Parameter der beiden anderen Kameras Einstellungen zwischen 40 und 1400 zulassen. Im Rahmen der Kalibrierung wurden für beide Kameratypen jeweils Bildfolgen des fokussierten Kalibriermusters aufgenommen, wobei eine schrittweise Erhöhung der Zoomparameter erfolgte. Dabei ist mit jeder der kalibrierten Zoomeinstellungen implizit ein spezifischer Fokusparameter verknüpft, so dass mögliche Auswirkungen einer Fokusänderung auf die Bildweite (bei konstantem Zoomparameter) zunächst unberücksichtigt bleiben.

Die Detektion der Punktreferenzen für die Kalibrierung erfolgt automatisch. Das verwendete Kalibriermuster (Abb. 2.6(a)) zeigt 52 schwarze und 12 farbkodierte Quadrate mit einer Kantenlänge und einem paarweisen Abstand von jeweils $2\,cm$, wobei die geschätzte Genauigkeit in den 3D-Positionen bei etwa $\pm 0,5\,mm$ liegt. Die insgesamt 256 Eckpunkte lassen sich mit einem Harris-Corner-Detektor [Har88] subpixelgenau detektieren. Ihre anschließende, eindeutige Zuordnung zu den durch das Kalibriermuster vorgegebenen 3D-Punktkoordinaten erfolgt anhand von äquidistant eingeteilten Rastern, die mit Hilfe der farbigen Quadrate automatisch generiert werden und die erwarteten 2D-Positionen der Eckpunkte aller Quadrate auf den einzelnen Ebenen vorgeben. Dazu findet zunächst eine Lokalisation der Farbquadrate in den Bildern durch eine bildspezifische Schwellwertbildung in den jeweiligen Farbkanälen statt (Abb. 2.6(b)), gefolgt von einer Regionensegmentierung. Die Anordnung der Quadrate auf dem Kalibriermuster ermöglicht dabei eine eindeutige Zuordnung der detektierten Regionen zu den einzelnen



(a) Originalbild des farbigen Kalibriermusters.

(b) Ergebnis der kanalweisen Farbsegmentierung.

(c) Detektionsresultat der Eckpunkte (Ausschnitt).

**Abbildung 2.6:** Offline-Kalibrierung: verwendetes Kalibriermuster (links) und Ergebnisse der Farbsegmentierung (mittig) und Eckpunkt-Detektion für die Quadrate des Musters (rechts).

Ebenen. Anschließend wird zu jeder Farbregion diejenige Harris-Ecke gesucht, die das Kalibriermuster der jeweiligen Ebene dort begrenzt und damit das Raster definiert. Dies geschieht, indem zunächst die Schwerpunkte aller Farbregionen berechnet und die jeweils vier nächstgelegenen Harris-Ecken als potenzielle Kandidaten ausgewählt werden. Durch eine senkrechte Projektion dieser vier Punkte auf eine Hilfslinie, die aus einer direkten Verbindung benachbarter Regionenschwerpunkte resultiert (in Abb. 2.6(c) gelb dargestellt), wird schließlich die gesuchte Ecke identifiziert. Sie liegt auf der vom Mittelpunkt der Ebene abgewandten Seite der Linie und ihre Projektion besitzt den größten Abstand zu deren Mittelpunkt. Auf Basis des durch diese Punkte definierten Rasters kann dann zu jedem 3D-Punkt des Kalibriermusters die korrespondierende 2D-Harris-Ecke im Bild gesucht werden. Ein vorgegebener Maximalabstand zu den durch das Raster implizierten Eckenpositionen soll dabei das Risiko fehlerhafter Zuordnungen vermindern.

Bei der Offline-Kalibrierung hat sich zunächst gezeigt, dass bei kleinen bis mittleren Zoomparametern der Kameras im Regelfall alle 256 Eckpunkte vom Harris-Corner-Detektor robust detektiert werden. Mit steigenden Werten für die Parameter, die mit großen Bildweiten korrelieren, hat sich die Kalibrierung jedoch als zunehmend schwieriger herausgestellt. Insbesondere eine abnehmende Bildqualität ist dabei als Ursache für schlechtere Resultate der Eckendetektion anzuführen. Als Konsequenz wurden nicht die vollständigen Parameterbereiche beider Kameras kalibriert, sondern nur Teilintervalle zwischen 0 und 900 bei der „EVI-D31“- und zwischen 50 und 1200 bei den „DFW-VL500“-Kameras. Die ausgeklammerten Bereiche, die großen Zoomstufen entsprechen, sind jedoch im Kontext dieser Arbeit nur von geringer Bedeutung, da die Kameras vornehmlich im Inneren von Gebäuden eingesetzt werden.

Die Detektion der Harris-Ecken wurde grundsätzlich, abgesehen von einer abnehmenden Bildqualität mit steigender Bildweite, auch durch Linsenverzerrungseffekte negativ beeinflusst. So wichen die 2D-Positionen der Ecken mitunter deutlich von den durch das Raster prädiktierten Positionen ab, was teilweise nur über eine empirisch gewählte, bildspezifische und auch lokal variierende Reskalierung der Punkteraster auszugleichen war. Abbildung 2.6(c) zeigt ein exemplarisches Detektionsergebnis. Die blauen Kreuze markieren die durch das Raster implizierten Positionen der einzelnen Ecken der Quadrate, während die roten Linien auf die jeweils zugeordneten Harris-Ecken verweisen.

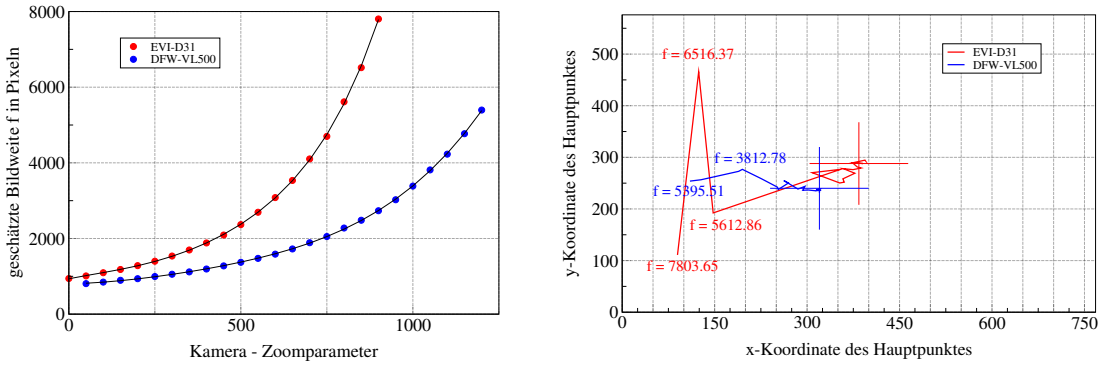



**Abbildung 2.7:** Offline-Kalibrierungsresultate für beide Kameratypen: Links sind die Abbildungsfunktionen der Zoomparameter auf die korrespondierenden Bildweiten gezeigt, während rechts die Ergebnisse der Hauptpunktschätzungen zu sehen sind. Dabei fallen insbesondere die großen Abweichungen der Werte von den Bildzentren mit einer zunehmenden Bildweite $f$ auf.

Die ermittelten Werte für $\alpha_x$ und $\alpha_y$ weisen für beide Kameratypen einen linearen Korrelationskoeffizienten von nahezu 1 auf, und interpolierte Regressionsgeraden besitzen in beiden Fällen eine Steigung von 1,0032 bzw. 1,0023. Dies rechtfertigt die Annahme quadratischer CCD-Elemente für beide Kameratypen, so dass jeweils eine einzelne Funktion genügt, um den Zusammenhang zwischen den Zoomparametern und der Bildweite zu beschreiben. Letztere wird als Mittelwert von $\alpha_x$ und $\alpha_y$ definiert. Aufgetragen gegen die Hardwareparameter der Kameras (Abb. 2.7, links) implizieren die Werte deutlich einen polynomialen Abbildungszusammenhang, wobei für den Kameratyp „EVI-D31“ ein interpoliertes Polynom fünften Grades den geringsten mittleren quadratischen Approximationsfehler aufweist, während die Werte für die Kameras des Typs „DFW-VL500“ durch ein Polynom vierten Grades am besten beschrieben werden können (schwarze Linien in Abb. 2.7, links). Die rechts in Abbildung 2.7 dargestellten Schätzungen der Hauptpunkte beider Kameras zeigen jeweils für kleine und mittlere Hardwareparameter eine deutliche Übereinstimmung mit den durch die Fadenkreuze markierten Bildzentren (der Kalibrierung der „EVI-D31“-Kamera lagen Bilder im Format $768 \times 576$ zu Grunde, mit den „DFW-VL500“-Kameras wurden Bilder im Format $640 \times 480$ aufgenommen). Je größer die Hardwareparameter jedoch gewählt werden, desto größer werden die Abweichungen. Gleichzeitig nimmt die nummerische Stabilität innerhalb der ermittelten Werte ab.

Vergleichbare Effekte sind auch bei Untersuchungen zur Reproduzierbarkeit der Kalibrierresultate zu beobachten, in deren Rahmen für ausgewählte Kameraeinstellungen wiederholt Kalibrierschritte durchgeführt wurden. Bei diesen Analysen zeigt sich mit zunehmenden Bildweiten $f$ bei beiden Kameras eine leicht ansteigende Varianz in den Ergebnissen, wobei die Standardabweichung im Verhältnis zu den geschätzen mittleren Werten jedoch zumeist noch immer deutlich unter 10% liegt. Lediglich die Hauptpunktkoordinaten fallen für große $f$ aus diesem Rahmen (vgl. auch Abb. 2.7, rechts).

Die aus dieser Offline-Kalibrierung resultierenden, intrinsischen Parameter der Kameras, d.h. insbesondere auch die Bildweite, können Online-Algorithmen grundsätzlich in Form von Look-Up-Tabellen oder direkt über die interpolierten Funktionen zugänglich gemacht werden. Dies bindet die Algorithmen jedoch an eine Verwendung von Kameratypen, die ein Auslesen ihrer aktuellen Parametereinstellungen zulassen und entsprechend

im Vorfeld einer Offline-Kalibrierung unterzogen werden müssen. Als flexiblere Alternative wurden daher auch die vorgestellten Online-Verfahren erprobt, wobei nur die Kamera des Typs „EVI-D31“ aufgrund der höheren Anzahl rotatorischer Freiheitsgrade zum Einsatz kam. Die direkte Berechnung der Bildweite $f$ mit dem von Shum und Szeliski vorgeschlagenen Ansatz erscheint dabei grundsätzlich zu instabil. Bei der Online-Kalibrierung mit Hilfe des absolute Conic ist dagegen eine variierende Güte und Stabilität der Ergebnisse in Abhängigkeit von den zu Grunde gelegten Annahmen und Kamerabewegungen zu beobachten. In allen Schätzungen wurden jeweils quadratische CCD-Elemente (begründet durch die Daten der Offline-Kalibrierung) und ein vernachlässigbar kleiner Skew angenommen.

Eine gleichzeitige Schätzung von Hauptpunkt und Bildweite hat sich zunächst als sehr unzuverlässig erwiesen. Dies bestätigt die Ausführungen von de Agapito und Kollegen, wonach eine Verschiebung des Hauptpunktes im Bild oftmals kaum von einer Rotation zu unterscheiden ist [dA00]. Die vorherige, und aufgrund der Offline-Kalibrierung auch gerechtfertigte Fixierung des Hauptpunktes im Bildzentrum führt demgegenüber zwar zu einer deutlichen Stabilisierung der Resultate, die Bildweite wird aber auch mit dieser Voraussetzung im Allgemeinen deutlich überschätzt. Die Werte weisen zudem eine hohe Varianz auf. Dies kann darauf hindeuten, dass die durchgeführten Kamerabewegungen möglicherweise keine hinreichende Grundlage für eine robuste Kalibrierung darstellen und daher Singuläritäten und Inkonsistenzen in den verwendeten Homographien vorliegen. De Agapito stellt in dieser Hinsicht insbesondere Rotationen um einzelne Achsen und nur kleine Rotationswinkel als problematisch heraus, da in diesen Fällen eine Bildweitenänderung oftmals nicht von einer Rotation unterschieden werden kann.

Eine Verbesserung der Qualität der aus einer Autokalibrierung resultierenden Kameraparameter ist einerseits bei größeren Kamerarotationen zu erwarten. Dadurch nimmt jedoch auch die Komplexität der Homographieschätzungen zu (Kap. 3), was wiederum eine Erhöhung des dortigen Fehlerrisikos und folglich eine weniger zuverlässige Datengrundlage nach sich zieht. Andererseits wurden bislang Effekte einer Linsenverzerrung in keiner Weise berücksichtigt, so dass auch dadurch verbesserte Resultate zu erzielen sein dürften [Tor00a]. Die Güte einer Autokalibrierung mit den diskutierten Ansätzen hängt allerdings in jedem Fall eng mit der Qualität der geschätzten Homographien zusammen. Da Schätzfehler insbesondere bei einer Online-Bildregistrierung, wie sie im Rahmen dieser Arbeit angestrebt wird, niemals vollständig ausgeschlossen werden können, sind Fehler in einer Bildweitenbestimmung mit Autokalibrierverfahren kaum zu vermeiden. In der Mosaikbildberechnung wird daher einstweilen von einer Bestimmung der Bildweite durch eine Autokalibrierung abgesehen und auf die insgesamt stabileren und auch vermeintlich zuverlässigeren Daten aus der Offline-Kalibrierung zurückgegriffen. Die für die Multi-Mosaikbilder notwendigen Bildweiten werden somit aus den jeweils aktuellen Hardwareparametern und den Interpolationspolynomen berechnet, die im Rahmen der Offline-Kalibrierung für die Abbildung der Zoomparameter auf die korrespondierenden Bildweiten bestimmt wurden. Auch dabei bleibt allerdings abschließend anzumerken, dass mit den geschätzten Werten zwar gute Ergebnisse erzielt werden konnten, eine explizite Korrektur von Linsenverzerrungen jedoch weitere Verbesserungen erwarten lässt.

# 3 Robuste Bildregistrierung

Ein wichtiger Forschungsschwerpunkt der digitalen Bildverarbeitung liegt auf der Analyse von Bildfolgen. Gegenüber unzusammenhängenden Einzelbildern bieten sie den Vorteil, nicht nur zeitlich punktuelle Informationen abbilden zu können, sondern auch visuelle Muster, die sich über einen längeren Zeitraum erstrecken. Auf diese Weise wird die für die Interpretation einer beobachteten Szene oder Situation zur Verfügung stehende Basis an visuellen Informationen breiter und ein Verständnis möglicherweise erleichtert. Allerdings steht dem Informationsgewinn ein erhöhter Aufwand bei der Extraktion der Daten aus der Bildfolge gegenüber. Eine der Kernaufgaben, die es dabei zu bewältigen gilt, besteht darin, die Bildinformationen einzelner Bilder zueinander in Beziehung zu setzen. Bei Verwendung einer bewegten Kamera muss dafür in der Regel zunächst die Kamerabewegung selbst geeignet kompensiert werden, bevor im Anschluss daran beispielsweise für eine Interpretation relevante Veränderungen zwischen den Bildern der Folge detektiert und ausgewertet werden können (vgl. auch Kap. 5).

Im Forschungsfeld der *Bildregistrierung* werden hierzu Ansätze und Methoden entwickelt. Die Grundlage bildet dabei im Allgemeinen ein adäquates Bewegungsmodell, für das aus den gegebenen Bilddaten Parameter ermittelt werden können. Sie ermöglichen die Bestimmmung der gesuchten, zueinander korrespondierenden Bereiche in den einzelnen Bildern. Die Komplexität der möglichen Bewegungsmodelle reicht von Transformationen mit verhältnismäßig wenigen Parametern (z.B. affin, biquadratisch, projektiv) bis hin zu elastischem Matching (z.B. [Mod04]) und optischem Fluss [Hor80], wo im Extremfall jedem Pixel ein eigener Parametervektor zugeordnet wird. Ebenso vielfältig wie die Modelle selbst sind auch die Verfahren zur Parameterschätzung, die sowohl direkt auf Basis der Intensitäts- oder Farbwerte, auf Korrespondenzen extrahierter Merkmale oder auch im Frequenzraum arbeiten können. In [Bro92] finden sich eine einordnende Übersicht der verschiedenen Modelle, Lösungsansätze und auch Anwendungsfelder.

Im Rahmen der vorliegenden Arbeit bildet eine Registrierung gegebener Bilder den ersten wichtigen Schritt bei der Erstellung eines Mosaikbildes aus einer Bildfolge. In diesem Kapitel werden die dabei verwendeten Ansätze und Algorithmen vorgestellt und diskutiert. Ausgehend von den Darstellungen im vorhergehenden Kapitel wird dabei als Bewegungsmodell für die stationären, rotierenden und zoomenden Kameras eine *projektive Transformation (Homographie)* zu Grunde gelegt. Unterkapitel 3.1 enthält zunächst einen allgemeinen Überblick über Verfahren zur Schätzung von Parametern eines Bewegungsmodells aus den Daten zweier Bilder. In dieser Arbeit findet im Kern das Verfahren des *Projective Flow* [Man96] Anwendung (Abschnitt 3.2), das jedoch in zweierlei Hinsicht erweitert wurde. Einerseits erfolgte eine Erhöhung der Robustheit durch eine gezielte Se-

lektion einzelner Pixel für die Schätzung (Abschnitt 3.2.1), und andererseits wurde eine optionale Initialisierungsphase zugefügt (Abschnitt 3.2.2). Während der Anwendungsbereich des *Projective Flow* aufgrund seines mathematischen Konzepts prinzipiell auf nur kleine Bewegungen der Kamera beschränkt ist, ermöglicht eine explizite Initialisierung die Verarbeitung von Bildfolgen, die auch große Kamerabewegungen einschließen. Dies ist insbesondere im Hinblick auf das skizzierte Anwendungsszenario des visuellen Speichers von Interesse, da im Allgemeinen auch Bilddaten repräsentiert werden sollen, die den kompletten Sichtbereich einer stationären, rotierenden Kamera umfassen und zu starke Einschränkungen der zulässigen Kamerabewegungen dabei nicht wünschenswert sind.

Neben diesen Aspekten besteht eine Bildfolge zudem üblicherweise aus mehr als nur zwei Bildern. Damit erwächst für die Erstellung eines Mosaikbildes die Notwendigkeit, mehrere Bilder zueinander zu registrieren. In Unterkapitel 3.3 wird diese Problematik näher beleuchtet, wobei eine robuste, inkrementelle Online-Schätzung von Parametern im Vordergrund steht. Dieser Ansatz ist vorrangig bei der inkrementellen Verarbeitung von erst nach und nach verfügbar werdenden Bilddaten von Bedeutung, wie sie in interaktiven Systemen anfallen. Er stellt damit eine wichtige Grundlage für den angestrebten Einsatz des visuellen Speichers in diesen Systemen dar. Das Kapitel schließt mit einem Abschnitt über die Transformation von Bildern in der Praxis und einer detaillierten, experimentellen Analyse der vorgestellten Algorithmen.

## 3.1 Verfahren der Parameterschätzung

Die Schätzung eines Parametervektors zu einem gewählten Bewegungsmodell aus den Daten zweier Bilder stellt grundsätzlich ein Optimierungsproblem dar. Gegeben sind das Bewegungsmodell $T_{\vec{p}}$, die beiden Bildfunktionen[1] $I_1$ und $I_2$ mit ihren jeweiligen Definitionsbereichen $\mathfrak{D}(I_1)$ und $\mathfrak{D}(I_2)$, sowie eine geeignete Zielfunktion $E$. Gesucht wird dann ein Parametervektor $\vec{p}^*$ für das Bewegungsmodell, so dass die Zielfunktion unter den gegebenen Bilddaten einen optimalen Wert[2] annimmt:

$$\vec{p}^* = \underset{\vec{p}}{\operatorname{argmin}}\, E(T_{\vec{p}}, I_1, I_2).$$

Grundsätzlich ist der Ausgangspunkt zur Lösung dieses Optimierungsproblems durch die Annahme gegeben, dass in den Daten der Bilder redundante Informationen enthalten sind, die Rückschlüsse auf die gesuchten Parameter des Bewegungsmodells zulassen. Allerdings sind die aus dieser Annahme resultierenden, praktischen Lösungsansätze vielfältig. Sie lassen sich dennoch anhand der Charakteristika der zu Grunde gelegten Daten grob in zwei Klassen einordnen: *merkmalsbasierte* und *merkmalslose* Ansätze.

[1] Zur einfacheren mathematischen Handhabung (etwa bei einer Berechnung von Ableitungen) werden Bilder in der vorliegenden Arbeit als diskrete, zweidimensionale Funktionen aufgefasst. Die Begriffe „Bild“ und „Bildfunktion“ werden dabei im Folgenden synonym verwendet.

[2] Hier wird davon ausgegangen, dass die Zielfunktion zu minimieren ist. Ein Maximierungsproblem lässt sich ggf. durch Multiplikation mit $-1$ in eine Minimierungsaufgabe überführen.

Algorithmen der ersten Klasse arbeiten auf Paaren korrespondierender Merkmale, die aus den Bildern extrahiert werden und damit von den exakten Funktionswerten der beiden Bildfunktionen abstrahieren. Demgegenüber legen Methoden der zweiten Klasse direkt die beiden Bildfunktionen $I_1$ und $I_2$ zu Grunde ohne sie einer weitergehenden Vorverarbeitung zu unterziehen. Eine Zuordnung von Verfahren zu einer dieser beiden Klassen ist dabei nicht immer eindeutig. In der Literatur finden sich auch hybride Ansätze, die die Vorteile beider Verfahrensklassen zu kombinieren versuchen (vgl. auch Abschnitt 3.2.1). Davis vollzieht im Vorfeld der Optimierung sogar einen vollständigen Wechsel des Basisbezugssystems vom Orts- in den Frequenzraum und abstrahiert damit gänzlich von den ursprünglichen Bildfunktionen [Dav98].

Die Zielfunktion $E$ hängt vorrangig von den der Optimierung zu Grunde gelegten Daten ab. In merkmalsbasierten Verfahren ist $E$ oft durch eine Distanzfunktion gegeben, die die Abstände korrespondierender Merkmale in den Bildern quantifiziert, bei merkmalslosen Algorithmen sind korrelationsbasierte Zielfunktionen verbreitet. Grundsätzlich kann eine Zielfunktion auch auf statistischer Basis begründet werden, wobei geeignete Wahrscheinlichkeitsdichten für die Modellparameter und Daten festzulegen sind. Die Optimierung entspricht dann beispielsweise einer Maximum-a-posteriori-Schätzung [Dae99].

## Merkmalsbasierte Verfahren

Die Datengrundlage merkmalsbasierter Ansätze bildet im Allgemeinen eine Menge $M$ korrespondierender Merkmalspaare zwischen den Bildern, auf deren Basis die Zielfunktion optimiert wird:

$$M(I_1, I_2) = \{\, (m_i, n_j) \,|\, m_i \in F(I_1), n_j \in F(I_2) \,\}.$$

$m_i$ und $n_j$ entsprechen dabei jeweils dem $i$-ten bzw. $j$-ten Element der nummerierten Mengen $F(I_1)$ und $F(I_2)$ extrahierter Merkmale aus beiden Bildern. Sie werden zunächst unabhängig voneinander in den Einzelbildern $I_1$ und $I_2$ detektiert und anschließend paarweise einander zugeordnet, wobei etwa geometrische Kriterien oder Korrelationsmaße Anwendung finden. Die Zielsetzung der Optimierung besteht dann in der Minimierung der Positionsabstände innerhalb aller ausgewählten Paare unter dem Bewegungsmodell.[3]

Die Merkmale selbst sind im Allgemeinen durch Strukturprimitiva wie Ecken oder Kanten gegeben. Zur Merkmalsextraktion eignen sich damit insbesondere gängige Operatoren wie der *Canny-Operator* [Can86] zur Detektion von Kanten oder auch der *SUSAN-Operator* [Smi97] für Ecken. Zum Quasi-Standard hat sich der *Harris-Corner-Detector* entwickelt. In [Har00] und [Cap04] bilden mit diesem Operator extrahierte und über lokale Korrelationsmaße einander zugeordnete Punkte die Basis zur Schätzung von Parametern für eine Homographie. Im Kern wird dabei ein aus den Korrespondenzen resultierendes, lineares Gleichungssystem mittels Singulärwertzerlegung gelöst. Dem Verfahren in [Zog97] liegen ebenfalls Harris-Ecken zu Grunde, wobei in die Zuordnung jedoch zusätzlich ein geometrisches Modell für die detektierten Ecken einbezogen wird.

[3] Neben einer reinen Minimierung der Abstände gegebener Merkmalspaare existieren auch Ansätze, die zusätzlich die exakten Positionen der Merkmale selbst bei der Optimierung verfeinern [Tor03].

Komplexere Merkmale, wie beispielsweise Kanten oder Linienzüge, sind gegenüber einzelnen Punkten schwieriger zu detektieren und zur Zuordnung werden aufwändigere Abstandsmaße benötigt. Die Homographieschätzung in [Kim00a] basiert dennoch auf extrahierten Linien, für eine robuste Zuordnung werden jedoch Kalibrierungsdaten der Kamera hinzugezogen. Ben-Ezra und Kollegen versuchen die mitunter fehlerträchtige, exakte punktweise Zuordnung von Merkmalen im Vorfeld der eigentlichen Schätzung gänzlich zu umgehen [BE98]. Sie detektieren zunächst im ersten Bild Punkte, die an prägnanten horizontalen und vertikalen Kanten liegen. Dann wird für jeden dieser Punkte über eine korrelationsbasierte Hough-Transformation eine korrespondierende Gerade im zweiten Bild festgelegt. Eine Minimierung der Abstände zwischen den korrespondierenden Punkten und Geraden unter Anwendung einer Homographie liefert final die gesuchten Parameter, ohne eine pixelweise Zuordnung vorauszusetzen.

Unabhängig von den konkret ausgewählten Merkmalen hat sich bei Verfahren dieser Klasse grundsätzlich eine hohe Sensitivität gegenüber Fehlern in der Position oder auch fehlerhaften Zuordnungen gezeigt. Im Allgemeinen wirkt sich dies negativ auf die Qualität der resultierenden Parametersätze aus. Da auch bei sehr robusten Merkmalsdetektoren Fehldetektionen und -zuordnungen nie gänzlich zu vermeiden sind, haben statistische Ansätze zur gezielten Auswahl verlässlicher Merkmalspaare für die Parameterschätzung eine große Bedeutung erlangt. Das wichtigste Verfahren ist dabei der *RANdom SAmple Concensus (RANSAC)* [Fis81], der bereits 1981 veröffentlicht und seitdem mehrfach erweitert wurde (z.B. [Tor00b]). Die Grundidee basiert auf einer iterativen, zufälligen Auswahl von Teilmengen aus $M(I_1, I_2)$. Für jede Teilmenge werden jeweils Parameter des Modells geschätzt, deren Qualität anschließend auf Basis geeigneter Maße bestimmt und als Bewertungskriterium der Teilmenge zu Grunde gelegt wird. Der optimale Parametersatz resultiert abschließend aus einer Schätzung auf Basis der am höchsten bewerteten Teilmenge und ist auf diese Weise im Allgemeinen robust mit einer hohen Genauigkeit zu ermitteln (vgl. auch [Tor97]).

## Merkmalslose Verfahren

Ein zweiter Ansatz zur Realisierung der Parameterschätzung für ein Bewegungsmodell $T_{\vec{p}}$ ist durch farb- bzw. intensitätsbasierte Verfahren gegeben, die ohne die Extraktion von Merkmalen aus den Bildern auskommen. Diese Algorithmen nutzen direkt die Funktionswerte der beiden Bildfunktionen, um optimale Parameter $\vec{p}^*$ zu ermitteln. Die Zielfunktion $E$ der Optimierungsaufgabe ist hier in der Regel durch eine Funktion gegeben, die Differenzen innerhalb der Farb- oder Intensitätswerte zwischen den Pixeln der beiden Bilder unter Anwendung des Bewegungsmodells quantifiziert. Diese Differenzen werden im Zuge der Optimierung minimiert. In der nachfolgenden Gleichung 3.1 ist als Beispiel hierzu eine oftmals verwendete Zielfunktion angegeben, in der die Summe der quadratischen Intensitätsdifferenzen über alle Pixel $u = (x, y, 1)^T$ gebildet wird, die im Überlappungsbereich der beiden Bilder $I_1$ und $I_2'$ liegen:

$$\vec{p}^* = \underset{\vec{p}}{\operatorname{argmin}} \sum_{u \in \mathfrak{D}(I_1) \cap \mathfrak{D}(I_2')} (I_1(u) - I_2'(u))^2 \quad , \quad I_2' = T_{\vec{p}}\,(I_2). \tag{3.1}$$

$I_2'$ bezeichnet dabei das aus der Anwendung der Transformation $T_{\vec{p}}$ auf das Bild $I_2$ resultierende, transformierte Bild. Die Lösung dieser Optimierungsaufgabe kann für Homographien über eine House-Holder-Transformation mit vorheriger Linearisierung der Zielfunktion durch eine Taylorreihen-Entwicklung [Syp99], oder auch mittels einer iterativen Levenberg-Marquardt-Minimierung [Sze96] geschehen. In [Ber92b], [Ira94] und [Man96] werden die Zielfunktionen zur merkmalslosen Schätzung von Parametern alternativ aus der Einschränkungsgleichung des optischen Flusses [Hor80] hergeleitet, wobei im letzten Fall eine Homographie als Bewegungsmodell dient. Das daraus resultierende Verfahren des *Projective Flow* bildet die Basis der Parameterschätzung in dieser Arbeit und wird daher im nachfolgenden Abschnitt genauer vorgestellt. Hansen und Mitarbeiter [Han94] schließlich errechnen zunächst mit Hilfe der Kreuzkorrelation ein Flussfeld zwischen den Bildern, an das sie dann ein affines Bewegungsmodell anpassen.

Die Optimierung der Zielfunktion merkmalsloser Verfahren erfolgt häufig durch iterative Algorithmen, wobei zumeist Ableitungen der Bildfunktionen einfließen. Diese werden im Allgemeinen auf Basis lokaler Intensitätsdifferenzen benachbarter Pixel approximiert. Insbesondere bei großen Bewegungen zwischen den Bildern, die über die in den Ableitungen berücksichtigten Nachbarschaften hinausgehen, lässt sich daher eine robuste Parameterschätzung nicht immer gewährleisten. Da bei großen Veränderungen außerdem das Risiko zunimmt, bei der Optimierung lediglich lokale Minima zu finden, sind merkmalslose Verfahren oft in einer Auflösungshierarchie organisiert [Ber92a]. Dabei werden die Bilder innerhalb einer Gauß-Pyramide in ihrer Auflösung reduziert. Anschließend lassen sich, von der Ebene mit der geringsten Auflösung an aufwärts, Parameter des gesuchten Bewegungsmodells ermitteln. Die in einer Ebene errechneten Parameter dienen dabei jeweils zur Initialisierung der Schätzung in der nächsthöheren Auflösungsebene, so dass die Bewegung stückweise rekonstruiert wird.

In [Saw99] erfolgt eine Schätzung der Parameter einer Homographie unter gleichzeitiger Korrektur von Linsenverzerrungen. Dabei werden nicht nur Parameter eines einzelnen Bewegungsmodells innerhalb einer Auflösungspyramide berechnet, sondern zusätzlich liegt dem Ansatz auch eine Hierarchie von Bewegungsmodellen zu Grunde. Angefangen bei einer 2D-Translation, über eine affine Transformation bis zu einer vollständigen Homographie werden jeweils Parameter bestimmt, wobei die für ein Modell ermittelten, optimalen Parameter als Ausgangspunkt der Schätzung des nächsten dienen. Eine solche Vorgehensweise ist insbesondere bei komplexen Bewegungsmodellen verbreitet, wo die direkte Schätzung der Parameter nichtlinear oder numerisch instabil ist. Homographien werden aus diesem Grund häufig durch einfacher zu handhabende Modelle approximiert (vgl. auch Unterkapitel 3.2).

Grundsätzlich bieten merkmalslose Verfahren gegenüber merkmalsbasierten Ansätzen den Vorteil, keine fehlerträchtige Extraktion und Zuordnung von Merkmalen aus den Bildern zu erzwingen. Allerdings führt ihre direkte Abhängigkeit von den Bildfunktionen zu einer hohen Sensitivität gegenüber Änderungen in der Gesamtenergie der Funktionen, wie sie z.B. bei wechselnden Lichtverhältnissen oder Kameras mit automatischer Blendenjustierung auftreten können. Um eine Parameterschätzung auch in diesen Fällen

zu ermöglichen, werden die Bildfunktionen im einfachsten Fall vor der Registrierung geeignet normiert. Darüber hinaus gehende Ansätze kombinieren z.B. eine Registrierung der Bilder mit einer simultanen, radiometrischen Korrektur (vgl. hierzu auch Abschnitt 4.1.1). Exemplarisch sei in diesem Zusammenhang die Arbeit von Jia and Tang [Jia05] angeführt, die auf einer leistungsfähigen, allerdings auch sehr aufwändigen Tensor-Voting-Strategie basiert.

## 3.2 Projective Flow

Die Grundlage zur Schätzung von Parametern einer Homographie in dieser Arbeit bildet das Verfahren des *Projective Flow*, das 1996 von Mann und Picard veröffentlicht wurde [Man96]. Es handelt sich dabei um einen merkmalslosen Ansatz. Ihm wurde der Vorzug gegenüber merkmalsbasierten Methoden gegeben, da auf diese Weise die mitunter schwierige und fehlerträchtige Merkmalsextraktion und -zuordnung im Vorfeld der Schätzung umgangen werden kann.

Den Ausgangspunkt des *Projective Flow* bildet der optische Fluss, durch den Veränderungen zwischen Bildern quantitativ beschrieben werden können. Er hat seinen Ursprung in der Hydrodynamik ([Jäh97], S. 431ff.), wo Strömungsbewegungen in Flüssigkeiten durch Kontinuitätsgleichungen modelliert werden. Ihnen liegt die Annahme zu Grunde, dass einzelne Teilchen einer Flüssigkeit in einem Zeitschritt lediglich ihre Position verändern, die Gesamtanzahl aber konstant bleibt. Übertragen auf Bilder folgt daraus analog, dass sich die Intensitätswerte der Pixel eines Bildes unter einer Kamerabewegung nicht verändern, sondern lediglich auf hinreichend eng benachbarte Pixel im Folgebild „transferiert“ werden sollten. Zu jedem Pixel $u = (x, y, 1)^T$ des ersten Bildes $I_t$, aufgenommen zum Zeitpunkt $t$, existiert dann ein korrespondierendes Pixel $u' = (x', y', 1)^T$ im nachfolgenden Bild $I_{t+\Delta t}$ mit identischem Grauwert, das lediglich um einen kleinen Betrag $\Delta u = (\Delta x, \Delta y, 0)^T$ in der Position verschoben ist. Dieser Zusammenhang wird durch die *Einschränkungsgleichung* des optischen Flusses formalisiert [Hor80]:

$$0 \;\approx\; (x' - x)\, I_t^{(x)} + (y' - y)\, I_t^{(y)} + I_t^{(t)} \;=\; \Delta x\, I_t^{(x)} + \Delta y\, I_t^{(y)} + I_t^{(t)}. \tag{3.2}$$

$\Delta x$ und $\Delta y$ geben die gesuchten, pixelweisen Verschiebungen pro Zeitschritt in $x$- und $y$-Richtung im Bild an, und $I_t^{(x)}$, $I_t^{(y)}$ und $I_t^{(t)}$ bezeichnen die lokalen, raum-zeitlichen Ableitungen der Bildfunktion. Das Bewegungsfeld des optischen Flusses lässt sich grundsätzlich durch diese Gleichung berechnen, wobei allerdings beschränkende Nebenbedingungen für das a priori unterbestimmte Optimierungsproblem hinzugezogen werden müssen.

Das Verfahren des *Projective Flow*, dessen Zielsetzung in der Berechnung einer Homographie zwischen zwei Bildern besteht, setzt auf der Einschränkungsgleichung des optischen Flusses auf. Zusätzlich wird dabei angenommen, dass sich die Pixelverschiebungen innerhalb eines Zeitschritts durch eine Homographie charakterisieren lassen und das Flussfeld somit durch das projektive Bewegungsmodell beschränkt werden kann. Für $\Delta x$ und $\Delta y$ eines Pixels gilt dann gemäß der Abbildungsgleichung für Homographien $H$

(Gl. 2.8) mit Parametervektoren $\vec{p} = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32})^T$ Folgendes:

$$\begin{bmatrix} \Delta x \\ \Delta y \\ 0 \end{bmatrix} = \Delta u = u' - u = H \cdot u - u \quad , \quad H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix}.$$

Aus der komponentenweisen Aufschlüsselung dieser Gleichung resultieren direkt

$$\Delta x = \frac{h_{11}\, x + h_{12}\, y + h_{13}}{h_{31}\, x + h_{32}\, y + 1} - x \quad \text{und} \quad \Delta y = \frac{h_{21}\, x + h_{22}\, y + h_{23}}{h_{31}\, x + h_{32}\, y + 1} - y.$$

Eingesetzt in Gleichung 3.2 lassen sich daraus pro Bildpixel zwei Bedingungsgleichungen als Grundlage zur Schätzung der optimalen Homographieparameter $\vec{p}^*$ gewinnen. Allerdings ist eine darauf aufbauende, direkte Berechnung mit Standardverfahren der Optimierung aufgrund der der Homographie inhärenten Nicht-Linearität schwierig. Die gesuchte Abbildung wird daher in der Praxis indirekt mit Hilfe eines iterativen, zweistufigen Verfahrens ermittelt, das im Folgenden skizziert wird. Die Grundidee besteht in einer schrittweisen Approximation der Homographie durch ein lineares Bewegungsmodell.

Grundsätzlich lässt sich eine nichtlineare Funktion durch eine Taylorreihe linear approximieren. Eine solche Entwicklung der Abbildungsgleichungen einer Homographie führt bei Vernachlässigung von Termen höherer Ordnung und einer zusätzlichen Parameterreduktion unter anderem auf eine *pseudo-perspektivische* Transformation:

$$\begin{aligned} x' &= q_1 + q_2\, x + q_3\, y + q_4\, x^2 + q_5\, xy \\ y' &= q_6 + q_7\, x + q_8\, y + q_4\, xy + q_5\, y^2. \end{aligned}$$

Sie kann zwar nicht alle unter einer Homographie zu beobachtenden Abbildungseffekte vollständig beschreiben, stellt jedoch für kleine Bewegungen nahe der Identität eine hinreichend exakte Approximation dar (vgl. auch [Man96]). Setzt man die daraus resultierenden Ausdrücke für $\Delta x$ und $\Delta y$ in die Einschränkungsgleichung 3.2 ein und summiert die Gleichungen aller Pixel im Überlappungsbereich der beiden Bilder quadratisch auf, so erhält man eine Zielfunktion für die Parameter $\vec{q} = (q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8)^T$ des pseudo-perspektivischen Bewegungsmodells:

$$\begin{aligned} \vec{q}^* = \underset{\vec{q}}{\operatorname{argmin}} \sum_{x,y} \Big( &\big(q_1 + q_2\, x + q_3\, y + q_4\, x^2 + q_5\, xy - x\big)\; I_t^{(x)} + \\ &\big(q_6 + q_7\, x + q_8\, y + q_4\, xy + q_5\, y^2 - y\big)\; I_t^{(y)} + I_t^{(t)} \Big)^2. \end{aligned} \quad (3.3)$$

Dieses Optimierungsproblem kann durch Ableiten und Nullsetzen bezüglich der einzelnen Komponenten $q_i$ auf ein lineares Gleichungssystem zurückgeführt und mit konventionellen Methoden eindeutig gelöst werden (Details zur Berechnung finden sich in [Syp99]). Aus der so ermittelten pseudo-perspektivischen Transformation lässt sich weiterhin durch die *4-Punkt-Methode* eindeutig eine korrespondierende Homographie bestimmen. Dazu werden unter Anwendung des optimalen Parametersatzes $\vec{q}^*$ vier Paare von Punkten[4]

[4]Diese Punkte dürfen nicht alle auf einer Geraden liegen.

mit ihren jeweiligen Abbildungen berechnet. Eine Einsetzung dieser Punktkorrespondenzen in die Transformationsgleichung für Homographien (Gl. 2.8) ergibt dann ein lineares Gleichungssystem mit acht Unbekannten bei acht Bedingungsgleichungen, das eindeutig bezüglich der gesuchten Parameter zu lösen ist. In der derzeitigen Implementierung bilden die vier Eckpunkte des Einheitsquadrates den Ausgangspunkt der Berechnungen, alternativ können aber auch die Eckpunkte des zu transformierenden Bildes verwendet werden. Im Mittel hat sich zwischen den aus beiden Ansätzen resultierenden Homographien lediglich ein geometrischer Fehler (s. Gl. 3.5) von maximal 0,04 Pixeln$^2$ gezeigt.

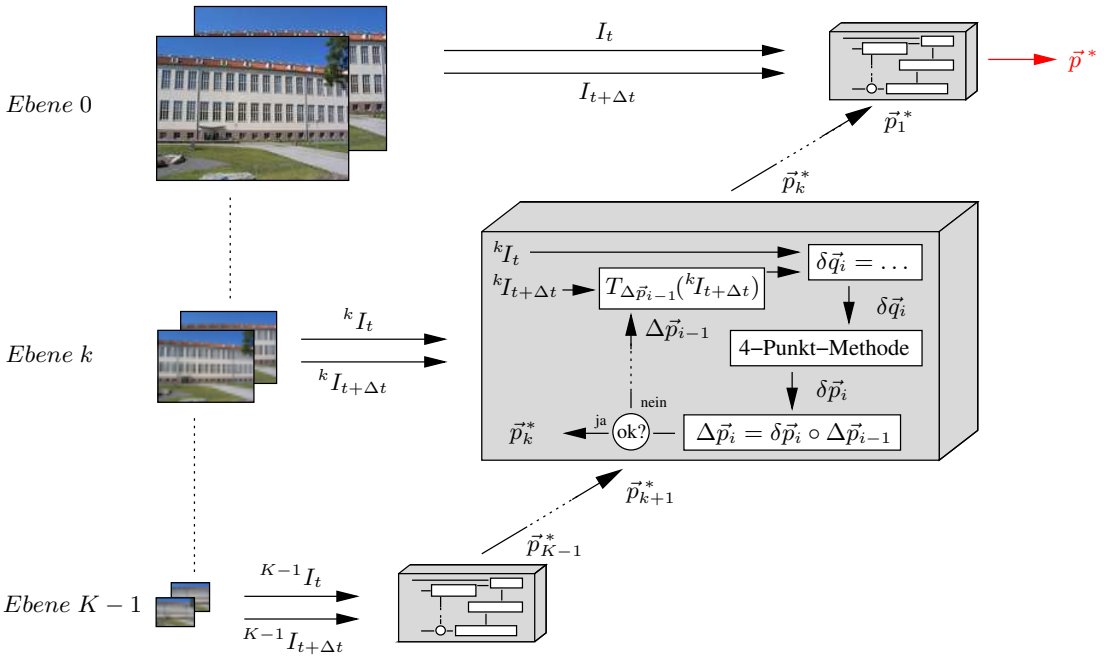



**Abbildung 3.1:** Schematischer Ablauf des *Projective Flow* in einer Pyramide mit $K$ Ebenen.

Mit der oben beschriebenen Vorgehensweise lässt sich eine Homographie berechnen, ohne dass das zuvor skizzierte, nichtlineare Optimierungsproblem gelöst werden muss. Da die pseudo-perspektivische Transformation die tatsächlich gesuchte Homographie jedoch im Allgemeinen nur approximiert, wird der zuvor beschriebene Schätzschritt in der Praxis wiederholt ausgeführt, bis die Homographie eine ausreichende Güte aufweist. Der vollständige Algorithmus, der in Abbildung 3.1 skizziert ist und im nachfolgenden Absatz noch genauer erläutert wird, arbeitet daher iterativ und zusätzlich innerhalb einer Auflösungspyramide. Durch letztere soll sichergestellt werden, dass nur kleine Bewegungen zwischen den jeweils betrachteten Bildern vorkommen und die stückweise Approximation der Homographie somit zulässig ist.

Die Grundlage der Parameterschätzung in jeder Pyramidenebene $k$ bilden gemäß der jeweiligen Ebene auflösungsreduzierte Versionen der Originalbilder ${}^kI_t$ und ${}^kI_{t+\Delta t}$. Das Bild ${}^kI_t$ dient dabei jeweils als Referenz, in dessen Koordinatensystem das Folgebild ${}^kI_{t+\Delta t}$ mit Hilfe der ermittelten Parameter überführt werden soll. In jeder Iteration $i \geq 1$ erfolgt zunächst eine Schätzung pseudo-perspektivischer Parameter $\delta\vec{q}_i$ zwischen dem

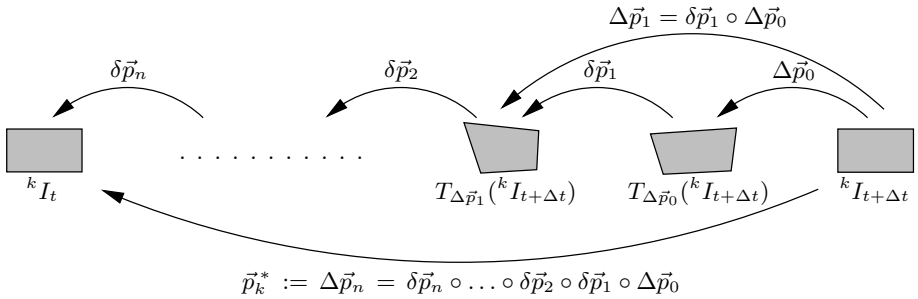



**Abbildung 3.2:** Stückweise Rekonstruktion der vollständigen Homographie $\vec{p}_k^*$ innerhalb einer Ebene $k$ der Auflösungspyramide durch Konkatenation der inkrementellen Parametersätze.

Referenzbild ${}^k I_t$ und dem aus der Anwendung des bis dato ermittelten Parametersatzes $\Delta\vec{p}_{i-1}$ auf ${}^k I_{t+\Delta t}$ resultierenden, transformierten Bild $T_{\Delta\vec{p}_{i-1}}({}^k I_{t+\Delta t})$. Diese werden dann über die 4-Punkt-Methode in projektive Parameter $\delta\vec{p}_i$ umgerechnet und mit dem bisherigen Parametersatz $\Delta\vec{p}_{i-1}$ zu neuen, vermeintlich besseren Parametern $\Delta\vec{p}_i = \delta\vec{p}_i \circ \Delta\vec{p}_{i-1}$ konkateniert[5] (Abb. 3.2). Dieser Zyklus wiederholt sich bis zum Erreichen eines festgelegten Abbruchkriteriums (s. unten). Der zugehörige Parametersatz $\vec{p}_k^*$ wird dann an die nächsthöhere Ebene zur Initialisierung der dortigen Schätzung $\Delta\vec{p}_0$ weitergereicht, wobei eine Reskalierung der Parameter die Auflösungsunterschiede in den Bildern zwischen den einzelnen Ebenen ausgleicht. In der untersten Pyramidenebene wird für $\Delta\vec{p}_0$ im Allgemeinen die Identitätsabbildung angenommen, aber auch alternative Initialisierungen sind möglich (vgl. Abschnitt 3.2.2).

Die Bewertung der Parameterqualität erfolgt nach jeder Iteration anhand eines geeigneten Maßes, auf dessen Basis über den Abbruch der Optimierung entschieden wird. Das Fehlermaß ist dabei im vorgestellten Verfahren von der eigentlichen Zielfunktion entkoppelt, da sich durch den zweistufigen Ansatz lediglich die Parameter der pseudoperspektivischen Abbildung direkt optimieren lassen, nicht jedoch die der Homographie. In der vorliegenden Arbeit findet der *normierte, mittlere quadratische Fehler (NMSE – Normalized Mean Squared Error)* in den pixelweisen Intensitätswerten der beiden Bilder Anwendung (Gl. 3.4), dessen Wahl im Rahmen einer detaillierten Diskussion seiner Eigenschaften in Abschnitt 3.5.1 begründet wird. Für seine Berechnung werden beide Bilder zunächst in ein einheitliches Koordinatensystem transformiert, das zumeist dem des Referenzbildes entspricht. Da die konkreten Fehlerwerte dabei stark von den Bildinhalten abhängen können, ist es schwierig, eine absolute Schwelle als Abbruchkriterium bei der Optimierung festzulegen. Aus diesem Grund dient die relative Verringerung des Fehlers pro Iterationsschritt als Kriterium, wobei in der vorliegenden Implementierung ein Abbruch bei einer minimalen Verbesserung des Fehlers um $10^{-6}$ erfolgt. Zusätzlich verhindert eine maximale Anzahl zulässiger Iterationen (üblicherweise 20) pro Pyramidenebene (von denen in der Regel 4 verwendet werden) eine Stagnation des Optimierungsprozesses.

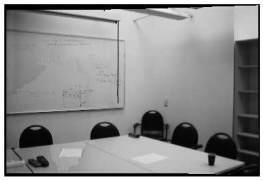
[5] Zur übersichtlicheren Darstellung werden die geschätzten Parametersätze an dieser Stelle mit den resultierenden Transformationen gleichgesetzt, so dass eine Konkatenation der Parametervektoren als Konkatenation der entsprechenden Homographien zu interpretieren ist.

### 3.2.1 Pixelselektion

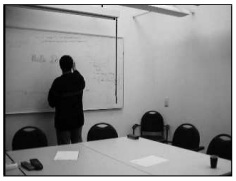
Das Verfahren des *Projective Flow* arbeitet, wie alle merkmalslosen Schätzverfahren, direkt auf den Bildfunktionen der zu registrierenden Bilder. Im Gegensatz zu merkmalsbasierten Ansätzen wird hier keine Existenz von markanten, robust detektierbaren Merkmalskorrespondenzen vorausgesetzt, sondern *alle* Pixel, die jeweils im Überlappungsbereich der beiden Bildfunktionen liegen, werden mit gleicher Gewichtung in die Schätzung einbezogen. Dabei bleibt oftmals unberücksichtigt, dass nicht alle Pixel denselben Informationsgehalt aufweisen und somit keine äquivalenten Beiträge zur Parameterschätzung leisten können. Insbesondere bei einer Verletzung der Grundannahme, dass Veränderungen in den Bildern allein durch die Kamerabewegung induziert werden, wirkt sich das negativ aus. Die Bewegungen von Bildbereichen, in denen dies zutrifft, lassen sich nicht durch das zu Grunde gelegte Modell beschreiben und korrumpieren damit den Schätzprozess. Weiterhin erfordert auch eine merkmalslose Schätzung, dass die Bildfunktionen eindeutig identifizierbare Charakteristika als Grundlage zur Registrierung der Bilder aufweisen. Vorrangig in homogenen, strukturarmen Bildregionen ist diese Bedingung oftmals nicht erfüllt, so dass dort eine eindeutige Identifikation korrespondierender Pixel allein auf Basis ihrer Intensitätswerte nicht möglich ist.

Die vorstehenden Ausführungen legen eine gezielte Selektion geeigneter Pixel für eine merkmalslose Parameterschätzung nahe. Insbesondere bei der Verarbeitung von Bildfolgen dynamischer Szenen verspricht eine Beschränkung auf Bildpunkte, die dem statischen Szenenhintergrund zuzurechnen sind, eine robuste Schätzung von Modellparametern [Ira94, Még99]. In [BE94] werden unabhängig bewegte Pixel in der Parameterschätzung maskiert, wobei zur Auswahl der Bildpunkte sowohl Bewegungsinformationen als auch strukturelle Daten in Form lokaler Gradientenbeträge ausgewertet werden. Auch Bober und Kollegen [Bob96] stellen ein verwandtes Verfahren zur Schätzung einer Fundamentalmatrix vor, in dem reichhaltig texturierte Bildregionen die Berechnungsgrundlage bilden. In den resultierenden, hybriden Algorithmen finden damit sowohl Daten aus den Bildfunktionen direkt als auch strukturelle Informationen Berücksichtigung, wodurch eine zuverlässigere Datenbasis für die Parameterschätzung zur Verfügung steht.

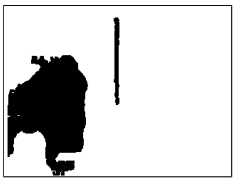
Der *Projective Flow* soll in der vorliegenden Arbeit durch die Bereitstellung robuster Transformationsparameter die Grundlage für eine Repräsentation von Bildfolgen statischer und dynamischer Szenen in Mosaikbildern bilden. Eine gezielte Selektion von „vertrauenswürdigen“ Pixeln für die Parameterschätzung leistet dazu einen wichtigen Beitrag und wurde daher in das Verfahren integriert. Dabei fließen vorrangig Bewegungsinformationen, aber auch strukturelle Daten auf Basis lokaler Gradienten in die Beurteilung der Bildpunkte ein [Möl03]. Die Bewegungsinformationen entstammen den Algorithmen zur Bewegungsanalyse aus Kapitel 5. Grundlage der Detektion unabhängig bewegter Pixel sind dort hohe Differenzen in den Intensitätswerten der registrierten Bilder, die als Eingangsdaten für eine binäre Klassifikation dienen. Die zusätzliche Integration der Daten über die Zeit reduziert dabei mittelfristig die Zahl von Fehlklassifikationen, so dass zuverlässige Bewegungsinformationen für die Pixelselektion zur Verfügung stehen. Zur Verringerung des Einflusses homogener Bildregionen in der Schätzung werden die lokalen



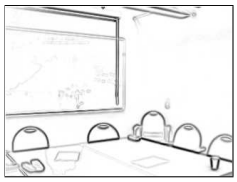
(a) Referenzbild



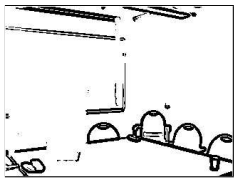
(b) aktuelles Bild



(c) Bewegungsdaten



(d) Bildgradienten



(e) Gesamtergebnis

**Abbildung 3.3:** Ergebnis der Selektion „vertrauenswürdiger“ Pixel (e) für zwei Bilder (oben) durch eine kombinierte Auswertung von Bewegungsinformationen (c) und lokalen Gradientenbeträgen (d).

Gradientennormen der Pixel analysiert, wobei die Annahme zu Grunde liegt, dass große Beträge mit ausgeprägten Bildstrukturen und damit einem hohen Informationsgehalt für die Registrierung korrelieren.

In der Implementierung erfolgt die Selektion geeigneter Pixel in jeder Iteration vor der Berechnung der pseudo-perspektivischen Parameter. Dabei wird für das Referenzbild eine Maske erstellt, in der Pixel mit geringem Informationsgehalt ausgeblendet werden. Der erste Schritt zur Selektion besteht in der Auswertung der Bewegungsinformationen aus der vorangegangenen Registrierung, wobei als bewegt klassifizierte Pixel direkt von der weiteren Verarbeitung ausgeschlossen werden. Die Menge dieser Pixel ist zwar unter Umständen nicht identisch mit den aktuell bewegten Bildpunkten, unter der Annahme nur kleiner Änderungen zwischen den Bildern fällt jedoch der überwiegende Teil unabhängig bewegter Pixel des aktuellen Bildes in den maskierten Bereich.

Für die verbleibenden Pixel $u$ des Referenzbildes werden die lokalen Gradientenbeträge $G(u)$ berechnet, wobei eine zuvor durchgeführte Glättung mit einer Gauß-Maske den Einfluss von Rauschen in den Bildern vermindert. Die finale Auswahl vertrauenswürdiger Pixel erfolgt schließlich anhand der Gradientenbeträge, wobei nur Pixel für die Parameterschätzung selektiert werden, deren Betrag einen festgelegten Schwellwert $\theta_G$ übersteigt (Abb. 3.3). Die Selektionsfunktion $S(u)$, die „vertrauenswürdigen“ Pixeln den Wert 1 zuweist, lässt sich damit insgesamt wie folgt zusammenfassen:

$$S(u) = \begin{cases} 0, & \text{falls} \quad R(u) > \theta_R \vee G(u) < \theta_G \\ 1, & \text{sonst} \end{cases}$$

Dabei entspricht $R(u)$ den lokal berechneten Intensitätsresiduen, anhand derer ein Pixel $u$ als bewegt eingestuft wird, wenn der Wert $R(u)$ den hierfür gewählten Schwellwert $\theta_R$ überschreitet (Details zur Bewegungsdetektion folgen in Kap. 5). Der Schwellwert $\theta_G$ für die lokalen Gradientenbeträge ergibt sich in Abhängigkeit vom Maximum $G_{\max} = \max_u\{G(u)\}$ aller Gradientenbeträge im Bild zu $\theta_G = \psi_G \cdot G_{\max}, \psi_G \in [0, 1]$.

Da die Verteilungen der Gradientenbeträge in den einzelnen Bildern einer Bildfolge stark variieren können, ist es schwierig einen allgemeingültigen Wert für $\psi_G$ anzugeben. Bei einem zu hohen Wert, der den Schwellwert $\theta_G$ sehr nah an den maximalen Gradientenbetrag im Bild legt, werden bei der Selektion nur wenige Pixel ausgewählt, so dass keine robuste Parameterschätzung mehr möglich ist. Ein zu niedrig gewählter Wert für $\psi_G$ schließt demgegenüber kaum Pixel von der Schätzung aus. Zur tatsächlichen Festlegung von $\psi_G$ findet daher eine adaptive Vorgehensweise Anwendung, die $\psi_G$ in Abhängigkeit von den aktuellen Bilddaten festlegt. $\psi_G$ wird dabei jeweils so gewählt, dass der Prozentsatz selektierter Pixel in einem Bild zwischen 10 und 25% aller gegebenen Pixel liegt. Dieses Intervall hat sich in empirischen Untersuchungen als geeignet herausgestellt.

Im Verlauf der Parameterschätzung wird der Wert von $\psi_G$ fortwährend durch eine Analyse des Anteils ausgewählter Pixel verifiziert. Liegt der Anteil dabei außerhalb des gültigen Bereichs, erfolgt eine schrittweise Erhöhung bzw. Verringerung von $\psi_G$ bis eine zulässige Pixelanzahl ausgewählt wird. Der in einem Schätzschritt für $\psi_G$ festgelegte Wert dient anschließend als Ausgangspunkt zur Pixelselektion im nachfolgenden Schritt, so dass lediglich zur initialen Bestimmung von $\psi_G$ eine umfangreichere Überprüfung des zulässigen Wertebereichs von $\psi_G$ notwendig werden kann. Im zeitlichen Verlauf sind dagegen im Allgemeinen schon geringe Modifikationen des Parameters jeweils ausreichend.

Die beiden vorstehend skizzierten Kriterien zur Pixelselektion erlauben zunächst nur eine binäre Klassifikation aller zur Verfügung stehenden Bildpunkte hinsichtlich ihrer Vertrauenswürdigkeit für eine Parameterschätzung. Auch zwischen den jeweils ausgewählten Pixeln können jedoch weitere Unterschiede in ihrer Relevanz bestehen. Dies motiviert eine zusätzliche, individuelle Gewichtung der einzelnen Pixel in der Parameterschätzung selbst (d.h. bei der Berechnung der pseudo-perspektivischen Transformationen nach Gl. 3.3), wodurch ihr Einfluss weiter differenziert werden kann.

Als mögliches Kriterium für eine derartige Gewichtung hat sich in der vorliegenden Arbeit der euklidische Abstand der einzelnen Punkte zum aktuell geschätzten Zentrum des Überlappungsbereichs der zu registrierenden Bilder bewährt. Pixel, die nahe an den Rändern dieses Bereichs liegen, können somit einen größeren Einfluss auf die Parameterschätzung ausüben. Dem Kriterium liegt die Beobachtung zu Grunde, dass mit einer korrekten Registrierung der Randbereiche eines Bildes zumeist auch eine weitgehend fehlerfreie Registrierung der zentralen Bildregionen verbunden ist, und den Randgebieten damit offensichtlich eine hohe Bedeutung bei einer Parameterschätzung zukommt.

### 3.2.2 Initialisierung

Eine grundsätzliche Voraussetzung für die Anwendbarkeit des *Projective Flow* zur Parameterschätzung leitet sich aus der lokalen Approximation der gesuchten Homographie durch pseudo-perspektivische Transformationen ab. Da dieses Vorgehen lediglich in der Nähe der Identitätsabbildung zulässig ist, lassen sich a priori auch nur für entsprechend kleine Bewegungen robust Parameter ermitteln. Durch die Auflösungspyramide können

zwar größere Bewegungen, die diesen Gültigkeitsbereich überschreiten, teilweise kompensiert werden, auch dieser Ansatz stößt aber bei sehr großen Bewegungen an seine Grenzen. Insbesondere wenn eine hohe Anzahl von Pyramidenebenen zur Kompensation der Bewegungen notwendig wird und die Bilder dadurch sehr weit in ihrer Auflösung reduziert werden müssen, ist die Datenbasis in unteren Ebenen oftmals nicht mehr ausreichend, um eine robuste Schätzung zu gewährleisten. Darüber hinaus nimmt mit größeren Rotationen der Kamera auch der Überlappungsbereich der Bilder ab, so dass das Risiko steigt, bei der Optimierung nur lokale Minima zu finden (vgl. hierzu aber auch den Ansatz des *Frame-to-Mosaic* in den Abschnitten 3.3.1 bzw. 6.4.3).

Eine mögliche Lösung des Problems besteht in einer gezielten Initialisierung der Parameterschätzung auf der untersten Pyramidenebene durch einen von der üblicherweise angenommenen Identität abweichenden Parametersatz. Wenn für die zwischen den Bildern $I_t$ und $I_{t+\Delta t}$ durchgeführte Bewegung eine hinreichend exakte initiale Schätzung $\Delta\vec{p}_0$ vorliegt, so ist nach deren Anwendung auf das aktuelle Bild nur noch eine kleine Kamerabewegung (nahe der Identität!) zwischen dem Referenzbild $I_t$ und dem transformierten Bild $T_{\Delta\vec{p}_0}(I_{t+\Delta t})$ zu ermitteln. Eine solche Initialisierung kann u.U. auch bei nur kleinen Bewegungen zu einer Erhöhung der Robustheit und zu einer Reduktion der Anzahl notwendiger Iterationen bis zum Erreichen des Optimierungsziels beitragen.

Grundsätzlich sind drei verschiedene Ansätze zur Initialisierung einer Parameterschätzung verbreitet. Eine manuelle Vorgehensweise, bei der der Benutzer zur Eingabe von Punktkorrespondenzen aufgefordert wird, ist in interaktiven Systemen nicht praktikabel. In [Saw99] wird die Schätzung von Parametern für ein komplexes Bewegungsmodell zunächst hierarchisch auf einfachere Modelle zurückgeführt und mit den daraus resultierenden Parametersätzen initialisiert. In der Regel sind jedoch auch zur Berechnung der einfacheren Modelle wiederum Korrespondenzprobleme zu lösen, deren Komplexität insbesondere bei großen Veränderungen zwischen den Bildern nicht unterschätzt werden sollte. Eine dritte Möglichkeit zur Parameterinitialisierung besteht schließlich in der Nutzung gänzlich von den Bilddaten unabhängiger, externer Informationen über die Kamerabewegung. Coorg und Kollegen [Coo00] setzen etwa zur Bildaufnahme eine mobile Plattform ein, die auch Informationen über Position und Orientierung der Kamera liefert. Mit diesen Daten lässt sich ihr globaler Registrierungsalgorithmus initialisieren.

Zur Bildaufnahme im Rahmen dieser Arbeit wurden im Wesentlichen zwei Kameratypen der Firma Sony$^{\text{TM}}$ verwendet (Abb. 2.5 in Abschnitt 2.4). Da sowohl beide Typen wie auch der zusätzlich eingesetzte Stereokamerakopf eine Abfrage ihrer aktuellen Konfigurationen und Positionen erlauben, wurde hier eine optionale Initialisierung der Parameterschätzung auf Basis dieser Daten und einer daraus abgeleiteten Schätzung für die Bildweite (vgl. Unterkap. 2.4) realisiert. Die Grundidee besteht darin, die Bildebene des neuen Bildes zunächst durch geometrische Operationen in ihre angenommene Position relativ zur jeweils zu Grunde gelegten Referenzbildebene zu transformieren und daraus eine projektive Abbildung zwischen der neuen Bildebene und dem Referenzkoordinatensystem abzuleiten (Abb. 3.4). Die Transformation wird dabei anhand von vier Punkten der aktuellen Bildebene durchgeführt, zu denen die korrespondierenden Punkte

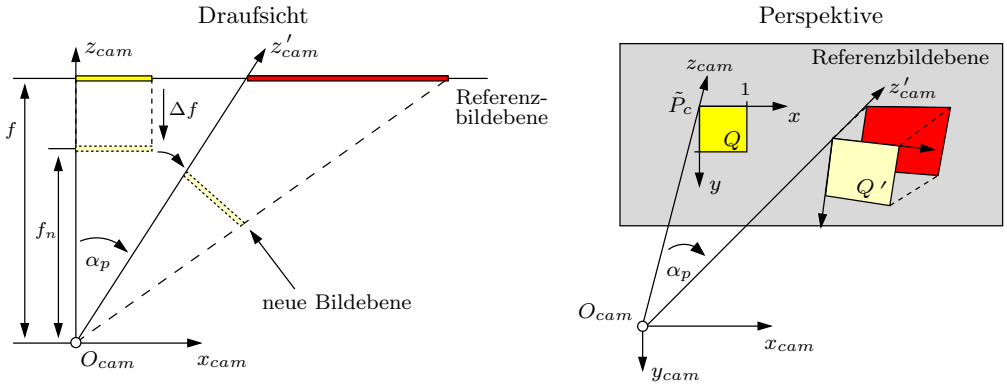



**Abbildung 3.4:** Skizze der geometrischen Zusammenhänge zur Berechnung einer projektiven Transformation für eine horizontale Kamerarotation um den Pan-Winkel $\alpha_p$ bei einer gleichzeitigen Bildweitenänderung um $\Delta f$: Zur Bestimmung der neuen Position der Bildebene wird sie zunächst um $\Delta f$ entlang ihres Normalenvektors verschoben und dann um $\alpha_p$ rotiert. Die Projektion der neuen Eckpunkte auf die Zielbildebene erlaubt schließlich die Rekonstruktion der zugehörigen Homographie.

auf der Referenzbildebene ermittelt werden, so dass sich die Homographie durch Anwendung der 4-Punkt-Methode (s. S. 31) berechnen lässt. Eine Menge möglicher Punkte, die leicht zu bestimmen ist, besteht dabei aus den vier Eckpunkten des Einheitsquadrats $Q$ der Ausgangsebene. Die Punkte werden in euklidische 3D-Weltkoordinaten[6] umgerechnet und gemäß den Änderungen in Bildweite und Orientierung der Kamera skaliert und rotiert. Abbildung 3.4 zeigt exemplarisch die bei einer horizontalen Kamerarotation um den Winkel $\alpha_p$ mit gleichzeitiger Bildweitenänderung von $f$ auf $f_n$ vorliegende Konfiguration.

Für die Berechnungen wird der Koordinatenursprung zur Vereinfachung in den Hauptpunkt der Ebene verlegt. Bei bekannter Position der Bildebene im Weltkoordinatensystem lassen sich aus den euklidischen 2D-Bildkoordinaten der Eckpunkte $\tilde{q}_i$ des Quadrats $Q$ direkt die zugehörigen Punkte $\tilde{Q}_i$ in 3D-Weltkoordinaten berechnen.

Diese werden im ersten Schritt gemäß der Bildweitenänderung der Kamera parallel zur optischen Achse verschoben, wobei $f$ und $f_n$ die alte und neue Bildweite und $\tilde{P}_c$ den Hauptpunkt in euklidischen 3D-Weltkoordinaten bezeichnen:

$$\tilde{Q}_i^s \;=\; \tilde{Q}_i - (1 - f_n/f) \cdot \tilde{P}_c \quad , \quad i \in \{1, 2, 3, 4\}.$$

Durch 3D-Rotationen können die resultierenden Punkte $\tilde{Q}_i^s$ dann in der Vertikalen um die x-Achse *(tilt)* und in der Horizontalen um die y-Achse *(pan)* in ihre neue Position gedreht werden, wobei die Drehwinkel $\alpha_t$ und $\alpha_p$ denen der Kamerabewegung entsprechen:

$$\tilde{Q}_i^r = \begin{bmatrix} \cos(\alpha_p) & 0 & \sin(\alpha_p) \\ 0 & 1 & 0 \\ -\sin(\alpha_p) & 0 & \cos(\alpha_p) \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha_t) & \sin(\alpha_t) \\ 0 & -\sin(\alpha_t) & \cos(\alpha_t) \end{bmatrix} \cdot \tilde{Q}_i^s.$$

Aus der Projektion dieser Punkte $\tilde{Q}_i^r$ auf die Referenzbildebene resultieren schließlich die Eckpunkte des transformierten Einheitsquadrats $Q'$ im Referenzkoordinatensystem.

[6] Das Weltkoordinatensystem kann dabei prinzipiell beliebig gewählt werden. Im Rahmen der Multi-Mosaikbilder erfolgt die Festlegung anhand der polyedrischen Grundkörper (vgl. Kap. 6).

Eine Rückrechnung ihrer Koordinaten in lokale 2D-Bildkoordinaten $\tilde{q}_i'$ ergibt gemeinsam mit den Ursprungspunkten $\tilde{q}_i$ vier Korrespondenzpaare, aus denen sich über die 4-Punkt-Methode eine projektive Transformation $T_{\vec{p}_0}$ zwischen dem neuen und alten Koordinatensystem ermitteln lässt.

Da in den Algorithmen üblicherweise von 'upper left'-Koordinaten ausgegangen wird, erfolgt abschließend noch eine Kompensation der Verlagerung des Koordinatenursprungs in den Hauptpunkt der Ebene. Dies geschieht durch Anwendung zweier zusätzlicher Abbildungen, die im Wesentlichen eine Verschiebung des Koordinatenursprungs um jeweils die halbe Bildbreite $w$ bzw. -höhe $h$ der Eingangsbilder realisieren:

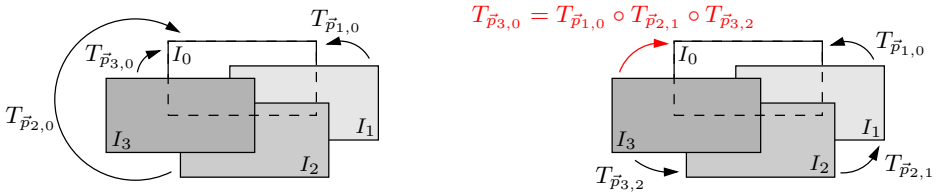
$$\Delta\vec{p}_0 = \begin{bmatrix} 1 & 0 & -w/2 \\ 0 & 1 & -h/2 \\ 0 & 0 & 1 \end{bmatrix} \cdot T_{\vec{p}_0} \cdot \begin{bmatrix} 1 & 0 & w/2 \\ 0 & 1 & h/2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Die so berechnete Homographie $\Delta\vec{p}_0$ kann nach einer Anpassung an die dortige Bildauflösung direkt zur Initialisierung der Parameterschätzung auf der untersten Ebene der Auflösungspyramide verwendet werden.

## 3.3 Online-Langzeitschätzung

Die Generierung eines Mosaikbildes aus einer Folge von mehreren Bildern setzt eine Transformation *aller* Bilder in das jeweils gewählte Referenzkoordinatensystem voraus. Damit ergibt sich die Notwendigkeit, die Schätzung von Parametern zwischen Paaren von Bildern auf eine vollständige, konsistente Registrierung einer kompletten Bildfolge auszuweiten. Wie bereits in der Einleitung skizziert wurde, sind dabei grundsätzlich Offline- und Online-Ansätze zu unterscheiden. Im ersten Fall ist die Datengrundlage zur Schätzung durch alle Bilder einer Folge gegeben, so dass bei der Optimierung eine globale (merkmalsbasierte oder merkmalslose) Zielfunktion zu Grunde gelegt werden kann. In [Kan00] bildet eine Graphrepräsentation der räumlichen Beziehungen zwischen den Bildern den Ausgangspunkt, wobei die intensitätsbasierte Parameterschätzung im Wesentlichen auf das Finden eines optimalen Pfades zurückgeführt wird. Oftmals erfolgt bei Offline-Algorithmen auch zunächst eine Berechnung lokaler Parameter zwischen einzelnen Paaren von Bildern, die dann im globalen Kontext verfeinert werden [Dav98, Gon98, Shu00]. Sowohl Sawhney et al. [Saw98] wie auch Gracias und Santos-Victor [Gra00] versuchen dazu, die topologische Anordnung der Bilder zu rekonstruieren und bei der gezielten Verbesserung der geschätzten Parametersätze zu berücksichtigen.

Offline-Verfahren weisen insbesondere im Hinblick auf einen Einsatz in interaktiven Systemen große Nachteile auf. Da die technische Ausstattung mobiler, interaktiver Systeme häufig noch nicht mit der aktueller, handelsüblicher PCs zu vergleichen ist, stehen in derartigen Systemen oftmals nur beschränkte Ressourcen zur Speicherung von Bilddaten zur Verfügung, die eine simultane Verarbeitung vollständiger Bildsequenzen nicht zulassen. Darüber hinaus ist in diesem Anwendungskontext häufig ein zeitlich nahezu uneingeschränkter Zugriff auf die Mosaikdaten erwünscht, den Offline-Verfahren in dieser




a) direkte Parameterschätzung b) inkrementelle Schätzung

**Abbildung 3.5:** Ansätze zur Online-Parameterschätzung: direkte Schätzung relativ zu einem gewählten Referenzbild $I_0$ (links) oder inkrementell mit anschließender Parameterkonkatenation (rechts).

Form nicht gewährleisten können. Aus diesem Grund werden dort zumeist Online-Ansätze benötigt, die eine inkrementelle Verarbeitung von Bildfolgen unterstützen. Dabei sind jeweils nur das aktuelle Eingangsbild, das Mosaikbild selbst und ggf. ein zusätzliches Bild für die Registrierung zu speichern. Im einfachsten Fall ist dieses zusätzliche Bild durch das Referenzbild $I_0$ gegeben, das das Referenzkoordinatensystem festlegt und zu dem die einzelnen Bilder direkt registriert werden können (Abb. 3.5, links). Da bei einer solchen Herangehensweise jedoch nicht immer ein ausreichender Überlapp zwischen $I_0$ und dem aktuellen Bild gegeben ist, wird zumeist eine inkrementelle Schätzung relativ zum jeweiligen Vorgängerbild bevorzugt *(Frame-To-Frame-Modus)*. Die vollständige Abbildung $T_{\vec{p}_t} := T_{\vec{p}_{t,0}}$ eines Bildes $I_t$ ins Referenzkoordinatensystem ergibt sich dabei aus der Konkatenation der Transformationen zwischen den einzelnen Bildern (Abb. 3.5, rechts):

$$T_{\vec{p}_t} := T_{\vec{p}_{t,0}} \;=\; T_{\vec{p}_{1,0}} \circ T_{\vec{p}_{2,1}} \circ \ldots \circ T_{\vec{p}_{t,t-1}}.$$

Bedingt durch die bei Online-Algorithmen reduzierte Datenbasis sind die geschätzten Transformationsparameter zumeist nicht global optimal. Insbesondere bei der Verarbeitung von langen Bildsequenzen ist häufig eine mit der Zeit abnehmende Parameterqualität zu beobachten. Durch die stetige Konkatenation der einzelnen Parametersätze werden Schätzfehler akkumuliert. Während eine minimale Abweichung vom gesuchten optimalen Parametersatz bei einer einzelnen Schätzung zwischen zwei Bildern meist vernachlässigbar ist, führt eine Verknüpfung mehrerer fehlerbehafteter Transformationen mittelfristig zu sichtbaren Registrierungsfehlern im Mosaikbild. Dies wird insbesondere bei Kamerabewegungen deutlich, die eine wiederholte Aufnahme verschiedener Szenenteile zu unterschiedlichen Zeitpunkten einschließen („*Looping Path*“). Ein Beispiel für eine solche Bewegung ist ein vollständiger, horizontaler 360 °-Scan, bei dem das erste und letzte Bild der Folge überlappen (vgl. auch Abschnitt 6.5.2).

Das Problem nicht-konsistenter Parametersätze bei einer inkrementellen Parameterschätzung wird in der Literatur häufig thematisiert. Konkrete Lösungsansätze, die *kein* Vorhandensein der kompletten Bildfolge verlangen, finden sich allerdings kaum. Eine Ursache hierfür liegt darin begründet, dass global optimale Parameter zu einem Zeitpunkt tatsächlich nur dann zu bestimmen sind, wenn alle im Mosaikbild zu repräsentierenden Daten gleichzeitig in der Optimierung berücksichtigt werden. Da letzteres bei Online-Algorithmen nicht gegeben ist, ziehen sich viele Autoren auf die Verwendung von globalen Techniken oder auf nachträgliche Korrekturschritte [Sze97] zurück. In den dort betrachteten Anwendungen hat im Allgemeinen aber auch die Fehlerfreiheit der

Mosaikbilder eine höhere Priorität als eine inkrementelle Berechnung. Darüber hinaus wird in vielen Arbeiten eine überschaubare Anzahl von Bildern zu Grunde gelegt und eine prinzipiell dauerhafte Verarbeitung von Bilddaten nicht angestrebt (Ausnahmen u.a. [Bur94, Ira95, Jet98]). Das mit den Mosaiks in dieser Arbeit verbundene Anwendungsfeld der interaktiven Systeme wird dabei insgesamt noch kaum berücksichtigt.

Einer der wenigen Ansätze, der zumindest eine Verringerung des Einflusses sich akkumulierender Fehler in den Transformationsparametersätzen verspricht und auch auf Online-Verfahren übertragen werden kann, ist die in verschiedenen Veröffentlichungen publizierte Schätzung von Parametern relativ zum aktuellen Mosaikbild (*Frame-to-Mosaic*, u.a. [Bur94, Ira95, Még99, Shu00]). Als Referenzbild dient dabei jeweils ein geeigneter Ausschnitt des aktuellen Mosaikbildes, so dass die Bilddaten früherer Zeitpunkte zumindest implizit in die aktuelle Schätzung einfließen. Grundsätzlich können zwar auch Ideen aus Offline-Verfahren zur Verbesserung der inkrementellen Parameterschätzung eingesetzt werden, wobei jedoch ein Kompromiss zwischen der Güte der Schätzung und einer Beeinträchtigung der Interaktivität durch die Offline-Berechnungen gefunden werden muss. Einen möglichen Lösungsansatz stellt beispielsweise eine blockweise Anwendung der Offline-Algorithmen auf Teilmengen der bis dato akquirierten Bilddaten dar. Dazu muss allerdings eine Mindestanzahl von Bildern gepuffert werden, was aufgrund der bereits skizzierten Ressourcenbeschränkung mobiler Systeme in der Praxis jedoch leicht zu Problemen führen kann. Weiterhin wird sich vermutlich auch durch diese Maßnahmen die Akkumulation von Registrierungsfehlern nur weiter zeitlich dehnen, nicht jedoch gänzlich beseitigen lassen. Aus diesem Grund wird in der vorliegenden Arbeit zunächst der Ansatz des *Frame-to-Mosaic* integriert (siehe nachfolgenden Abschnitt), kombiniert mit Fehler ausgleichenden Integrationsheuristiken (Abschnitt 4). Die online berechneten Mosaikbilder des visuellen Speichers stellen damit nicht zwangsläufig eine absolut fehlerfreie Repräsentation der akquirierten Bilddaten bereit. Sie lassen sich vielmehr als dynamische, adaptive Repräsentationsdatenstruktur interpretieren, die grundsätzlich vollständige Bilddaten der betrachteten Szene zur Verfügung stellt, in kleinen Teilbereichen jedoch zwischenzeitlich inkonsistente Daten enthalten kann.

### 3.3.1 Frame-to-Mosaic

Die Registrierung eines neuen Bildes mit inkrementellen Verfahren der Online-Parameterschätzung basiert zumeist nur auf dem aktuellen Bild und seinem direkten Vorgänger. Durch den *Frame-to-Mosaic*-Modus wird diese Datenbasis erweitert, ohne jedoch eine explizite Vorhaltung aller Bilder der Folge erforderlich zu machen. Vielmehr finden bei der Parameterschätzung auch Daten des bis zum aktuellen Zeitpunkt berechneten Mosaikbildes Berücksichtigung, so dass auf diese Weise eine verbesserte globale Konsistenz der Parametersätze erzielt werden kann.

Der Ansatz des *Frame-to-Mosaic* wird direkt in das in Abb. 3.1 skizzierte Verfahren des *Projective Flow* integriert. Dabei bleibt die vollständige hierarchische Struktur des Algorithmus erhalten (Abb. 3.6). Die wesentliche Änderung besteht darin, dass in den einzelnen Pyramidenebenen $k$ die auflösungsreduzierte Version des Vorgängerbildes ${}^{k}I_t$

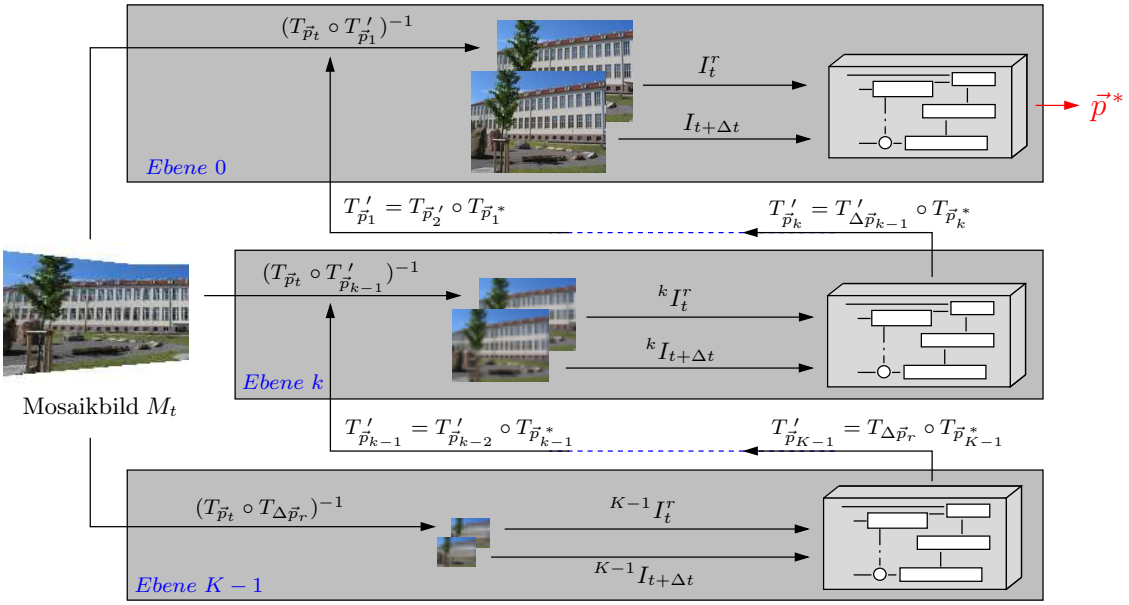



**Abbildung 3.6:** Parameterschätzung im *Frame-to-Mosaic*-Modus: im Vergleich zum ursprünglichen Verfahren (Abb. 3.1) werden lediglich andere Referenzbilder in der Schätzung zu Grunde gelegt.

durch einen geeigneten, transformierten Ausschnitt ${}^kI_t^r$ des bis zum aktuellen Zeitpunkt generierten Mosaikbildes $M_t$ ersetzt wird. Die Daten resultieren im Wesentlichen aus einer Anwendung des inversen, globalen Parametersatzes $T_{\vec{p}_t}^{-1}$ des Vorgängerbildes $I_t$ auf das Mosaikbild und stammen somit aus dessen Integrationsbereich. Damit sind sie mit denen von $I_t$ selbst vergleichbar und bilden eine gute Grundlage zur Parameterschätzung für kleine Bewegungen zwischen den Bildern. Bei größeren Bewegungen kann wie zuvor in der untersten Pyramidenebene eine initiale Schätzung $\Delta\vec{p}_r$ für die Bewegung einbezogen werden. Das Referenzbild wird in diesem Fall an der durch die Schätzung implizierten Stelle des Mosaikbildes ausgeschnitten (für $\Delta\vec{p}_0$ wird folglich die Identität angenommen). Auf diese Weise lassen sich selbst dann Parameter schätzen, wenn zwischen aufeinander folgenden Bildern kein Überlapp mehr gegeben ist, der betreffende Szenenbereich aber bereits im Mosaikbild repräsentiert ist (vgl. auch Abschnitt 6.4.3).

Das extrahierte Referenzbild und das aktuelle Bild werden mit Hilfe des bereits bekannten, iterativen Verfahrens (s. S. 32) registriert. Die resultierenden Parameter dienen anschließend zur Extraktion eines neuen Referenzbildes mit angepasster Auflösung in der nachfolgenden Pyramidenebene. Dabei ist zu berücksichtigen, dass die geschätzten Parameter $\vec{p}_k^*$ das aktuelle Bild lediglich bezüglich des zu Grunde liegenden Referenzbildes registrieren. Beim Ausschneiden eines neues Bildes in der nachfolgenden Ebene müssen daher neben $\vec{p}_k^*$ auch die Parameter durch Konkatenation einbezogen werden, die bei der Extraktion des vorherigen Referenzbildes angewendet wurden (vgl. Abb. 3.6). Durch die daraus resultierende, sukzessive Verkettung der Parametersätze wird das Referenzbild fortwährend an $I_{t+\Delta t}$ angeglichen und die Schätzung somit schrittweise verfeinert.

In der derzeitigen Implementierung wird in jeder Ebene nur zu Beginn der Schätzung ein Referenzbild ausgeschnitten, nicht jedoch nach einzelnen Iterationen. Dies geschieht vorrangig aus Effizienzgründen, da das Ausschneiden eines neuen Referenzbildes zusätz-

liche, aufwändige Interpolationen der einzelnen Farbwerte erforderlich macht (Kap. 3.4). Weiterhin sind aber auch die zu erwartenden Änderungen innerhalb des Referenzbildes zwischen einzelnen Iterationen klein, wenn eine hinreichend exakte initiale Schätzung für den gesuchten Parametersatz vorliegt. Eine zwischenzeitliche Aktualisierung des Referenzbildes lässt dann keine großen Verbesserungen der Schätzung vermuten.

## 3.4 Bildtransformation in der Praxis

Die Schätzung von Parametern für eine Homographie legt den Grundstein für die Überführung eines Bildes in ein anderes Koordinatensystem. Dabei wird das jeweilige Bild nicht nur nach Ablauf der Schätzung mit dem finalen Parametersatz ins Zielkoordinatensystem transformiert, sondern auch mehrfach im Verlauf der Schätzung selbst. Dies geschieht vorrangig, um die Qualität des aktuellen Parametersatzes zu bewerten (Abschnitt 3.5). Darüber hinaus werden auch die Referenzbilder im *Frame-to-Mosaic*-Modus über eine Bildtransformation aus dem aktuellen Mosaik extrahiert (Abschnitt 3.3.1). In diesem Abschnitt wird die praktische Realisierung dieser Transformationen beschrieben.

Ein Bild $I$ wird durch Anwendung einer Transformation $T_{\vec{p}}$ in das durch den Parametersatz $\vec{p}$ kodierte Zielkoordinatensystem projiziert. Die Rasterpunkte $u = (x, y, 1)^T$ des Zielkoordinatensystems, für die dabei Farbwerte bestimmt werden müssen, sind im Allgemeinen durch ganzzahlige Koordinaten $x, y \in \mathbb{N}$ gegeben. Eine direkte Anwendung der Homographie auf einzelne, ebenfalls ganzzahlige Pixelkoordinaten des Ausgangsbildes *(Vorwärtsprojektion)* führt dabei nur in Ausnahmefällen zu einer direkten Abbildung auf die gewünschten Rasterpunkte. Häufiger liegen die Projektionspunkte außerhalb des Rasters und die gesuchten Farbwerte der ganzzahligen Pixelkoordinaten können erst durch eine Interpolation ermittelt werden. Dazu sind jedoch geeignete Heuristiken notwendig, die eine Festlegung des Einflusses der Farbe eines spezifischen Projektionspunktes auf die Farbwerte der umliegenden Rasterpunkte ermöglichen. Insbesondere muss garantiert werden, dass für jeden gesuchten Rasterpunkt auch ein adäquater Farbwert ermittelt werden kann. Aufgrund der mit dieser Vorgehensweise offenkundig verbundenen Schwierigkeiten wird daher stattdessen oftmals eine *Rückwärtsprojektion* durchgeführt.

Dabei wird zunächst der Zielbereich der Projektion im Zielkoordinatensystem ermittelt. Dies geschieht in der Regel durch eine Vorwärtsprojektion der Eckpunkte des umschließenden Rechtecks des zu transformierenden Bildes. Alternativ kann der Bereich auch durch externe Anforderungen festgelegt werden. Im *Frame-to-Mosaic*-Modus beispielsweise resultiert das Zielgebiet der Projektion bei der Extraktion eines neuen Referenzbildes aus dem Definitionsbereich des aktuellen Bildes.

Im Anschluss an die Festlegung des Zielbereichs werden alle darin enthaltenen Pixel unter Anwendung der inversen Transformation $T_{\vec{p}}^{-1}$ ins Ausgangsbild zurückprojiziert. Aus den resultierenden Abbildungen lassen sich leicht die Pixel des Ausgangsbildes ermitteln, die einen Einfluß auf den Farbwert eines gesuchten Pixels $(x, y, 1)^T$ im transformierten Bild $I_t'$ haben (Abb. 3.7). Die Farbwerte selbst werden durch eine bilineare Interpolation berechnet, wobei die relevanten Bildpunkte gemäß ihres jeweiligen Anteils

$a_i$ an der Gesamtfläche $A$ des projizierten Pixels eingehen (Details siehe Abb. 3.8):

$$I_t'(x,y) = \frac{1}{A}\left(a_1 I_t(\lfloor x' \rfloor, \lfloor y' \rfloor) + a_2 I_t(\lceil x' \rceil, \lfloor y' \rfloor) + a_3 I_t(\lfloor x' \rfloor, \lceil y' \rceil) + a_4 I_t(\lceil x' \rceil, \lceil y' \rceil)\right)$$

$$(x',y') \;=\; T_{\vec{p}}^{-1}(x,y) \;\;,\;\; A \;=\; a_1 + a_2 + a_3 + a_4$$

Alternativ können die Farbwerte der Pixel in den transformierten Bildern auch über Splines oder vergleichbare Interpolationstechniken (s. z.B. [Pre92], Kap. 3) berechnet werden, solche Ansätze sind jedoch oftmals mit einem höheren Aufwand verbunden.

Bei Bildtransformationen ist grundsätzlich zu berücksichtigen, dass jede Interpolation Unschärfen verursacht und damit die Bildqualität vermindert. Mehrere aufeinander folgende Transformationen eines Bildes sollten daher vermieden werden. Im Rahmen der Parameterschätzung und -bewertung werden Parametersätze, die nacheinander auf ein einzelnes Bild anzuwenden wären, aus diesem Grund konkateniert und dann direkt auf das Originalbild angewendet. Auf diese Weise lassen sich negative Einflüsse durch mehrere aufeinander folgende Transformationen eines Bildes in der Optimierung ausschließen.

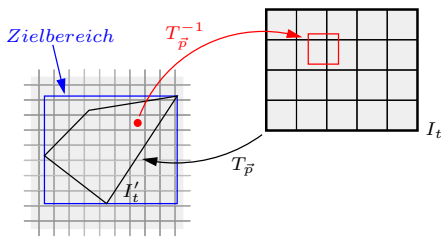



**Abbildung 3.7:** Rückwärtsprojektion: Die Pixel des Zielbereichs im transformierten Bild $I_t'$ werden jeweils über die inverse Transformation ins Ausgangsbild $I_t$ zurückprojiziert.

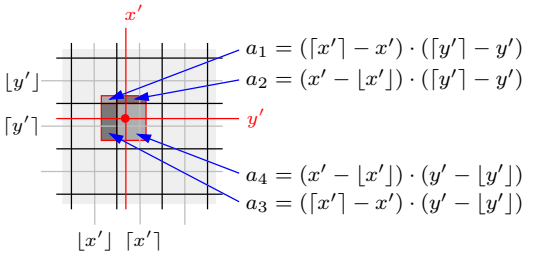



**Abbildung 3.8:** Bilineare Interpolation: Der Farbwert eines Pixels $(x',y',1)$ resultiert aus einer gewichteten Mittelung der Farbwerte seiner vier direkten Nachbarn.

## 3.5 Auswertung

Das in den vorangegangenen Abschnitten vorgestellte Verfahren des *Projective Flow* zur Schätzung von Homographieparametern stellt einen wichtigen Baustein im Prozess der Generierung von Mosaikbildern dar. Hochwertige Bilder lassen sich nur auf der Basis hinreichend exakter Transformationen berechnen. In diesem Kapitel werden die mit dem implementierten Algorithmus und den Erweiterungen erzielten Ergebnisse bei der Online-Parameterschätzung vorgestellt und diskutiert. Dabei steht die Güte der resultierenden Transformationen im Vordergrund. Da in Bezug auf das Anwendungsfeld der interaktiven Systeme außerdem die Effizienz der Algorithmen eine wichtige Rolle spielt, werden auch Laufzeitanalysen präsentiert.

### 3.5.1 Qualitätsmaße

Eine qualitative und quantitative Beurteilung von Schätzverfahren und der jeweils resultierenden Parameter (sowie damit auch der finalen Mosaikbilder) ist derzeit ein noch

unzureichend gelöstes Problem. In der Literatur finden sich neben Verfahren zur direkten Bewertung der Transformationen verschiedene Ansätze, die aus Ähnlichkeiten und Differenzen zwischen registrierten Bildern auf die Qualität der Parameter rückzuschließen versuchen. In diesem Abschnitt wird ein Überblick über die unterschiedlichen Herangehensweisen gegeben, aus denen sich die in dieser Arbeit verwendeten Maße motivieren.

Der grundlegende Ansatz zur Evaluation eines Verfahrens besteht zunächst in einer theoretischen Abschätzung der erzielbaren Genauigkeit. Dabei fließen in Abhängigkeit von den zu Grunde gelegten Daten wahlweise Annahmen über die Verteilung von Lokalisations- und Zuordnungsfehlern in merkmalsbasierten Verfahren, oder aber Hypothesen über den Rauschanteil im Bildsignal bei merkmalslosen Ansätzen in die Analyse ein. Da die Komplexität theoretischer Betrachtungen vorrangig mit dem Bewegungsmodell skaliert, waren solche Analysen lange Zeit auf verhältnismäßig einfache Modelle wie reine Translationen [McG76, Man93] konzentriert. Inzwischen existieren zumindest für die Berechnung von Homographien über merkmalsbasierte Verfahren ebenfalls Untersuchungen [Kan99, Har00]. Allerdings lässt sich aus den theoretisch optimalen Grenzen eines Algorithmus nicht zwangsläufig auch auf seine praktische Leistungsfähigkeit schließen. Die reale Performanz hängt gerade in merkmalslosen Ansätzen oftmals eng mit dem Informationsgehalt und Frequenzspektrum der Bilder zusammen [Rob03a], so dass die theoretischen Optima nur selten erreicht werden. Eine für die Praxis relevantere Alternative erscheint daher ein Vergleich mit kalibrierten Referenzdaten zu sein. Diese lassen sich jedoch oftmals nur unter sehr großem Aufwand gewinnen [Cap04].

In vielen Arbeiten zur Bildregistrierung und Mosaikbildberechnung wird die Qualität der Transformationen und Mosaikbilder neben manuellen, visuellen Inspektionen häufig anhand eines indirekten Bewertungsschemas beurteilt. Dazu wird ein Qualitätsmaß ausgewertet, das Ähnlichkeiten und Unterschiede zwischen den zu registrierenden Bildern quantifiziert. Die Entwicklung geeigneter Maße ist dabei bereits seit langem ein aktives Forschungsgebiet. Nichtsdestotrotz ist es noch immer schwierig, Unterschiede in Bildern adäquat zu charakterisieren. Insbesondere das oftmals formulierte Ziel, durch die Qualitätsmaße die vom Menschen wahrgenommenen Sinneseindrücke möglichst exakt widerzuspiegeln, ist schwierig zu erreichen. Ein Grund dafür besteht sicher darin, dass die komplexe menschliche Verarbeitung visueller Informationen erst unzureichend verstanden und damit auch nur schwer in Algorithmen nachzubilden ist.

Viele heute verbreitete Distanz- und Ähnlichkeitsmaße haben ihren Ursprung in der Übertragungs- und Netzwerktechnik, wo Qualitätseinbußen durch verlustbehaftete Kompressionsverfahren quantitativ beschrieben werden sollen. Nach [Wan03] ist das am häufigsten benutzte, gleichzeitig aber wohl auch am eindringlichsten kritisierte Distanzmaß der *Mean Squared Error (MSE)* in den Intensitätswerten zweier Bilder $I_1$ und $I_2$:

$$\mathrm{MSE}(I_1, I_2) = \frac{1}{|\mathfrak{D}(I_1) \cap \mathfrak{D}(I_2)|} \sum_{u \in \mathfrak{D}(I_1) \cap \mathfrak{D}(I_2)} (I_1(u) - I_2(u))^2.$$

Die Kritik begründet sich in erster Linie dadurch, dass Paare von Bildern, deren Differenzen jeweils durch den gleichen MSE charakterisiert werden, oftmals sehr unterschiedli-

che Arten von Fehlern aufweisen können. Darüber hinaus ist der MSE nur selten mit den Einschätzungen von Testpersonen vergleichbar, dem so genannten *Mean Opinion Score (MOS)* [Wan03], der zumeist als Referenz dient. Aus diesem Grund gewinnen zunehmend Ansätze an Bedeutung, die wahrnehmungspsychologische Erkenntnisse über das visuelle System des Menschen *(Human Visual System - HVS)* stärker berücksichtigen.

In die HVS-basierten Distanzmaße werden beispielsweise Adaptionsvorgänge und Maskierungseffekte einbezogen, die bei der menschlichen Wahrnehmung zu beobachten sind [Wan03]. Xu und Hauske [Xu94] gründen auf einem solchen Ansatz ein komplexes Modell, in dem eine Unterscheidung verschiedener Fehlertypen den Grundstein bildet. Viele HVS-basierte Qualitätsmaße sind gegenüber konventionellen Ansätzen allerdings aufwändiger in ihrer Berechnung. Dennoch ist ihre Leistungsfähigkeit umstritten, wie z.B. in den Arbeiten von [Esk95] und [Wan02] deutlich wird. Wang und Kollegen schlagen alternativ für eine Bewertung den *Universal Image Quality Index (UIQI)* vor, der ebenfalls perzeptiv motiviert, dabei aber auch effizient zu berechnen sein soll:

$$UIQI(I_1, I_2) = \frac{\sigma_{1,2}}{\sigma_1 \sigma_2} \cdot \frac{2\mu_1\mu_2}{\mu_1^2 + \mu_2^2} \cdot \frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}.$$

Dabei entsprechen $\mu_1$ und $\mu_2$ den mittleren Farbwerten der beiden Bilder, $\sigma_1^2$ und $\sigma_2^2$ den zugehörigen Varianzen. $\sigma_{1,2}^2$ schließlich quantifiziert die Kovarianz in den Farbwerten, so dass auch der UIQI im Wesentlichen auf der Auswertung der Farbvarianzen innerhalb der Bilder beruht.

Insgesamt ist die Suche nach adäquaten Qualitätsmaßen oftmals stark durch die Zielapplikation geprägt. Insbesondere im Hinblick auf die Datenübertragung fließen häufig Annahmen über potenzielle Störeinflüsse und zu erwartende Kompressionsartefakte (z.B. Blockbildung bei der JPEG-Kompression) direkt in die Entwicklung möglichst universeller Distanzmaße ein. Registrierungsfehler, wie sie bei einer Parameterschätzung auftreten können, sind jedoch kaum mit solchen Störungen zu vergleichen. Sie wirken selten gleichmäßig auf ein ganzes Bild ein, sondern treten vielmehr vereinzelt an Bildrändern oder in strukturarmen Regionen auf. Dementsprechend schwierig ist die Übertragung der oben skizzierten Qualitätsmaße auf die Bildregistrierung. Die Tatsache, dass in den meisten Arbeiten eine finale Beurteilung der Mosaikbilder dem Betrachter überlassen wird (z.B. [Man96, BE98, Dav98]) und objektive Beurteilungskriterien neben dem MSE so gut wie nicht existieren (einer der wenigen Ansätze in diesem Bereich ist in [Kim00b] zu finden), untermauert die damit offensichtlich verbundenen Schwierigkeiten.

In der vorliegenden Arbeit ist die Qualität eines Parametersatzes vorrangig wiederholt im Verlauf der iterativen Optimierung zu bewerten. Das dabei verwendete Fehlermaß sollte einerseits möglichst gute Hinweise für einen Abbruch der Optimierung geben, andererseits aber auch effizient zu berechnen sein. Eine manuelle, graphische oder visuelle Inspektion ist nicht praktikabel, HVS-basierte Maße scheinen ihren Berechnungsaufwand noch nicht zu rechtfertigen. Daher wird auch in dieser Arbeit trotz der vermeintlichen Nachteile im Wesentlichen eine normierte Version des MSE (NMSE) zu Grunde gelegt:

$$NMSE(I_1, I_2) = \frac{1}{|\mathfrak{D}(I_1) \cap \mathfrak{D}(I_2)|} \sum_{u \in \mathfrak{D}(I_1) \cap \mathfrak{D}(I_2)} \left( \frac{I_1(u)}{\mu_1} - \frac{I_2(u)}{\mu_2} \right)^2. \tag{3.4}$$

Die Normierung der Intensitätswerte der beiden Bilder mit ihren jeweiligen Mittelwerten $\mu_1$ und $\mu_2$ dient dabei vorrangig einem Ausgleich von großen Differenzen in den Bildenergien. Auf diese Weise wird näherungsweise der Effekt einer Gamma-Korrektur der beiden Bilder erzielt (vgl. [Jäh97], S. 249ff.), wie sie üblicherweise zum Ausgleich variierender Helligkeiten verwendet wird, deren Parameter sich jedoch im Rahmen eines automatisierten Verfahrens nur schwer bestimmen lassen.

Häufig wird für die Praxis eine blockweise Berechnung der verschiedenen Maße auf den Bildern vorgeschlagen, um eine verbesserte Detektion lokaler Fehler zu ermöglichen. Die Werte der einzelnen Blöcke werden dabei durch eine Mittelung zu einer Maßzahl zusammengefasst. Für die Bewertung von Transformationsparametern hat sich daraus in Vorversuchen jedoch kein Vorteil ergeben. Ebenso wenig ließen sich aussagekräftige Kriterien finden, um vollständige Fehlerhistogramme mit ihrem größeren Informationsgehalt der Analyse zu Grunde zu legen. Diese Ansätze werden daher in der nachfolgenden Auswertung nicht weiter berücksichtigt.

Abschnitt 3.5.2 stellt nun Ergebnisse der implementierten Parameterschätzung anhand ausgewählter, repräsentativer Beispielmosaikbilder und Bildsequenzen vor.[7] Gemäß der vorstehenden Ausführungen basiert die Berechnung der Bilder dabei auf dem NMSE als Fehlermaß. Allerdings stehen die konkreten Fehlerwerte in den nachfolgenden Auswertungen nicht im Vordergrund, da sie nur selten ein geeignetes Kriterium zur Beurteilung der tatsächlichen Bildqualität darstellen (s. oben). Sie können zwar grobe Registrierungsfehler zufriedenstellend quantifizieren, kleinere lokale Unstimmigkeiten spiegeln sich jedoch nur selten in den Werten wider. In den Beispielen, in denen absolute Fehlerwerte angegeben sind, sollten diese daher primär als Anhaltspunkte für eine grobe qualitative Einordnung der geschätzten Parametersätze aufgefasst werden.

Für die Auswertung der Pixelselektion wurde eine Bildfolge aufgenommen, in der auf eine Kamerabewegung verzichtet wurde, so dass mit der zwischen den Bildern vorliegenden Identitätsabbildung eine Referenz zur Bewertung der geschätzten Parameter gegeben ist. Zum Vergleich der ermittelten Homographien mit der Identitätsabbildung dient dabei der so genannte *geometrische Fehler* $G_\epsilon$ als Qualitätsmaß ([Har00], S. 77). Er quantifiziert die Unterschiede zwischen zwei Transformationen $T^r$ und $T$ durch den mittleren quadratischen Abstand zwischen den aus der Anwendung der Transformationen auf eine ausgewählte Punktemenge $x_i, i = 1 \dots n$, resultierenden, projizierten Punkten im Zielkoordinatensystem:

$$G_\epsilon = \frac{1}{n} \sum_{i=1}^{n} (T^r(x_i) - T(x_i))^2. \tag{3.5}$$

$T^r$ entspricht dabei der Referenztransformation, während $T$ die zu evaluierende, geschätzte Transformation kodiert. Als Punktemenge wurde in der vorliegenden Arbeit jeweils der Definitionsbereich des zu registrierenden Bildes zu Grunde gelegt.

[7] Die finalen Mosaikbilder sowie alle Einzelbilder der verschiedenen Sequenzen finden sich auch im Internet unter `http://www.informatik.uni-halle.de/~moeller/phd/`.

### 3.5.2 Ergebnisse & Diskussion

Die Implementierung des *Projective Flow*-Algorithmus folgt den Ausführungen in den vorangegangenen Abschnitten, wobei die in den Abbildungen 3.1 bzw. 3.6 skizzierten Ablaufschemata den Ausgangspunkt bilden. Das System ist grundsätzlich in der Lage, sowohl Grauwert- wie auch Farbbilder zu verarbeiten. Die eigentliche Schätzung der Parameter wird jedoch, analog wie auch die Bewegungsdetektion (vgl. Abschnitt 5.1.1 auf Seite 70), auf Grauwerte beschränkt. Eine Nutzung der vollständigen Farbinformationen im Verlauf der Schätzung hat sich in Vorversuchen als deutlich aufwändiger erwiesen, ohne jedoch eine Verbesserung der Parameterqualität zu bewirken.

Zur Berechnung der Ableitungen der Bildfunktionen, die zur Bestimmung der pseudo-perspektivischen Transformationen in jedem Schritt benötigt werden (vgl. S. 31), dienen die von Horn und Schunck in [Hor80] vorgeschlagenen Formeln. Sie berücksichtigen im Gegensatz zu vielen anderen Ansätzen *beide* Bilder nicht nur bei der Berechnung der zeitlichen Ableitungen, sondern auch zur Bestimmung der räumlichen Ableitungen. Dies hat sich im Kontext der vorliegenden Arbeit als vorteilhaft erwiesen. Die Bilder werden darüber hinaus vor der Berechnung der Ableitungen mit einer Gauß-Maske der Größe $3 \times 3$ und $\sigma = 1$ geglättet. Auch dies hat sich in Vorversuchen bewährt und bestätigt damit u.a. die Ausführungen zur Robustheit von merkmalslosen Schätzungen in [Rob03a].


**Abbildung 3.9:** Mosaikbild einer überwiegend planaren Szene mit künstlicher Beleuchtung, berechnet aus einer Bildfolge mit 224 Bildern (ausgewählte Beispielbilder s. Anhang A.1). Die eingezeichnete blaue Trajektorie markiert die Mittelpunkte der integrierten Bilder und deutet damit die bei der Bildaufnahme durchgeführten Kamerabewegungen an.

In Abbildung 3.9 ist zunächst ein Mosaikbild gezeigt, dem eine Sequenz mit 224 Bildern zu Grunde liegt (s. Anhang A.1). Die dargestellte Szene ist überwiegend planar und wurde künstlich ausgeleuchtet, um einen wechselnden Lichteinfall bei natürlichem Tageslicht auszuschließen. Das Mosaik weist insgesamt eine hohe Qualität auf, wie der direkte

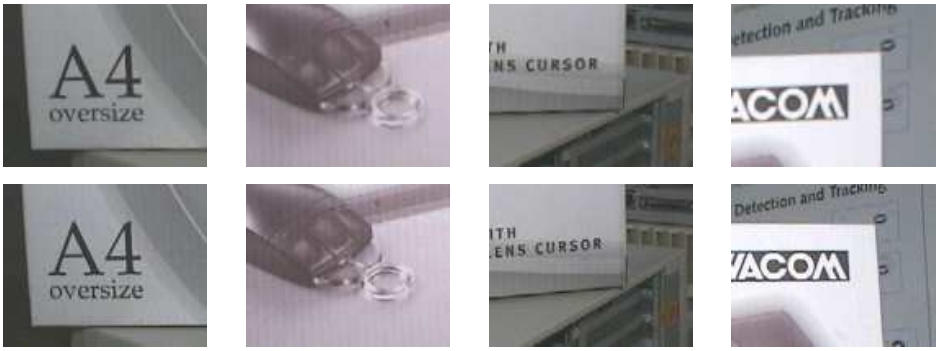



**Abbildung 3.10:** Veranschaulichung der Qualität des Mosaikbildes aus Abb. 3.9: Die obere Zeile zeigt Ausschnitte des Mosaiks, während in der unteren Zeile korrespondierende Ausschnitte aus den Einzelbildern der Bildsequenz (s. Anhang A.1) zu sehen sind. Das Mosaikbild weist eine leicht geringere Schärfe auf, die mit einer zunehmenden Bildverzerrung jedoch zunimmt (s. etwa Ausschnitt rechts).

Vergleich einiger Referenzausschnitte des Bildes mit den entsprechenden Originalbildern zeigt (Abb. 3.10). Insbesondere die linken drei Ausschnitte des Mosaikbildes weisen nur eine unwesentlich geringere Schärfe gegenüber den Originalbildern auf. Der Ausschnitt ganz rechts zeigt allerdings eine verminderte Bildqualität. Obwohl die korrespondierenden Bilder aufgrund der Kameratrajektorie erst spät in das Mosaikbild integriert werden (Abb. 3.9 bzw. A.1), spielt der Einfluss von Registrierungsfehlern in der Schätzung dabei durch den hier angewendeten Frame-to-Mosaic-Modus nur eine untergeordnete Rolle. Die abnehmende Qualität ist vielmehr auf das leichte Wachstum des Mosaiks bei zunehmenden Kamerarotationen und die damit einhergehende Bildverzerrung zurückzuführen.

Dieser Effekt ist auch im zweiten Beispielmosaik deutlich zu beobachten (Abb. 3.11). Dort wurde ein Mosaikbild von einer 3D-Szene aufgenommen. Auch dieses Bild weist im Vergleich zu den Originalbildern der Bildsequenz eine hohe Qualität auf (Abb. 3.12).

**Abbildung 3.11:** Dieses Mosaikbild einer 3D-Szene wurde aus einer Folge von 147 Bildern berechnet (Anhang A.2). Die blaue Trajektorie veranschaulicht wiederum die Kamerabewegungen, wobei insbesondere die Verzerrung des Mosaiks mit einer zunehmenden Entfernung der Kamera von ihrer initialen Position in der linken unteren Ecke auffällt.

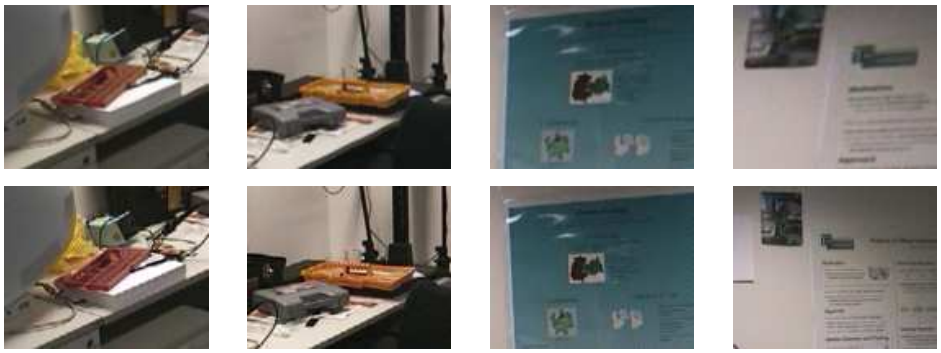


**Abbildung 3.12:** Oben sind einige ausgewählte Ausschnitte des Mosaikbildes aus Abb. 3.11 gezeigt, die im Vergleich zu den Originaldaten (unten) die gute Qualität des Mosaikbildes unterstreichen. Die Mosaikqualität nimmt allerdings bei großen Rotationswinkeln der Kamera deutlich ab (Bilder rechts).

Analog zu dem vorangegangenen Beispiel führt die wachsende Verzerrung des Mosaikbildes mit einer zunehmenden Kamerarotation jedoch auch hier zu einer deutlichen Unschärfe, wie insbesondere die beiden rechten der vergrößerten Ausschnitte zeigen.

Die beiden vorstehenden Beispiele illustrieren die hohe Qualität der Mosaikbilder, die sich mit den eingesetzten Verfahren erzielen lässt. Gleichzeitig wird aber auch deutlich, dass eine einzelne Ebene keine adäquate Basis zur Projektion von Bilddaten einer rotierenden Kamera darstellt. Insbesondere große Sichtwinkel können nicht adäquat widergespiegelt werden, so dass geeignetere Ansätze zur Repräsentation solcher Daten notwendig sind, wie sie im Rahmen dieser Arbeit entwickelt wurden (vgl. Kap. 6).

Bei der Berechnung der beiden zuvor diskutierten Mosaikbilder lagen der Parameterschätzung jeweils alle gegebenen Bildpunkte im jeweiligen Überlappungsbereich der zu registrierenden Bilder zu Grunde. Die beiden aufgenommenen Szenen enthalten allerdings auch keine bewegten Objekte und weisen in den meisten Teilbereichen eine hinreichende Struktur auf, um eine robuste Rekonstruktion der Homographien zu ermöglichen. In Szenen, in denen diese Voraussetzungen nicht gegeben sind, kann insbesondere die in Abschnitt 3.2.1 vorgestellte Durchführung der Parameterschätzung auf Basis einer ausgewählten Teilmenge aller verfügbaren Pixel hilfreich sein. Um die Auswirkungen der Pixelselektion im Hinblick auf unabhängige Bewegungen in einer Szene beurteilen zu können, wurde eine exemplarische Bildfolge mit einer statischen Kamera aufgenommen, in der ein bewegtes Objekt zunehmend größere Teile der Szene verdeckt (Abb. 3.13). Durch die statische Kamera ist die gesuchte Transformation zwischen den Bildern bekannt, so dass die geschätzten Homographien direkt mit der Identitätsabbildung verglichen werden können und sich ihre Güte damit auch quantitativ beurteilen lässt.

In den Diagrammen der Abbildung 3.14 sind die Ergebnisse dieser Untersuchung gezeigt. Links sind die geometrischen Fehler (Gl. 3.5) der jeweils geschätzten Homographien ohne und mit Ausblendung der bewegten Pixel relativ zur Identität gezeigt. Während ohne eine Beschränkung der Schätzung auf statische Pixel schon nach wenigen Bildern signifikante Differenzen zwischen der geschätzten Homographie und der vorliegenden Identitätsabbildung resultieren (und die Schätzung schließlich aufgrund vollständig degenerierter Parameter manuell abgebrochen wurde), sind die Ergebnisse mit einer Pi-

(a) Bild 1 (b) Bild 12 (c) Bild 17 (d) Bild 30 (e) Bild 42

**Abbildung 3.13:** Fünf Bilder aus einer Sequenz mit insgesamt 42 Bildern, die eine Szene mit einem dominierenden, bewegten Objekt zeigt und mit einer statischen Kamera aufgenommen wurde.

xelselektion deutlich stabiler. Zwar steigt auch hier der Schätzfehler über die Zeit an, selbst bei einem hohen Überdeckungsgrad der Szene (von fast 50%) können jedoch noch verhältnismäßig robust Parameter berechnet werden. Dabei ist insbesondere zu berücksichtigen, dass in der betrachteten Sequenz durch das bewegte Objekt auch große Teile der für die Parameterschätzung unerlässlichen Szenenstruktur verdeckt werden.

Die mit bewegten Pixeln in einer Szene bei einer Parameterschätzung verbundenen Probleme zeigen sich auch in der Anzahl von Iterationen (Abb. 3.14, rechts), die das Verfahren jeweils durchläuft. Ohne eine Selektion von Pixeln werden deutlich mehr Iterationen benötigt, um einen stabilen Optimierungszustand zu erreichen, so dass die Pixelselektion hier insgesamt schneller zu besseren Minima der Zielfunktion führt.

Zur Analyse der Wirksamkeit einer Selektion von Pixeln auf Basis lokaler Gradientenbeträge dienten verschiedene Bildfolgen, in denen großflächige, homogene Regionen vorhanden sind (Abb. 3.15(a)). Eine Pixelselektion in derartigen Bildern beschränkt die Daten zur Schätzung der Parameter im Wesentlichen auf Bildpunkte an signifikanten Kanten, wie die exemplarische Selektionskarte in Abb. 3.15(b) zeigt. Bei der Schätzung selbst hat sich durch diese Selektion in verschiedenen Testläufen allerdings kein Qualitätsgewinn gegenüber einer Schätzung ohne Pixelselektion gezeigt. Der zu beobachtende Registrierungsfehler ist zumeist nahezu gleich, wie das Diagramm rechts in Abb. 3.15 beispielhaft veranschaulicht. Allerdings wurden die Schätzungen durch die reduzierte Datenbasis teilweise um bis zu 20% beschleunigt.

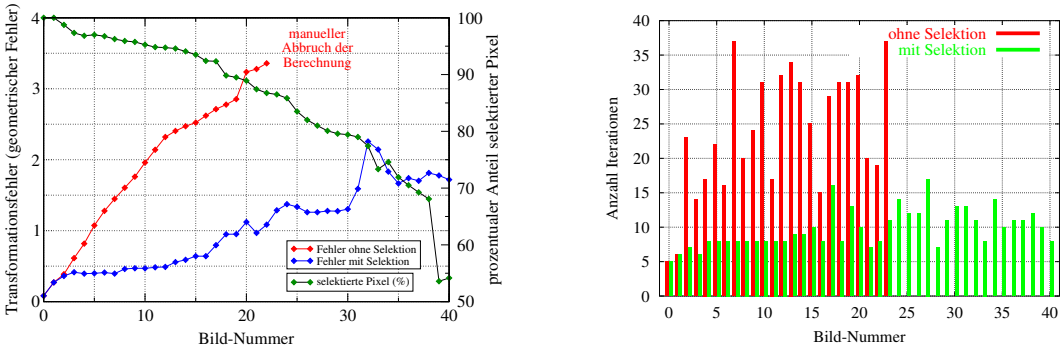
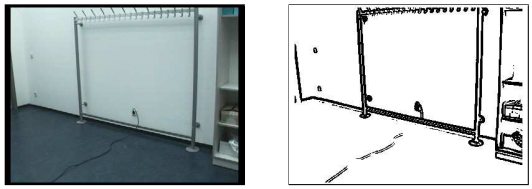



**Abbildung 3.14:** Vergleich der Güte geschätzter Homographien innerhalb einer Bildsequenz mit einem unabhängig bewegten Objekt, das bis zu 50% der Szene verdeckt (grüne Kurve links): Ohne eine Selektion statischer Pixel resultieren deutlich größere Abweichungen der geschätzten Transformationen von der Identität, wobei zusätzlich auch mehr Iterationen bei der Optimierung durchgeführt werden.



(a) Beispielbild (b) Selektionskarte

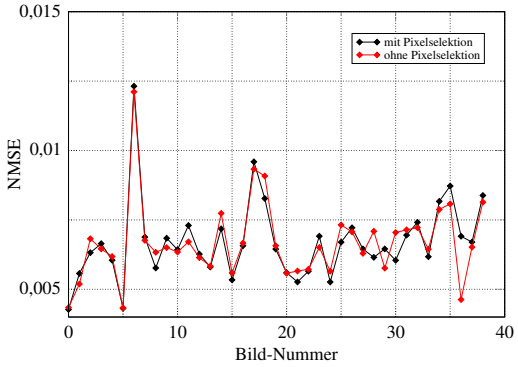



**Abbildung 3.15:** Untersuchung der Auswirkungen einer struktur-basierten Pixelselektion: Der NMSE bei der Registrierung einer exemplarischen Bildfolge (s. Beispielbild und zugehörige Karte mit selektierten Pixeln links) weist mit und ohne Pixelselektion kaum Unterschiede auf.

Die bislang diskutierten Beispiele zeigen eine im Allgemeinen gute Qualität der Transformationsparameter, die mit Hilfe des *Projective Flow* geschätzt werden können. In allen Bildsequenzen waren die Kamerabewegungen dabei hinreichend klein ($\approx 1-2°$), um eine robuste Parameterschätzung im Rahmen der Auflösungspyramide gewährleisten zu können (vgl. S. 32). Derart kleine Kamerabewegungen haben allerdings den Nachteil, dass zur großflächigen Aufnahme einer Szene eine hohe Anzahl an Bildern erforderlich ist und die Aufnahme damit unter Umständen sehr lange dauern kann. Aus diesem Grund ist eine Unterstützung von größeren Kamerabewegungen wünschenswert, die in dieser Arbeit durch eine explizite Initialisierung der Parameterschätzung auf Basis von Bewegungsdaten der Kamera realisiert wurde (Abschnitt 3.2.2).

Zur Auswertung dieses Ansatzes dienten verschiedene Bildsequenzen, die unter Verwendung der „EVI“-Kamera mit Hilfe von horizontalen bzw. vertikalen Kameraschwenks oder schrittweisen Veränderungen des Zooms aufgenommen wurden (exemplarische Bilder einer solchen Folge sind in Abb. 3.16 zu sehen). Anschließend erfolgte für jedes Bild dieser Sequenzen eine Parameterschätzung relativ zu den jeweils ersten Bildern der entsprechenden Sequenz, die als Referenz zu Grunde gelegt wurden.

Exemplarische Ergebnisse dieser Analysen für ausgewählte Bildfolgen sind in Abbildung 3.17 dargestellt. In den beiden Diagrammen sind jeweils die normierten, quadratischen Fehler (NMSE) zwischen den einzelnen Bildern der Sequenzen und dem Referenzbild nach der Registrierung mit und ohne eine explizite Initialisierung des Optimierungsprozesses aufgetragen. Obgleich die ermittelten Fehlerwerte nicht als absolute

(a) $\Delta\alpha = 0°$ (b) $\Delta\alpha \approx 10°$ (c) $\Delta\alpha \approx 17°$ (d) $\Delta\alpha \approx 25°$ (e) $\Delta\alpha \approx 30°$

**Abbildung 3.16:** Fünf Bilder einer Beispielsequenz zur Evaluation der expliziten Parameterinitialisierung. Die Bildfolge wurde mit einem horizontalen Kameraschwenk aufgenommen, so dass die Bilder eine zunehmende Winkeldifferenz $\Delta\alpha$ zum Referenzbild der Bildfolge (ganz links) aufweisen.

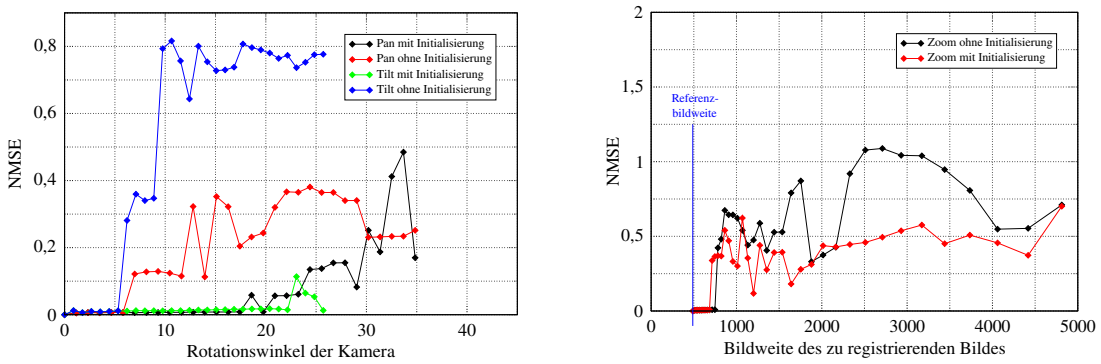



**Abbildung 3.17:** Evaluation der Parameterschätzung bei großen Differenzen zwischen zwei Bildern mit und ohne explizite Parameterinitialisierung aus Hardwareparametern der Kamera: Links sind die Fehler in der Registrierung der Bilder im Verhältnis zum jeweiligen Rotationswinkel der Kamera dargestellt, rechts ist die Fehlerentwicklung bei Bildweitenänderungen veranschaulicht.

Gütemaße zur Bewertung der Parameterqualität interpretiert werden sollten (vgl. Abschnitt 3.5.1), lassen sich aus den Diagrammen dennoch wichtige Schlussfolgerungen ziehen und grundsätzliche Tendenzen ableiten.

Bei rotatorischen Kamerabewegungen zeigen sich deutlich die Vorteile der Parameterinitialisierung. Zunächst sind bei kleinen Rotationen weder in der Güte der Parameter noch in der Performanz der Schätzung insgesamt (d.h. auch der Anzahl notwendiger Iterationen) signifikante Differenzen zwischen beiden Vorgehensweisen zu beobachten. Übersteigen die Rotationswinkel jedoch etwa 5°, so ist eine Parameterschätzung ohne explizite Initialisierung kaum mehr möglich. Im Gegensatz dazu gelingt mit der Initialisierung auch noch eine Registrierung von Bildern, zwischen denen Winkel von bis zu 20° vorliegen können.[8] Die beschränkende Größe der Parameterschätzung bei derart großen Rotationen zwischen den Bildern scheint dabei auch vorrangig durch den stetig abnehmenden Bildüberlapp gegeben zu sein und weniger durch die Initialisierung selbst. Während die aus den Hardwareparametern der Kameras ableitbaren, initialen Homographien offensichtlich auch bei großen Winkeln um 20° noch eine hinreichende Güte aufzuweisen scheinen, liegt bei diesen Konstellationen nur noch ein geschätzter Bildüberlapp von knapp 40% vor (vgl. Abb. 3.16), der eine robuste Schätzung offenkundig deutlich erschwert (vgl. hierzu auch das Konzept einer erweiterten Initialisierung der Schätzung auf Basis der Multi-Mosaikbilder in Abschnitt 6.4.3).

Während die Parameterinitialisierung insgesamt bei Kamerarotationen deutliche Vorteile im Hinblick auf zulässige Kamerabewegungen bietet, zeigt sich bei Veränderungen innerhalb der Bildweite ein anderes Bild (Abb. 3.17, rechts). Die ohne eine explizite Initialisierung der Schätzung schon bei verhältnismäßig geringen Differenzen in den Bildweiten zu beobachtenden, großen Registrierungsfehler werden auch mit einer Initialisierung nicht vermindert. Dies lässt einerseits auf eine stärkere Sensitivität der Parameterschätzung hinsichtlich von Differenzen in den Bildweiten der Bilder schließen, deutet andererseits

[8] Die Bildfolgen hier wurden mit minimaler Bildweite aufgenommen. Bei größeren Bildweiten sind mit den ermittelten Winkeln deutlich größere Veränderungen zwischen aufeinander folgenden Bildern verbunden, so dass die maximal zulässigen Winkel dort entsprechend kleiner zu wählen sind.

aber auch auf möglicherweise zu ungenaue Schätzwerte für die Bildweiten in diesem Anwendungskontext hin (vgl. Abschnitt 2.4.1). Zusätzlich bedingen große Differenzen in den Bildweiten aber auch grundsätzlich signifikante Skalierungen der Bilder, die eine Registrierung erschweren (vgl. auch Abschnitt 6.5.1). Abschließend kann damit gefolgert werden (vgl. Abb. 3.17), dass zur Mosaikbildberechnung bei Verwendung von Kameras, die ein Auslesen ihrer extrinsischen Bewegungsparameter erlauben, unter einer expliziten Initialisierung Rotationen von bis zu 20° durchaus zulässig sind. Veränderungen in den Bildweiten sollten dagegen 250 Pixel nicht übersteigen.

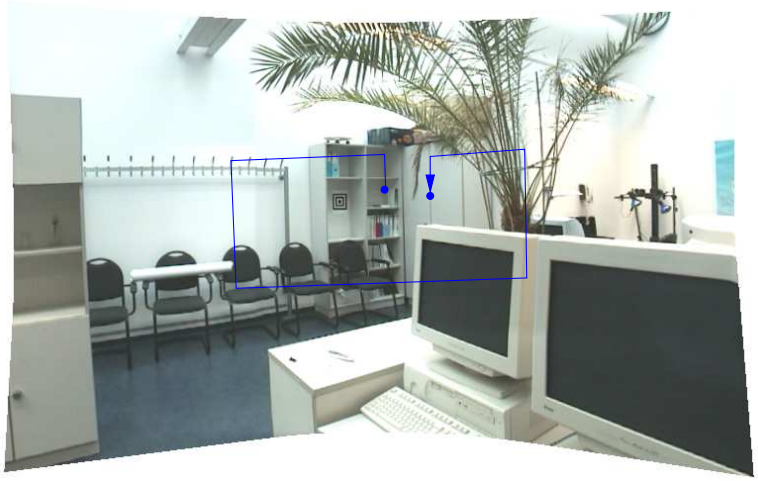


**Abbildung 3.18:** Mosaikbild der Bildsequenz aus Abb. A.3 im Anhang, berechnet im Frame-to-Mosaik-Modus und bei Tageslicht. Die blaue Trajektorie kennzeichnet grob die Mittelpunkte der Einzelbilder.

Eine Online-Langzeitberechnung von Mosaikbildern ist gegenüber einer Offline-Generierung vorrangig mit einer Akkumulation von Registrierungsfehlern verbunden. In interaktiven Systemen ist eine solche Herangehenweise dennoch zumeist unerlässlich, da beschränkte Ressourcen vielfach eine Schätzung globaler Parameter auf Basis vollständiger Bildsequenzen verhindern, und dies im Allgemeinen auch mit einer kontinuierlichen Verarbeitung von Bilddaten nicht zu vereinbaren ist.

Als Ansatz zur Verminderung des Einflusses der im Online-Modus unvermeidlichen Registrierungsfehler wurde in dieser Arbeit der in Abschnitt 3.3.1 skizzierte Frame-to-Mosaic-Modus implementiert. Dabei wird eine Schätzung aktueller Parameter jeweils relativ zum bislang berechneten Mosaikbild durchgeführt, so dass nicht allein das Vorgängerbild als Referenz zur Registrierung eines neuen Bildes herangezogen wird. Die Vorteile, die aus einer solchen Vorgehensweise resultieren und sich insbesondere bei Kamerabewegungen zeigen, bei denen einzelne Szenenausschnitte wiederholt zu unterschiedlichen Zeitpunkten aufgenommen werden, sind anhand einer Beispielbildfolge mit 98 Bildern (Abb. A.3) exemplarisch veranschaulicht. Das unter Anwendung des Frame-to-Mosaic-Modus aus dieser Bildfolge resultierende Mosaikbild ist in Abbildung 3.18 gezeigt, während in Abbildung 3.19 einzelne Ausschnitte des Mosaikbildes im Verlauf seiner Berechnung zu sehen sind. Sie sind zum Vergleich den korrespondierenden Ausschnitten aus einer Berechnung im Frame-to-Frame-Modus (S. 40) gegenübergestellt.

Die Resultate zeigen deutlich die Vorteile einer Schätzung relativ zum Mosaikbild auf. Während Registrierungsfehler zwar nicht in allen Fällen vollständig vermieden werden können (s. z.B. das obere Beispiel in Abb. 3.19), so weisen die linken Bildausschnitte des Frame-to-Mosaic-Modus im Vergleich zu denen des Frame-to-Frame-Modus dennoch eine deutlich höhere Qualität auf. Insbesondere in der unteren Zeile der Abbildung wurde das neue Bild im Frame-to-Frame-Modus mit einem deutlichen Versatz integriert, während der Übergang zwischen den Bildern im Mosaikbild des Frame-to-Mosaic-Modus kaum zu erkennen ist. Der Modus hat sich damit zur Reduktion des Einflusses von Registrierungsfehlern über die Zeit als geeignet herausgestellt und führt, insbesondere in Bildfolgen, die mehrere mit einem größeren zeitlichen Versatz aufgenommene Bilder einzelner Szenenausschnitte beinhalten, zu einer deutlich höheren Qualität des Mosaiks.

**Abbildung 3.19:** Vergleich der Mosaikqualität bei einer Online-Berechnung im Frame-to-Mosaic-Modus (links) und im Frame-to-Frame-Modus (rechts).

Zusammenfassend zeigen die Untersuchungen zur erzielbaren Qualität der Parametersätze unter Verwendung des *Projective Flow*-Algorithmus und der zusätzlich realisierten Ergänzungen (Pixelselektion, explizite Initialisierung, Frame-to-Mosaic-Modus) die Eignung dieses Ansatzes zur Berechnung qualitativ hochwertiger Mosaikbilder auf. Obgleich die Schätzung aufgrund der durch den implizierten Anwendungskontext der interaktiven Systeme vorgegebenen Rahmenbedingungen online durchgeführt wird, unterscheiden sich die berechneten Bilder oftmals kaum von den Originalbildern der verarbeiteten Sequenzen. Zwar dürfen keine absolut fehlerfreien Mosaikbilder erwartet werden, mit dem Frame-to-Mosaic-Modus konnte jedoch ein Ansatz gefunden werden, der eine zufriedenstellende Reduktion von sich akkumulierenden Schätzfehlern verspricht. Eine weitere Verbesserung der Bildqualität lässt sich zwar, wie eingangs diskutiert (Abschnitt 3.3), etwa durch eine simultane Verarbeitung mehrerer Bilder erzielen, derartige Ansätze sind jedoch mit einem höheren (Zeit-)Aufwand verknüpft und daher nicht notwendigerweise eine adäquate Alternative in interaktiven Systemen. Einzig eine parallelisierte Verarbeitung könnte Perspektiven in dieser Hinsicht aufzeigen, wobei jedoch die in den ins Auge gefassten Systemen verfügbaren Kapazitäten nicht außer Acht gelassen werden dürfen.

In den durchgeführten Experimenten ist eine verminderte Qualität der Mosaikbilder oftmals auch mit Unschärfen verbunden, die aus signifikanten geometrischen Verzerrungen der Mosaikbilder bei großen Kamerarotationen resultieren. Dies zeigt, dass die Wahl einer geeigneten Projektionsbasis für die Bilddaten einen signifikanten Einfluss auf die spätere Bildqualität ausübt. Eine einzelne Bildebene, die allen Beispielen zu Grunde lag, ist dabei zur Darstellung von Bilddaten rotierender Kameras offenkundig ungeeig-

net. Abhilfe verspricht diesbezüglich jedoch das Konzept der Multi-Mosaikbilder, das explizit zur Repräsentation von Daten rotierender Kameras entwickelt wurde (Kap. 6).

Im Hinblick auf die Robustheit der Parameterschätzung bei großen Rotationen konnten die Vorteile einer expliziten Initialisierung der Schätzung aus Kameradaten aufgezeigt werden. Demgegenüber hat sich eine gezielte Selektion geeigneter Pixel für die Parameterschätzung vorrangig bei unabhängigen Bewegungen in einer Szene als vorteilhaft erwiesen. Hinsichtlich einer Pixelselektion auf der Grundlage lokaler Bildstrukturen ist primär der Geschwindigkeitsgewinn bedeutsam, während ein signifikanter Einfluss auf die Parameterqualität nicht zu beobachten war.

Neben der Qualität der berechneten Mosaikbilder, die für die angestrebte Weiterverarbeitung der Mosaikbilder von hoher Bedeutung ist, spielt auch der Berechnungsaufwand der Registrierung eine entscheidende Rolle. Das angestrebte Anwendungsfeld der interaktiven Systeme erfordert eine möglichst schritthaltende Verarbeitung aufgenommener Bilddaten, die die Interaktivität der Systeme nicht beeinträchtigt und keine zeitlichen Verzögerungen verursacht. In der Literatur wird zur Echtzeitberechnung von Mosaikbildern oftmals spezielle Hardware eingesetzt [Bur94, Han94], während sich rein softwarebasierte Ansätze nur vereinzelt finden lassen [Jet98]. Zudem werden dort nur in Ausnahmefällen vollständige Homographien geschätzt [BE98, Rob03b]. Die Implementierung der Algorithmen in der vorliegenden Arbeit erfolgte auf einem handelsüblichen PC[9], wobei das System derzeit für eine Verarbeitung von Grauwertbildern optimiert ist. Zur Bewertung der erzielbaren Verarbeitungsraten bei der Parameterschätzung wurden verschiedene Bildsequenzen (insgesamt 740 Bilder der Größe $384 \times 288$) statischer Szenen analysiert, die jeweils im Frame-to-Mosaic-Modus ohne Selektion und Initialisierung verarbeitet wurden. Zur detaillierteren Aufschlüsselung der Laufzeiten wurden dabei die Zeiten in den einzelnen Ebenen der Auflösungspyramide gesondert gemessen. Die Ergebnisse sind in der nachfolgenden Tabelle zusammengefasst:

| Bildgröße $384 \times 288$ | Ebene 3 | Ebene 2 | Ebene 1 | Ebene 0 | gesamt[10] |
|---|---|---|---|---|---|
| mittlere Laufzeit $[ms]$ | 16,2 | 44,0 | 147,7 | 706,9 | 1015,1 |
| Standardabweichung | 3,0 | 11,4 | 28,1 | 225,1 | 229,5 |

Es zeigt sich, dass die derzeitige Implementierung eine Registrierung von Grauwertbildern etwa im Sekundentakt zulässt. Dabei ist jedoch zu berücksichtigen, dass die den Messungen zu Grunde liegende Bildgröße eine hohe Qualität bedingt. In Abhängigkeit vom jeweiligen Anwendungskontext kann es jedoch genügen, kleinere Bilder aufzunehmen. Da sich der Aufwand zur Registrierung von Bildern annähernd proportional zu ihrer Größe verhält, sind in diesen Fällen deutliche Geschwindigkeitsgewinne zu erwarten. Abgesehen davon ist die Bildregistrierung allerdings auch nur ein Teil innerhalb des gesamten Systems, dessen Performanz damit auch nur unter Berücksichtigung aller Teile abgeschätzt werden kann. In Abschnitt 6.5.3 folgt daher eine vollständige Evaluation.

[9] Rechnerkonfiguration: Prozessor AMD Athlon$^{TM}$ XP 1800+, 1 GB Hauptspeicher.

[10] Laufzeit über alle Ebenen inkl. der einmaligen Berechnung der Auflösungspyramiden beider Bilder.

# 4 Bildintegration

Die Registrierung der Einzelbilder einer gegebenen Bildfolge legt den Grundstein zur Etablierung von Korrespondenzen zwischen den Bildern innerhalb der Sequenz und damit auch für eine Eliminierung redundanter Informationen, wie sie den Ausgangspunkt zur Berechnung von Mosaikbildern darstellt. Als Ergebnis der Registrierung resultieren für jedes Bild der Folge Parameter eines zu Grunde gelegten Bewegungsmodells, mit dessen Hilfe die Bilder anschließend in das gemeinsame Mosaikkoordinatensystem transformiert werden können. Die Transformation bildet die Grundlage der eigentlichen Datenreduktion, die durch die *Integration* der Bilder in das Mosaikbild erfolgt. Dabei werden die Daten der Einzelbilder fusioniert, so dass redundante Bereiche miteinander verschmelzen und im finalen Mosaikbild nur noch einfach auftreten. Das Volumen der Eingangsdaten verringert sich auf diese Weise in der Regel erheblich.

Bei der Integration werden die Farb- bzw. Intensitätswerte der einzelnen Pixel des Mosaikbildes aus den Daten der einzelnen Bilder der Sequenz berechnet. Prinzipiell ist die Grundlage zur Festlegung des Wertes eines Mosaikpixels dabei durch alle Bildpunkte aus den einzelnen Bildern gegeben, die unter der Transformation auf das entsprechende Mosaikpixel abgebildet werden. Sie lassen sich unter Anwendung einer geeigneten Integrationsheuristik zu einem Wert zusammenfassen, so dass ein bezüglich der Eingangsbilder möglichst „repräsentatives“ Mosaikbild entsteht. Die Wahl der Heuristik wird dabei durch verschiedene Faktoren beeinflusst. Einerseits sind nicht alle Verfahren für Online- und Offline-Ansätze gleichermaßen geeignet, so dass bereits die grundlegende Verarbeitungsweise der Bildfolgen die möglichen Vorgehensweisen eingrenzt. Darüber hinaus gibt auch die angestrebte Verwendung der Mosaikbilder Rahmenbedingungen vor.

So muss bei der Wahl eines Integrationsverfahrens beispielsweise explizit berücksichtigt werden, ob in dynamischen Umgebungen nur der statische Szenenhintergrund oder auch unabhängig bewegte Objekte im Mosaikbild repräsentiert werden sollen. Die gewählte Integrationsheuristik hat damit direkten Einfluss auf die Qualität des Mosaikbildes. Dieser Faktor ist insbesondere bei einer *kontinuierlichen* Registrierung und Integration von Bilddaten bedeutsam. In diesem Kapitel werden zunächst grundlegende Verfahren zur Integration von Bildern vorgestellt (Unterkap. 4.1). Vor dem Hintergrund des in dieser Arbeit präsentierten visuellen Speichers geht Unterkapitel 4.2 dann gezielt auf Ansätze ein, die sich auch für den Einsatz in Online-Verfahren eignen. Die in der Implementierung verwendeten Heuristiken werden abschließend in Abschnitt 4.3 diskutiert.

## 4.1 Grundlegende Ansätze zur Integration

Bei der Integration von Bildern in ein Mosaikbild besteht das Ziel in der Fusion der Farbwerte der einzelnen Bilder, so dass das Mosaikbild die Daten der Bildfolge in geeigneter Weise widerspiegelt. Die dabei zu Grunde liegende Integrationsheuristik ist formal durch eine Funktion $m$ gegeben, die jedem Pixel $u$ des Mosaikbildes $M$ in Abhängigkeit von den zur Verfügung stehenden $N$ Bildern $I_i', i = 1 \ldots N$, einen Farbwert zuordnet.[1]

Grundsätzlich lassen sich bei der Festsetzung von $m$ zwei verschiedene Herangehensweisen unterscheiden. Einerseits können für die Berechnung eines Mosaikpixels *alle* Bilder einer Folge gleichermaßen berücksichtigt werden, während sich andererseits auch nur jeweils ein einzelnes Bild als Datengrundlage auswählen lässt. Im ersten Fall ist $m$ durch eine Funktion gegeben, die die $N$ Farbwerte zu einem einzelnen Wert zusammenfasst. Dies geschieht oftmals beispielsweise durch Mittelwertbildung, wobei der Einfluss einzelner Pixel über individuelle Gewichte $\gamma_i$ gezielt gesteuert werden kann:

$$M(u) = m(I_1'(u), \ldots, I_N'(u)) = \frac{1}{\sum \gamma_i} \sum_{i=1}^{N} \gamma_i I_i'(u). \tag{4.1}$$

Die $\gamma_i$ lassen sich unter anderem aus den Abständen der einzelnen Pixel zu den Bildzentren oder aus Bewegungsdaten motivieren [Ira96]. Dabei weisen im Allgemeinen mehrere Gewichte einen Wert größer als 0 auf. Setzt man demgegenüber nur genau ein Gewicht auf den Wert 1 und allen anderen auf 0, so führt dies zu Verfahren der zweiten Kategorie. Jedem Mosaikpixel liegt dann genau ein Eingangsbild zu Grunde und $m$ ist durch eine *Projektionsfunktion* gegeben, die auf Basis einer *Selektionsfunktion* $s$ für jedes Pixel $u$ genau dieses eine Bild als Datenquelle auswählt (vgl. auch [Még99]):

$$M(u) = m_{s(u)}(I_1'(u), \ldots, I_N'(u)) = I_{s(u)}'(u). \tag{4.2}$$

Da durch $s$ für Gruppen benachbarter Pixel oftmals dasselbe transformierte Bild als Datenbasis selektiert wird, resultiert aus dieser Vorgehensweise vielfach eine Segmentierung des Mosaikbildes in zusammenhängende *Regionen*. Dabei ist der Abstand der Mosaikpixel zu den Mittelpunkten der transformierten Bilder ein gängiges Kriterium zur Selektion. Ihm liegt die Annahme zu Grunde, dass der Einfluss von Linsenverzerrungen in Bildern mit steigendem Abstand vom Bildmittelpunkt zunimmt[2] und sich daher im Mosaikbild minimieren lässt, wenn die Bilddaten möglichst aus zentrumsnahen Bildregionen stammen. Das Ergebnis ist in diesem Fall eine Voronoi-Tesselierung des Mosaikbildes [Pel97], wobei die Zentren der einzelnen Zellen durch die Mittelpunkte der transformierten Bilder gegeben sind. Alternativ können auch die Aufnahmezeitpunkte

[1] Im Folgenden wird angenommen, dass alle Bilder $I_i$ bereits unter Anwendung des Bewegungsmodells $T_{\vec{p}}$ transformiert wurden und im Koordinatensystem des Mosaikbildes vorliegen. Die zu einem Mosaikpixel $u$ korrespondierenden Pixel in diesen Bildern $I_i' = T_{\vec{p}}(I_i)$ weisen damit identische Koordinaten auf und ihre Farbwerte sind direkt durch $I_1'(u), \ldots, I_N'(u)$ gegeben.

[2] Hierbei wird davon ausgegangen, dass der Hauptpunkt der Kamera im Bildzentrum liegt (Kap. 2).

der Einzelbilder für eine Selektion herangezogen werden, wobei jeweils das Bild als Informationsquelle ausgewählt wird, welches die aktuellsten Daten bereithält (Abb. 4.1(b)).

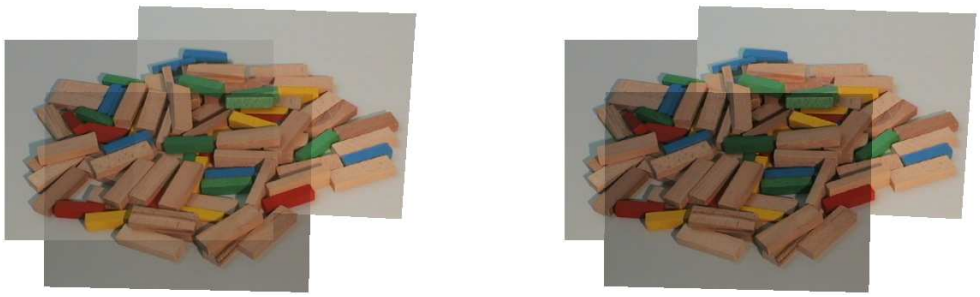
Grundsätzlich ist bei einer Integration von Bildern, unabhängig von der angewandten Heuristik, zu berücksichtigen, dass zwischen einzelnen Bildern einer Folge auch nach einer erfolgreichen Registrierung signifikante Unterschiede bestehen können. Solche Differenzen sind in Bildfolgen statischer Szenen vorrangig auf Änderungen der Bildenergien innerhalb der Folge zurückzuführen, wie sie beispielsweise durch eine automatische Adaption der Kameraeinstellungen während der Aufnahme (Blendenjustierung, o.ä.) oder Änderungen in den Beleuchtungsverhältnissen (z.B. wechselnde Sonneneinstrahlung oder Schattenwurf) hervorgerufen werden (Abb. 4.1). Verbliebene Unterschiede zwischen den Bildern von Folgen dynamischer Szenen lassen sich oftmals durch unabhängig bewegte Objekte erklären (Kap. 5), deren Positionen und Sichtbarkeiten innerhalb der registrierten Bilder variieren (z.B. Abb. 5.1(b)). Beide Klassen von Differenzen verursachen bei der Integration (zumeist unerwünschte) Artefakte im Mosaikbild, die im Allgemeinen durch geeignete Algorithmen explizit korrigiert werden (Abschnitte 4.1.1 und 4.1.2).

### 4.1.1 Farbadaption

Größere Abweichungen in den Farbwerten korrespondierender Pixel in Bildfolgen statischer Szenen, die während einer Registrierung nicht ausgeglichen werden, sind im Allgemeinen durch Änderungen in den Bildenergien bedingt. Sie führen in pixelbasierten Integrationsansätzen als Folge der Mittelung zu großflächigen Farbverfälschungen und sichtbaren Unstetigkeiten entlang der Bildränder (Abb. 4.1(a)). Bei Heuristiken, die in einer Regionensegmentierung des Mosaiks münden, werden die Pixelwerte im Mosaikbild gegenüber denen der Eingangsbilder zwar weniger gravierend verfälscht, Diskontinuitäten an den Übergängen zwischen einzelnen Regionen sind jedoch deutlich stärker ausgeprägt (Abb. 4.1(b)). Beide Arten von Differenzen können durch geeignete Algorithmen ausgeglichen werden, um die Qualität der resultierenden Mosaikbilder zu erhöhen. Der nachfolgende Abschnitt enthält eine Übersicht einiger exemplarischer Ansätze.

Im Rahmen einer gewichteten Mittelung über alle Bilder einer Folge kann eine Korrektur von Farbdifferenzen durch eine Angleichung der Farbwerte korrespondierender Pixel realisiert werden. Dazu ist einerseits eine gezielte Adaption der einzelnen Gewichte $\gamma_i$ bei der Fusion der Bildpunkte denkbar (s. Gl. 4.1). Gümüştekin führt alternativ eine Histogrammnormierung der beteiligten Bilder durch, gekoppelt an die Suche eines optimalen Pfades entlang minimaler Diskontinuitäten [Güm96]. In [Uyt01] schließlich wird ein Ansatz zum Ausgleich von Helligkeitsschwankungen vorgeschlagen, bei dem vor der Fusion zunächst eine Anpassung der Einzelbilder an ihre lokale Nachbarschaft über quadratische Interpolationsfunktionen stattfindet.

In regionenbasierten Ansätzen erfolgt der Ausgleich von Differenzen im Allgemeinen durch eine lokale Überblendung benachbarter Bilder in Grenzbereichen. Dazu wird dort bei der Festlegung von Werten für einzelne Mosaikpixel keine harte Entscheidung zu Gunsten je eines Bildes getroffen, sondern die Bilder werden durch eine gewichtete Mittelung entlang der Regionengrenzen allmählich ineinander überführt. Die Grundlage dabei



(a) Integration von Bildern unter einer Mittelwertbildung.

(b) Regionenbasierte Integration der jeweils aktuellsten Daten.

**Abbildung 4.1:** Auswirkungen wechselnder Beleuchtungsverhältnisse in den Bildern einer Folge bei der Integration in ein Mosaikbild unter Anwendung verschiedener Heuristiken: bei einer Mittelwertbildung (a) sind großflächige Farbverfälschungen zu beobachten, während eine regionenbasierte Integration (b) vorrangig mit ausgeprägten, harten Übergängen an Regionengrenzen verbunden ist.

bilden z.B. lineare oder sigmoide Überblendungsfunktionen (Abb. 4.2), mit denen die Pixel der Bilder jeweils in Abhängigkeit ihres Abstandes von der Grenze gewichtet werden. Die Breite des Übergangsbereichs, in dem diese Überblendung stattfindet, ist dabei im Allgemeinen fest vorgegeben und nicht adaptiv (z.B. [Jet98]). Burt und Adelson [Bur83b] schlagen als Alternative einen flexibleren Algorithmus vor, der die spektralen Charakteristika der Bildfunktionen stärker berücksichtigt, dabei aber auch deutlich aufwändiger ist. Im Wesentlichen basiert das Verfahren auf einer Zerlegung der Bilder in einzelne Frequenzbänder, die getrennt voneinander mit optimalen Überlappungsbereichen fusioniert und dann wieder zu einem einzelnen Bild zusammengesetzt werden.

Grundsätzlich spielt bei der Auswahl eines Verfahrens zur Glättung von Übergängen der Berechnungsaufwand im Verhältnis zum erzielbaren Qualitätsgewinn eine entscheidende Rolle. Da die umschließenden Vierecke von transformierten Bildern zumeist weder rechteckig noch achsenparallel sind, lassen sich ihre Grenzen insbesondere bei einer simultanen Überlagerung von mehreren Bildern oftmals nur durch komplexe Algorithmen ermitteln, deren Aufwand den Qualitätsgewinn deutlich relativieren kann.

Exemplarische Blendfunktionen:

- linear:
  $f(x) = \frac{1}{d} \cdot x \ , \ d \in \mathbb{N}$
- sigmoid (Fermi-Funktion):
  $f(x) = \frac{1}{1+e^{-a \cdot (x-0.5 \cdot d)}} \ , \ a \in \mathbb{R}, d \in \mathbb{N}$

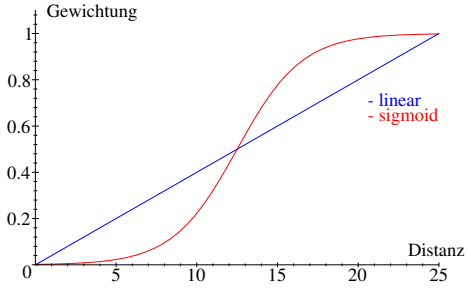


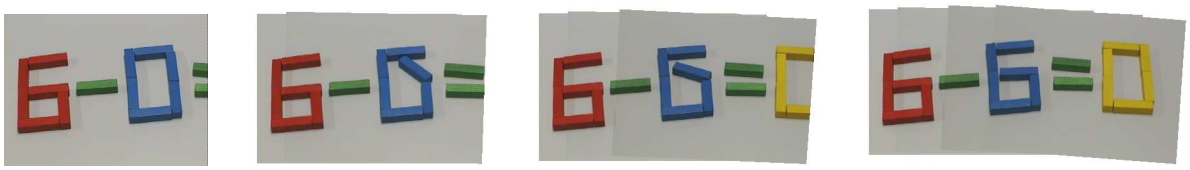

**Abbildung 4.2:** Lineare und sigmoide Gewichtungsfunktionen zum Überblenden von Bildern: auf der linken Seite sind die allgemeinen Definitionen angegeben, während rechts die Graphen der Funktionen für einen beispielhaften Überblendungsbereich der Breite 25 gezeigt werden.

### 4.1.2 Unabhängige Bewegungen

Bildfolgen dynamischer Szenen sind durch unabhängig bewegte Objekte charakterisiert, deren Bewegungen nicht vom globalen Bewegungsmodell der Kamera erfasst werden (vgl. auch Kap. 5). Die Abbildungen dieser Objekte in den einzelnen Bildern lassen sich nicht durch die geschätzten Transformationen ineinander überführen, so dass nach der Registrierung konkurrierende Positionen der Objekte in den transformierten Bildern verbleiben. Der Umgang mit derart inkonsistenten Bilddaten bei einer Integration in ein Mosaikbild hängt im Wesentlichen davon ab, welche Daten repräsentiert werden sollen. Irani und Kollegen [Ira96] unterscheiden dazu zwischen statischen und dynamischen Mosaikbildern. In statische Mosaiks werden alle Bilder einer Folge direkt integriert, ohne dynamische Änderungen in der Szene zu berücksichtigen. Das Mosaikbild entspricht damit einer zeitlichen Mittelung über der kompletten Bildfolge, wobei unabhängig bewegte Objekte verwischt erscheinen (Abb. 4.3(b)). Diese Artefakte sind jedoch insbesondere im Hinblick auf eine weitere Verarbeitung der Bilder oftmals störend.

Eine Berechnung dynamischer Mosaikbilder umgeht diese Problematik. Die Bilder bestehen im Grunde aus einer Folge von mehreren Mosaiks, die nur die jeweils aktuellsten Daten einer Szene bereitstellen. Unabhängig bewegte Objekte werden dabei ausschließlich durch ihre gegenwärtige Position repräsentiert, so dass Verwischungen und Unschärfen nicht auftreten. Zur Speicherung dieser Mosaikbilder erfolgt zumeist keine Sicherung aller Einzelbilder, sondern nur des initialen Mosaikbildes und explizit extrahierter, paarweiser Differenzen zwischen aufeinander folgenden Bildern im zeitlichen Verlauf. Jedes einzelne Mosaikbild der Sequenz kann auf dieser Basis später aus seinen Vorgängern rekonstruiert werden und den aktuellen Zustand der Szene *einschließlich* der jeweils aktuellen Positionen bewegter Objekte korrekt wiedergeben. Abbildung 4.3(a) zeigt exemplarisch verschiedene Mosaikbilder einer solchen Sequenz. In der aufgenommenen Szene ist eine aus Bausteinen gelegte Gleichung zu sehen, die anfangs einen Fehler enthält und erst im Verlauf der Bildaufnahme korrigiert wird. Die Mosaiksequenz repräsentiert diese schrittweisen Veränderungen in der Szene vollständig, wobei jeweils bewegte und unbewegte Objekte zeitgleich in den Bildern präsent sind.

Im Hinblick auf eine adäquate Repräsentation der *statischen* Anteile einer Szene als Ziel dieser Arbeit sind, im Gegensatz zu den vorstehenden Darstellungen, vorrangig statische „Momentaufnahmen“ einer Szene von Interesse. Dies bedeutet, dass bei dynamischen Veränderungen in einer Szene nicht alle Zwischenschritte repräsentiert werden sollen, sondern dass das Mosaikbild jeweils den aktuellen statischen Anteil einer Szene *vor* bzw. *nach* Abschluss der Modifikationen widerspiegelt. Daraus ergibt sich primär die Notwendigkeit einer expliziten Identifikation und Trennung statischer und dynamischer Szenenanteile. In die Mosaikbilder werden nur die statischen Anteile übernommen, während die dynamischen Informationen in einer separaten Datenstruktur abgelegt werden (Kap. 5.2). Abbildung 4.3(c) veranschaulicht ein solches Mosaikbild, das den Zustand der aufgenommenen Szene *vor* der Korrektur der Gleichung zeigt. Auf eine Aktualisierung der Hintergrundrepräsentation nach Abschluss der Modifikationen innerhalb der Szene wurde an dieser Stelle verzichtet, sie wird aber durch eine Analyse der extrahierten



(a) dynamisches Mosaikbild nach Irani [Ira96]

(b) statisches Mosaik

(c) explizite Ausblendung bewegter Bereiche

**Abbildung 4.3:** Exemplarischer Vergleich verschiedenener Integrationsheuristiken bei einer Verarbeitung von Bildfolgen mit unabhängig bewegten Objekten: In (a) ist ein dynamisches Mosaik gezeigt, das als Momentaufnahme der jeweils aktuellen Situation in der Szene aufgefasst werden kann. (b) veranschaulicht die Auswirkungen einer Mittelung, bei der bewegte Objekte verwischt erscheinen, während in (c) nur unbewegte Objekte integriert und die dynamischen Daten vollständig ausgeklammert werden.

dynamischen Daten prinzipiell unterstützt (vgl. Abschnitt 5.2.2 bzw. Abb. 5.5).

Im Allgemeinen werden zur Berechnung von statischen Mosaikbildern pixelbasierte Integrationsheuristiken verwendet, da sich durch sie die angestrebte, temporale Mittelung über die Bilder einer Folge leicht umsetzen lässt. Bei der Generierung von (dynamischen) Mosaikbildern, die bewegte Objekte jeweils durch eine ausgewählte Position repräsentieren, empfiehlt sich dagegen ein Einsatz von Integrationsheuristiken, die zu einer Zerlegung des Mosaikbildes in kompakte Regionen führen. Die Regionengrenzen werden dabei zumeist so gewählt, dass die Bilddaten für ein bewegtes Objekt jeweils nur einem Quellbild entstammen und sich inkonsistente Bilddaten somit weitgehend ausschließen lassen. In [Uyt01] wird eine solche Zerlegung aus Korrespondenzen zwischen bewegten Regionen in einzelnen Bildern abgeleitet. Davis [Dav98] schlägt alternativ eine Segmentierung entlang optimaler Pfade vor, die Diskontinuitäten minimieren und durch eine Suche in einer „Residuen-Landschaft“ ermittelt werden können. Die vollständige Ausblendung bewegter Objekte in Mosaikbildern lässt sich abschließend sowohl in pixel- wie auch in regionenbasierten Ansätzen durch eine explizite Maskierung von bewegten Bildpunkten bei der Integration realisieren (z.B. [Még99, Möl01a]).

## 4.2 Integration in Online-Verfahren

Bei der Auswahl einer geeigneten Heuristik zur Integration von Bilddaten in ein Mosaikbild ist neben der Charakteristik der zu repräsentierenden Daten (vgl. vorherigen Abschnitt) insbesondere die grundlegende Vorgehensweise bei der Berechnung der Mosaikbilder von entscheidender Bedeutung. In Bezug auf die vorliegende Arbeit sind somit vorrangig Integrationsheuristiken interessant, die sich in geeigneter Weise mit einer Online-Verarbeitung von Bilddaten kombinieren lassen.

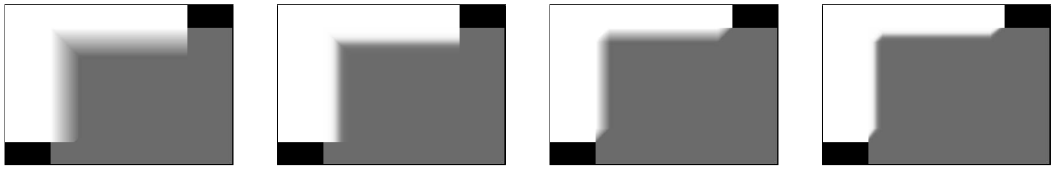
Grundsätzlich schränkt eine Online-Generierung von Mosaikbildern die Menge potenziell anwendbarer Heuristiken ein. Im Gegensatz zur Integration von Bildern in Offline-Ansätzen stehen als Grundlage zur Berechnung der Mosaikpixel nur das einzufügende Bild selbst und das bis zum aktuellen Zeitpunkt generierte Mosaikbild zur Verfügung. Heuristiken, die auf einer Auswertung von mehr als zwei Farbwerten beruhen (z.B. eine Medianberechnung), sind damit prinzipiell nicht einsetzbar. Neben diesem Aspekt gibt auch die angestrebte zeitliche Dauer einer Verarbeitung Randbedingungen für die Auswahl einer geeigneten Heuristik vor. Da eine Online-Berechnung von Mosaikbildern grundsätzlich dauerhaft erfolgen kann und nicht auf eine beschränkte Anzahl von Bildern festgelegt ist, sind Verfahren notwendig, die auch über größere Zeiträume hinweg eine hohe Qualität der Mosaikbilder gewährleisten. Aus diesem Grund eignet sich eine pixelbasierte Integration mit sukzessiver, gewichteter Mittelwertbildung im Allgemeinen nicht für eine Verwendung in diesem Anwendungskontext. Insbesondere bei (unvermeidlichen) Registrierungsfehlern im Verlauf der Berechnungen ist mit diesem Ansatz eine zunehmende Unschärfe der Mosaikbilder zu erwarten.

Regionenbasierte Algorithmen weisen in dieser Hinsicht deutliche Vorteile auf. Unschärfen und Verwischungen werden bereits durch die Grundidee der regionenbasierten Verfahren auf ein Minimum reduziert. Da für die Werte der einzelnen Pixel im Mosaikbild, bedingt durch die Selektionsfunktion $s$ (Gl. 4.2), nur jeweils ein einzelnes Bild als Grundlage dient, beschränken sich Inkonsistenzen im Wesentlichen auf die Übergangsbereiche zwischen den einzelnen Regionen. Allerdings ist dabei zu berücksichtigen, dass bei hinreichend kleinen Verschiebungen zwischen aufeinander folgenden Bildern und verhältnismäßig großen Überblendungszonen auch hier schnell grossflächige, unscharfe Regionen resultieren können, die langfristig eine verminderte Mosaikqualität bedingen. Die jeweils zuletzt integrierten Daten werden jedoch in jedem Fall in hoher Qualität integriert, so dass regionenbasierten Verfahren in dieser Hinsicht stets der Vorzug zu geben ist.

Ein wesentlicher Unterschied bei der Anwendung von regionbasierten Verfahren im Online-Modus gegenüber einem Einsatz in Offline-Algorithmen besteht darin, dass die Selektion des Ursprungsbildes für ein Mosaikpixel auf eine binäre Entscheidung zwischen dem Mosaik und dem neu zu integrierenden Bild beschränkt wird. Die Farbwerte des Mosaikbildes sind dadurch prinzipiell anfälliger gegenüber zwischenzeitlich auftretendem Bildrauschen. Darüber hinaus kann eine Anwendung regionenbasierter Verfahren im Online-Modus mit einem erhöhten Berechnungsaufwand verbunden sein. Beispielsweise erfordert die inkrementelle Voronoi-Tesselierung eines Mosaiks beim Einfügen eines neuen Bildes eine wiederholte Prüfung aller Pixel im Hinblick auf ihren Abstand zum Mittelpunkt des neuen Bildes und gegebenenfalls eine Aktualisierung der Farbwerte.

## 4.3 Auswertung & Diskussion

Die Anforderungen an die Integrationsheuristik für den ikonischen Speicher in dieser Arbeit ergeben sich primär aus der a priori zeitlich uneingeschränkten Online-Verarbeitung der Bildfolgen. Wie im vorherigen Abschnitt verdeutlicht wurde, bieten sich für diesen



(a) linear (b) sigmoid (c) linearN (d) sigmoidN

**Abbildung 4.4:** Vergleich verschiedener Überblendheuristiken bei der Bildintegration: Die linken beiden Mosaikbilder resultieren aus einer Anwendung linearer und sigmoider Überblendungsfunktionen. Die verbleibenden harten Kanten links unten und rechts oben in den Überblendungsbereichen lassen sich u.a. durch eine alternative Heuristik vermindern, die das Verhältnis valider Pixel in beiden Bildern zu Grunde legt (Abb. 4.4(c) und 4.4(d)), Details s. Text.

Ansatz vorrangig regionenbasierte Heuristiken an. Da mit der angestrebten Repräsentation ausschließlich statischer Szenenbereiche weiterhin eine explizite Ausblendung unabhängig bewegter Objekte verbunden ist, muss auch dies bei der Auswahl einbezogen werden. Unter Berücksichtigung dieser Aspekte scheint sich damit eine regionenbasierte Integration stets neuer Daten unter expliziter Ausblendung bewegter Bereiche am besten für eine Verwendung im Rahmen des visuellen Speichers zu eignen. Das Selektionskriterium der Ursprungsbilder für einzelne Mosaikpixel ist dabei durch Zeitstempel gegeben, so dass die Daten im Mosaikbild fortschreitend ersetzt werden, sobald neue Informationen in den entsprechenden Szenenbereichen zur Verfügung stehen.

Die mit regionenbasierten Integrationsheuristiken verbundenen harten Übergänge zwischen einzelnen Regionen werden durch geeignete Überblendungsfunktionen gemildert. Im Rahmen der Arbeit wurden dazu vier Ansätze erprobt, die anhand eines synthetisierten Beispiels in Abb. 4.4 veranschaulicht werden. Dort erfolgt eine Integration zweier einfarbiger Bilder mit unterschiedlichen Intensitäten. Das graue Bild enthält dabei die vermeintlich neueren Daten und dominiert bei der Integration (schwarze Bereiche markieren Gebiete außerhalb der Definitionsbereiche beider Bilder). In allen vier Mosaikbildern wird eine Überblendung der Bilder entlang der Ränder des neueren Bildes mit einer Breite von $d = 50$ Pixeln durchgeführt. Die Abbildungen 4.4(a) und 4.4(b) zeigen die Auswirkungen der linearen und sigmoiden Blendfunktionen, die in Abbildung 4.2 eingeführt wurden. Dabei ergibt sich die Distanz der einzelnen Pixel zur Regionengrenze allein aus dem senkrechten Abstand zum Rand des neu zu integrierenden Bildes und wird direkt analytisch über das umschliessende Viereck des neuen Bildes bestimmt. Da bei dieser Vorgehensweise in Integrationsgebieten, die nahe an den Rändern beider Bilder liegen, scharfe Kanten im Mosaikbild verbleiben (in den Abbildungen jeweils links unten bzw. rechts oben), wurde alternativ eine Bestimmung der Gewichte durch eine Analyse der umliegenden Nachbarschaften erprobt. Dazu wird in einer durch die Breite des Integrationsbereichs festgelegten, quadratischen Nachbarschaft eines jeden Bildpunktes das Verhältnis $\kappa$ valider Pixel im neuen und im alten Bildes bestimmt (Details s. Abb. 4.5).

Die daraus folgenden Gewichtungsfaktoren $\gamma_\kappa$ können anschließend direkt einer linearen Gewichtung zu Grunde gelegt („linearN“, Abb. 4.4(c)) oder noch zusätzlich sigmoid transformiert werden („sigmoidN“, Abb. 4.4(d)). In beiden Fällen ergeben sich in den zuvor skizzierten Problemzonen sanftere Übergänge zwischen beiden Bildern. Es ist aller-

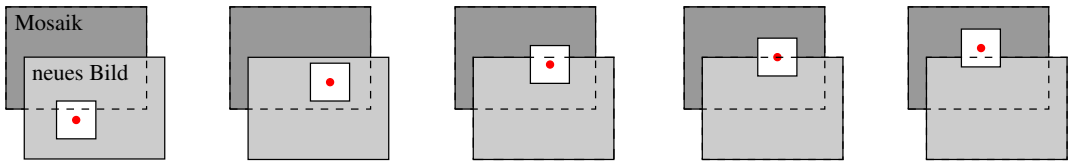
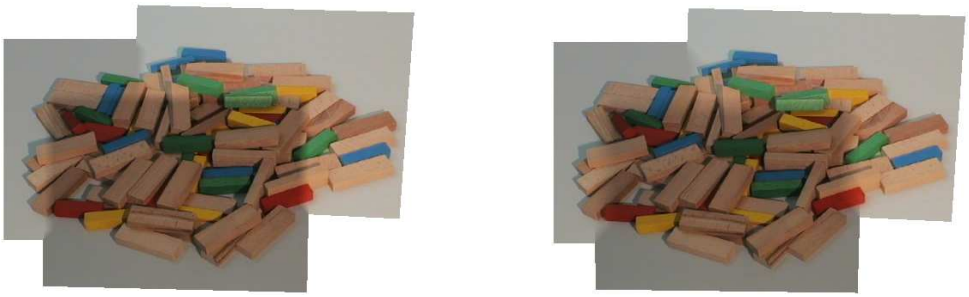



**Abbildung 4.5:** Direkte Bestimmung der lokalen Gewichte $\gamma_\kappa$ beim Überblenden zweier Bilder durch Auswertung des Verhältnisses $\kappa$ der validen Pixel in beiden Bildern innerhalb einer betrachteten Nachbarschaft (weißes Quadrat).

dings zu berücksichtigen, dass eine Betrachtung lokaler Nachbarschaften zur Bestimmung der Gewichte aufwändiger ist als eine analytische Distanzberechnung. Der Aufwand lässt sich zwar durch eine geschickte Zählung der Bildpunkte zumindest partiell reduzieren, die Effizienz analytischer Berechnungen ist aber auch damit nicht zu erreichen. In der praktischen Anwendung muss daher in Abhängigkeit vom Anwendungskontext zwischen den Vorteilen glatterer Übergänge und dem damit verbundenen Mehraufwand abgewogen werden. Abbildung 4.6 zeigt die Auswirkungen der linearen, nachbarschaftsbasierten Überblendung bei der Integration realer Bilder (vgl. auch Abb. 4.1).



(a) Überblendung mit einer Breite von $d = 25$ Pixeln.

(b) Überblendung mit einer Breite von $d = 50$ Pixeln.

**Abbildung 4.6:** Vergleich der Bildintegration unter Anwendung der Überblendungsheuristik „linearN“ (vgl. Abb. 4.4) mit unterschiedlichen Breiten $d$ für den Überblendungsbereich: Je größer $d$ gewählt wird, desto sanfter erscheinen die Übergänge zwischen den Regionen.

Die Ausblendung bewegter Objektpixel wird durch Anwendung einer Maske realisiert, die sich aus den Daten der Bewegungsdetektion des vorangegangenen Zeitschritts generieren lässt und bewegte Pixel (zur Detektion, s. Kap. 5) bei der Integration ausschließt. Da das Mosaikbild möglichst den jeweils aktuellen statischen Anteil einer Szene repräsentieren soll, ist es dabei wichtig, auch Änderungen der Bewegungscharakteristika von Objekten geeignet zu berücksichtigen. Beginnt oder endet beispielsweise die Bewegung eines Objektes innerhalb der Bildfolge (wie z.B. die des Bauklotzes in Abb. 4.3), so ist es wünschenswert, dass derartige Änderungen erkannt werden und eine Aktualisierung des statischen Szenenhintergrundes erfolgt. In Bezug auf das zuvor angeführte Beispiel würde der Bauklotz damit final von seiner ursprünglichen Position entfernt und in seiner

**Abbildung 4.7:** Auswirkungen einer Überblendung bei einer Integration fehlerhaft registrierter Bilder: Während ohne Überblendung (links) harte „Kanten“ im Mosaikbild hervorgerufen werden, verbessert die Überblendung den visuellen Eindruck deutlich (rechts).

neuen Position repräsentiert werden, ohne jedoch alle Zwischenschritte explizit darzustellen. Die der direkten Bewegungsdetektion nachgeschalteten Schritte zur Analyse der Bewegungscharakteristik von Objekten (Abschnitt 5.2.2) liefern hierfür entsprechende Daten, die in der Integration berücksichtigt werden können. Das Mosaikbild lässt sich damit fortwährend (mit wenigen Zeitschritten Versatz) aktualisieren und an die aktuellen Gegebenheiten in der repräsentierten Szene anpassen.

Abschließend sei darauf verwiesen, dass durch eine regionenbasierte Integration mit lokaler Überblendung an den Rändern auch lokale Registrierungsfehler in begrenztem Maße ausgeglichen werden können, wie die Beispielbilder in den Abbildungen 4.7 und 4.8 zeigen. Das erste Beispiel zeigt den Ausschnitt eines online berechneten Mosaikbildes, in dem eine Akkumulation von Registrierungsfehlern zu einem signifikanten Versatz der Bilder bei der Integration geführt hat. Demgegenüber enthält die zweite Abbildung ein synthetisches Beispiel, das durch Addition von horizontalen und vertikalen Offsets auf die geschätzten Transformationsparameter generiert wurde.

In beiden Fällen führt die Überblendung der Bilder bei der Integration, im Gegensatz zu einer regionenbasierten Integration ohne Überblendung, zu einer deutlichen Verbesserung des visuellen Eindrucks der resultierenden Mosaikbilder. Zwar lassen sich die Registrierungsfehler auf diese Art nicht gänzlich korrigieren, durch die verbesserte Qualität der Bilder wird jedoch beispielsweise nachfolgenden Analysemodulen der Umgang mit den Mosaikbildern erleichtert.

**Abbildung 4.8:** Ein synthetisches Beispiel zu den Vorteilen einer Überblendung bei einer regionenbasierten Integration von Bildern. Die zwischen den Bildern geschätzten Transformationsparameter wurden durch einen künstlichen Offset verfälscht. Durch die Überblendung (rechts) wird dieser Fehler jedoch, im Gegensatz zu einer rein regionenbasierten Integration (links), gut kompensiert.

# 5 Dynamische Szenen

Bildfolgen enthalten im Allgemeinen sowohl statische als auch dynamische Informationen einer betrachteten Szene. Die weitgehend konstanten, statischen Daten geben Aufschluss über den Aufbau und die Struktur der Szene (statischer Szenenhintergrund), während dynamische Daten aktuelle Veränderungen beschreiben (dynamischer Szenenvordergrund). Sie können indirekt durch eine Auswertung von Differenzen zwischen einzelnen Bildern einer Folge extrahiert werden. Dabei lassen sich die Differenzen grundsätzlich auf verschiedene Ursachen zurückführen, wie beispielsweise variierende Beleuchtungsverhältnisse im Verlauf der Bildaufnahme (vgl. auch Abschnitt 4.1.1) oder auch Objektbewegungen in der Szene. Insbesondere in Bildfolgen aktiver Kameras ist die Kamerabewegung selbst häufig ein Grund für Veränderungen zwischen einzelnen Bildern (Abb. 5.1(a)). Da sie zumeist keine verwertbaren Rückschlüsse für eine adäquate Interpretation der visuellen Daten zulässt[1] und unter Umständen sogar bedeutsame Informationen (wie z.B. Objektbewegungen) maskiert, wird sie im Regelfall vor einer genaueren Analyse der Bildfolgen zunächst mit Hilfe geeigneter Vorverarbeitungsschritte kompensiert (vgl. Kap. 3).

Eine Mosaikbild-basierte Repräsentation von Bildfolgen aktiver Kameras, die auch dynamische Daten enthalten, stellt besondere Anforderungen an die eingesetzten Verfahren. Grundsätzlich führt die explizite Modellierung der Kamerabewegungen im Rahmen der Bildregistrierung und ihre damit verbundene Kompensation zu einer Stabilisierung des statischen Szenenhintergrundes und unterstützt die Detektion relevanter Objektbewegungen. Da Veränderungen in derartigen Bildfolgen jedoch nicht allein auf die Bewegungen der Kamera zurückzuführen sind, ist eine Grundvoraussetzung vieler Registrierungsalgorithmen nicht mehr erfüllt und der Parameterschätzung liegen inkonsistente Daten zu Grunde. Das globale Bewegungsmodell erfasst insbesondere dominante Veränderungen, die nicht durch die Kamerabewegung induziert werden, nur unzureichend, so dass die Registrierung nachhaltig gestört werden kann (vgl. auch Abschnitt 3.2.1).

Sollte sie dennoch gelingen (beispielsweise aufgrund nur kleiner Bildregionen, die dem Modell widersprechen, oder unter einer expliziten Maskierung unabhängig bewegter Teilbereiche, Abschnitt 3.2.1) verbleiben oftmals Differenzen in den registrierten Bildern, die bei einer Integration ohne explizite Berücksichtigung zu Unschärfen und Verwischungen führen und die Qualität des resultierenden Mosaikbildes vermindern (siehe auch Abschnitt 4.1.2). Abbildung 5.1(b) veranschaulicht diese Auswirkungen eines unabhängig bewegten Objektes in einer Bildfolge („Ghosting“). In Abbildung 5.1(c) sind der Vollständigkeit halber ergänzend Differenzen gezeigt, die aus verletzten Mo-

[1] Eine Ausnahme bildet in diesem Zusammenhang eine explizite Analyse der Kamerabewegungen, beispielsweise als Grundlage für eine 3D-Szenenrekonstruktion *(„Structure from Motion“)* [Wen93].

(a) nicht kompensierte Kamerabewegung

(b) unabhängig bewegtes Objekt (Auto) in einer Bildfolge

(c) fehlerhafte Bildregistrierung durch Parallaxe aufgrund einer Translation der Kamera

**Abbildung 5.1:** Verschiedene Ursachen für Differenzen zwischen den Bildern einer Sequenz.

dellannahmen (z.B. Parallaxe durch unzulässige Kamerabewegungen) bei der Homographieschätzung resultieren, hier jedoch nicht näher betrachtet werden. Weitere Details zu diesem Thema finden sich beispielsweise in [Ira98] oder [Pel00].

Ein nahe liegender Ansatz zur Lösung der vorstehend skizzierten Probleme bei der Berechnung von Mosaikbildern dynamischer Szenen, der im Rahmen dieser Arbeit verfolgt wird, besteht in einer expliziten Trennung der statischen und dynamischen Anteile einer Bildfolge und ihrer jeweils gesonderten Repräsentation (s. beispielsweise auch [Gel98]). Den Schwerpunkt in dieser Arbeit bildet dabei eine adäquate Modellierung des statischen Szenenhintergrundes, ohne jedoch die dynamischen Anteile vollständig auszuklammern.[2]

Die gewählte Vorgehensweise beruht auf einer Detektion unabhängig bewegter Teilbereiche und ihrer Maskierung bei der Integration, wodurch die Erstellung eines Mosaikbildes der statischen Anteile einer gegebenen Bildfolge ohne Verwischungen ermöglicht wird (Abschnitt 4.1.2). Die dabei nicht berücksichtigten, dynamischen Informationen in der Sequenz werden anschließend durch eine zeitliche Integration der Detektionsresultate extrahiert und in einer ergänzenden Graphdatenstruktur abgelegt. Gemeinsam mit dem Mosaikbild des statischen Szenenhintergrundes kann interaktiven Systemen somit eine weitgehend vollständige Repräsentation der verschiedenen Daten einer Szene zur Verfügung gestellt werden. Die zeitliche Integration eröffnet darüber hinaus Möglichkeiten, die Zuverlässigkeit der Detektionsresultate zu überprüfen.

Nachfolgend wird zunächst ein Überblick über verschiedene Verfahren zur Detektion unabhängiger Bewegungen in Bildsequenzen gegeben. Der Schwerpunkt liegt dabei auf residuenbasierten Algorithmen, die in der vorliegenden Arbeit zum Einsatz kommen. Im Anschluß daran präsentiert Unterkapitel 5.2 den verwendeten Ansatz zur Extraktion und Repräsentation dynamischer Daten, der auf einer Verfolgung von Zusammenhangskomponenten über die Zeit basiert (Abschnitt 5.2.1). Eine erweiterte Analyse der daraus folgenden Trajektorien, die sowohl eine Konsistenzprüfung der Detektionsresultate als auch eine rudimentäre Interpretation der visuellen Daten erlaubt, wird in Abschnitt 5.2.2 vorgestellt. Unterkapitel 5.3 fasst Ergebnisse des Umgangs mit dynamischen Daten und deren Diskussion zusammen.

[2]Dabei liegt die Annahme zu Grunde, dass in den ersten Bildern einer Sequenz der statische Szenenhintergrund eindeutig identifiziert und zur Initialisierung des Mosaiks herangezogen werden kann.

## 5.1 Detektion unabhängiger Bewegungen

Im Rahmen einer Berechnung von Mosaikbildern existieren verschiedene Ansätze zur Detektion unabhängiger Bewegungen in einer gegebenen Bildsequenz. Einerseits kann die Extraktion der verschiedenen Bewegungsmuster direkt durch eine simultane Parameterschätzung realisiert werden, wobei die Kamerabewegung als *ein* Element aus einer Menge mehrerer Bewegungskomponenten interpretiert und nicht explizit von den übrigen unterschieden wird. Sawhney und Kollegen [Saw96] stellen dazu ein aufwändiges, statistisches Verfahren vor, das auf der expliziten Beschreibung der einzelnen Bewegungen durch geeignete Modelle und einer gleichzeitigen Schätzung aller Parametersätze beruht. Sofern sich die Anteile der Bewegungsmuster an der Gesamtbewegung hinreichend deutlich unterscheiden, bietet sich auch eine gestaffelte Segmentierung der verschiedenen Komponenten an. Dabei werden schrittweise Parameter der jeweils dominierenden Bewegung in der Bildfolge geschätzt, unter Maskierung der zuvor bereits modellierten Anteile. Während dem Ansatz in [Ber92b] dabei eine implizite Modellierung der verschiedenen Bewegungen im Rahmen eines iterativen Algorithmus zu Grunde liegt, basiert das in [Ira94] publizierte Verfahren auf einer sukzessiven, expliziten Detektion und Maskierung der einzelnen Bewegungsanteile im Verlauf der Schätzung. In beiden Fällen wird durch diese Vorgehensweise sogar eine Separation überlagerter, transparenter Bewegungen erreicht, wie sie beispielsweise aus Spiegelungen bewegter Objekte in Glas resultieren.

Eine zweite Klasse von Methoden zur Bewegungsdetektion setzt eine erfolgreiche Registrierung der Bilder einer Folge voraus. In registrierten Bildern lassen sich unabhängig bewegte Teilgebiete durch eine Analyse von verbliebenen Differenzen *(Residuen)* ermitteln. Im Gegensatz zu den im vorhergehenden Absatz beschriebenen Verfahren steht dabei die Detektion der bewegten Regionen im Vordergrund und weniger eine exakte, mathematische Beschreibung ihrer Bewegungscharakteristika. Residuen können sowohl zwischen aufeinander folgenden Bildern einer Sequenz als auch relativ zu einem bereits generierten Mosaikbild berechnet werden, das sich als Modell des statischen Szenenhintergrundes auffassen lässt. Da die zweite Variante im Regelfall zu vollständigeren Bewegungsdaten führt, wird sie in den meisten Fällen bevorzugt (vgl. auch [Möl01a]).

Das zu Grunde gelegte Mosaikbild wird im Allgemeinen aus den zuvor registrierten Bildern der Folge erzeugt, wobei sich bei einer Veränderung der statischen Anteile einer Szene über die Zeit (z.B. veränderte Lichtverhältnisse im Verlauf eines Tages bei Außenaufnahmen) eine kontinuierliche Adaption empfiehlt. Diese kann einerseits indirekt durch die sukzessive Integration neuer Daten, und andererseits direkt über eine explizite Modellierung der Veränderungen realisiert werden. In [Lip99] und [Bha00] werden die Eigenschaften des Szenenhintergrundes im zeitlichen Verlauf jeweils durch Filter modelliert, wobei in [Bha00] zusätzlich auch die Klassifikationsregel für die unabhängigen Bewegungen selbst adaptiv ist. Mittal gründet sein Hintergrundmodell ausschließlich auf den Bilddaten, wobei die einzelnen Pixelwerte jeweils über Gaußsche Mischverteilungen vorhergesagt werden [Mit00]. Dabei schließt die Modellierung Heuristiken zur Bewegungsdetektion direkt ein, da Pixel, an denen die beobachteten Werte deutlich von den Prädiktionen abweichen, als bewegt eingestuft werden.

Die Detektion unabhängiger Bewegungen über Residuen lässt sich sowohl bei einer inkrementellen Verarbeitung von Bilddaten als auch in Offline-Ansätzen anwenden. Wenn zum Zeitpunkt der Detektion die vollständige Bildfolge als Datenbasis zur Verfügung steht, führt dies allerdings auf zuverlässigere Bewegungsdaten. Für jedes Pixel sind dann prinzipiell Informationen aus mehreren Bildern der Folge verfügbar [Még99], die z.B. über Voting-Schemes [Par94] miteinander verknüpft werden können. Daraus lassen sich verlässlichere Klassifikationsresultate ableiten als das in Online-Verfahren auf Basis einzelner Differenzwerte zwischen dem Referenzbild und nur einem registrierten Bild möglich ist.

### 5.1.1 Residuenbasierte Detektion

Im Allgemeinen bilden das Mosaikbild des statischen Szenenhintergrundes und ein relativ zu dessen Koordinatensystem registriertes Bild den Ausgangspunkt einer residuenbasierten Detektion von Bewegungen in einer Szene. Die unabhängig bewegten Bereiche, die dabei zu ermitteln sind, werden unter Anwendung der zuvor geschätzten Transformation nur unzureichend an die vorhandenen Daten des Mosaikbildes angepasst. Als Folge resultieren pixelweise markante Residuen $R(u)$ zwischen den Farb- bzw. Intensitätswerten des registrierten Bildes $I'$ und des aktuellen Mosaikbildes $M$. Diese Residuen können einer binären Klassifikation $K_b$ mit einem Schwellwert $\theta_R$[3] zu Grunde gelegt werden:

$$K_b(u) = \begin{cases} 0, & \text{falls } R(u) \leq \theta_R, \qquad \text{(unbewegt)} \\ 1, & \text{sonst} \qquad\qquad\qquad\quad \text{(bewegt)} \end{cases} \tag{5.1}$$

Zur Berechnung der Residuen $R$ sind in der Literatur zahlreiche Distanzmaße zu finden. Im Folgenden wird ein Überblick über einige wichtige Ansätze gegeben, die im Rahmen der vorliegenden Arbeit implementiert wurden. Details und ausführlichere Analysen dazu finden sich in [Möl01a]. Die Abstandsmaße werden grundsätzlich wahlweise separat auf alle Farbkanäle eines Bildes angewendet und anschließend zu einer Maßzahl pro Pixel verrechnet, oder bereits im Ansatz auf einzelne Kanäle beschränkt. In der vorliegenden Arbeit basiert die Bewegungsdetektion allein auf Grauwertbildern (vgl. Abschnitt 3.5.2), so dass sich die Darstellung im Folgenden auf einen einzelnen Intensitätskanal bezieht.

Im einfachsten Fall lassen sich Residuen zwischen Bildern durch eine pixelweise Berechnung von Intensitätsdifferenzen ermitteln. Dabei ermöglicht eine geeignete Normierung der Bildintensitäten beider Bilder den Ausgleich von Schwankungen der Bildenergie. Da die pixelweise Berechnung der Residuen sensitiv auf Bildrauschen reagiert, hat sich eine Mittelung der Werte in einer quadratischen Nachbarschaft $V_u$, mit $\sqrt{|V_u|} \in \{3, 5, 7\}$, eines jeden Pixels $u$ als vorteilhaft erwiesen:

$$\bar{R}_{I_n}(u) \;=\; \frac{1}{|V_u|} \cdot \sum_{v \in V_u} R_{I_n}(v) \;\;,\;\; R_{I_n}(v) \;=\; \left| \frac{M(v)}{\mu_M} - \frac{I'(v)}{\mu_{I'}} \right| .$$

[3] Die zu wählenden Schwellwerte hängen einerseits direkt von dem zur Residuenberechnung verwendeten Distanzmaß ab und werden andererseits auch durch die Bildqualität der zu analysierenden Sequenzen beeinflusst, so dass hier nur schwer allgemeingültige Richtwerte angegeben werden können.

$\mu_M$ und $\mu_{I'}$ bezeichnen in der vorstehenden Gleichung, analog zu Gleichung 3.4 in Abschnitt 3.5.1, jeweils die mittleren Intensitätswerte der beiden betrachteten Bilder.

Ein alternatives Distanzmaß ist über den Betrag des *Normal Flow* $\vec{f}$ gegeben, der durch die Norm der senkrecht zum lokalen Gradienten $\nabla M(u)$ stehenden Komponente des optischen Flusses definiert wird (z.B. [Coh99, Ira98, Ple99]):

$$R_F(u) = \| \vec{f}(u) \| = \frac{|M(u)^{(t)}|}{\| \nabla M(u) \|}.$$

$M(u)^{(t)}$ entspricht dabei der zeitlichen Ableitung von $M$ an der Position $u$, die durch die pixelweise Differenz zwischen $M$ und $I'$ approximiert wird. Im Gegensatz zur Intensitätsdifferenz bezieht dieses Distanzmaß die lokalen Bildstrukturen in die Berechnungen ein und setzt die Differenzen zu ihnen ins Verhältnis. Auch hierbei hat sich eine Mittelwertbildung in einer Nachbarschaft $V_u$ bewährt. Da $\vec{f}$ allerdings nur an Bildpunkten mit einem nicht verschwindenden Gradienten definiert ist, können nicht zwingend flächendeckend Residuen berechnet werden. Lückenlose Bewegungsdaten lassen sich erst durch die Ergänzung einer zusätzlichen Konstanten $C$ im Nenner des Bruches oder eine geeignete Nachbearbeitung der binären Bewegungsbilder gewinnen (s. Unterkap. 5.2.1). Darüber hinaus wird der Normal Flow in der Praxis oftmals mit den lokalen Gradientennormen gewichtet [Ira94, Még99]:

$$\bar{R}_{F_n}(u) = \frac{1}{\sum\limits_{v \in V_u} \| \nabla M(v) \|^2 + C} \sum_{v \in V_u} \frac{|M(v)^{(t)}|}{\| \nabla M(v) \|} \cdot \| \nabla M(v) \|^2 \quad , \quad C \in \mathbb{R}, C > 0.$$

Sowohl der Normal Flow als auch die normierte Intensitätsdifferenz weisen eine hohe Abhängigkeit von den konkreten Pixelwerten der Bilder auf. Die enge Verknüpfung wird gelöst, wenn statt der Differenzen selbst ihre Konvergenzeigenschaften unter Anwendung der berechneten Transformation betrachtet werden [BE94]. Dies geschieht durch einen Vergleich der pixelweisen Residuen zwischen Mosaik $M$ und Bild $I$ jeweils vor und nach Anwendung der Transformation:

$$R_C(u) = \frac{\delta(u) - \delta'(u)}{\delta(u) + \delta'(u)} \quad \text{mit} \quad \delta(u) = |M(u) - I(u)| \ , \ \delta'(u) = |M(u) - I'(u)|.$$

An Bildpunkten, die dem globalen Bewegungsmodell unterliegen, verringert sich die zu beobachtende Differenz tendenziell (d.h. $R_C(u) \in ]0, 1]$), während sie an unabhängig bewegten Pixeln eher konstant bleibt oder zunimmt ($R_C(u) \in [-1, 0]$). Durch eine binäre Klassifikation mit einem Schwellwert zwischen $-1$ und 0 lassen sich somit unabhängig bewegte Pixel lokalisieren. Allerdings erfordert das Konvergenzmaß $R_C$ die Anwendung von Transformationen ungleich der Identität, da nur in diesen Fällen Änderungen in den pixelweisen Differenzen induziert werden. Damit ist eine Bewegung der Kamera zwingende Voraussetzung für seine Anwendbarkeit, wodurch die praktische Relevanz beschnitten wird. Im Kontext des visuellen Speichers finden daher überwiegend die normierte Intensitätsdifferenz und der Normal Flow Verwendung.

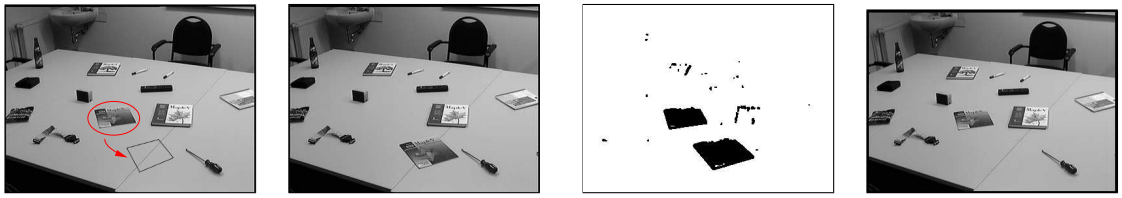
## 5.2 Zeitliche Integration von Bewegungsdaten

Die Ausblendung als bewegt klassifizierter Teilbereiche eines Bildes bei der Integration in ein Mosaikbild fokussiert die Repräsentation auf statische Anteile einer Szene. Da die vollständige visuelle Information jedoch im Allgemeinen auch dynamische Komponenten umfasst, führt diese Vorgehensweise zu einer unerwünschten Einschränkung der repräsentierten Daten. Ein Ansatz, um diesem schwerwiegenden Informationsverlust entgegenzuwirken, besteht darin, die dynamischen Anteile in einer Bildfolge explizit zu extrahieren und parallel zum Mosaikbild des statischen Szenenhintergrundes in einer ergänzenden Datenstruktur zu repräsentieren. Im Kontext der vorliegenden Arbeit sind die dynamischen Szenenanteile im Wesentlichen durch die bewegten Objekte und ihre Bewegungsmuster definiert. Sie werden durch Trajektorien beschrieben, die aus einer Verfolgung *(Tracking)* bewegter Zusammenhangskomponenten über die Zeit resultieren. Eine exaktere Beschreibung der Objektbewegungen durch komplexere mathematische Modelle war nicht Ziel der Arbeit und wird daher hier nicht weiter betrachtet.

Der skizzierte Ansatz erlaubt neben der Erhaltung der dynamischen Daten zusätzlich eine Überprüfung ihrer Verlässlichkeit. So korrespondieren nicht in allen Fällen als bewegt klassifizierte Teilgebiete auch tatsächlich zu realen Bewegungen in einer Szene. Wird beispielsweise ein Objekt innerhalb einer Szene verschoben, das zuvor dem statischen Szenenhintergrund zugerechnet wurde (vgl. Abb. 5.2(a) und Abb. 5.2(b)), so resultieren daraus im Allgemeinen zwei kompakte Mengen bewegter Pixel (Abb. 5.2(c)). Einerseits treten Unterschiede gegenüber dem Referenzmosaik an der aktuellen Position des Objektes auf, andererseits aber auch an seiner initialen Position in der Szene. Eine Ausblendung beider Gebiete verhindert zwar die unerwünschte Integration des bewegten Objektes ins Mosaikbild, konserviert dort aber gleichzeitig auch seine ursprüngliche, nicht länger gültige Initialposition (Abb. 5.2(d)). Dies führt zu einer lokal inkonsistenten Repräsentation des Szenenhintergrundes (vgl. auch Abb. 4.3(c)). Die zeitliche Integration der Daten bildet eine gute Grundlage zur Identifikation solcher fehlklassifizierten Bereiche und ermöglicht somit eine lokale Korrektur des Mosaikbildes. In den beiden folgenden Abschnitten werden das Tracking und die anschließende Trajektorienanalyse im Überblick beschrieben. Weiterführende Details zu beiden Ansätzen finden sich in [Möl01a], [Möl01b] oder auch [Möl02].

### 5.2.1 Tracking von Zusammenhangskomponenten

Den Ausgangspunkt zur Extraktion der dynamischen Parameter von bewegten Objekten in einer Bildsequenz bilden die binären Bewegungskarten, die sich aus der pixelweisen Klassifikation berechneter Residuen ergeben (Gl. 5.1). Auf diesen Binärbildern wird zunächst eine morphologische Dilatation durchgeführt, wobei quadratische strukturierende Elemente mit Größen von $5 \times 5$ oder $7 \times 7$ Anwendung finden. Durch die Dilatation lassen sich die Daten einserseits lokal glätten. Darüber hinaus werden Bewegungsdaten einzelner Pixel dabei in eine lokale Nachbarschaft propagiert, so dass kleine Lücken in



(a) Referenzbild (b) neues Bild (c) Bewegungsbild (d) neues Mosaik

**Abbildung 5.2:** Ergebnis der Bewegungsdetektion zwischen einem Referenzmosaik und einem neuen Bild: Innerhalb der Szene wird ein Objekt auf eine neue Position verschoben ((a),(b)), wodurch sowohl an dessen Initialposition als auch an der neuen Position Differenzen gegenüber dem Referenzbild hervorgerufen werden, die zu zwei Clustern bewegter Pixel führen (c).

den Bewegungsbildern geschlossen und auch an initial undefinierten Bildpunkten Klassifikationsresultate verfügbar gemacht werden können. Letzteres ist vorrangig bei einer Berechnung von Residuen auf Basis des Normal Flow bedeutsam (s. S. 71).

Im Anschluss an die Vorverarbeitung erfolgt eine Segmentierung der einzelnen Gruppen bewegter Pixel in Regionen mit Hilfe eines *"Component Labeling"*-Algorithmus. Resultierende Regionen, deren Größen einen extern spezifizierten Schwellwert $\nu_R$ unterschreiten, werden dabei als Bildrauschen interpretiert und von der weiteren Betrachtung ausgeschlossen. Geeignete Werte für $\nu_R$ liegen üblicherweise zwischen 150 und 200 Pixeln, wobei jedoch in Abhängigkeit vom jeweiligen Anwendungskontext auch deutlich davon abweichende Werte sinnvoll sein können.

Die verbleibenden Regionen bilden die Basis des Trackings. Um unvermeidlichen Varianzen innerhalb der Regionensegmentierung über die Zeit vorzubeugen, werden die Regionen dabei nicht einzeln betrachtet, sondern zuvor in *Zusammenhangskomponenten* räumlich benachbarter Regionen gruppiert. Dadurch lässt sich eine robuste Verfolgung selbst bei Objekten erzielen, die partiell nur einen geringen Kontrast gegenüber dem Szenenhintergrund aufweisen und aufgrund dessen in zeitlich variierende Regionen zerfallen (vgl. etwa die Regionen in den beiden rechten Teilbildern der Abb. 5.3). Die räumliche Nachbarschaft einzelner Regionen leitet sich aus dem minimalen Punktabstand zwischen ihnen ab. Als Grundlage des Trackings wird jede Zusammenhangskomponente durch spezifische Merkmale charakterisiert. Im Wesentlichen sind dies ihr Schwerpunkt, ihre Größe und das Histogramm über alle Grauwerte enthaltener Pixel. Die Merkmale lassen sich direkt auf Basis der jeweils eingeschlossenen Regionen bestimmen.

Während des Trackings werden detektierte Zusammenhangskomponenten aufeinander folgender Bilder jeweils paarweise miteinander verglichen. Eine Zuordnung als korrespondierend erfolgt, wenn die normierten Intensitätshistogramme eine festgelegte Mindestähnlichkeit (beispielsweise etwa 70%) aufweisen und eine maximale Größendifferenz ($\approx 15 - 20\%$) nicht überschritten wird. Die Ähnlichkeit der Histogramme gründet dabei auf dem Grad der Überdeckung ihrer Flächen. Zur Reduktion des a priori quadratischen Aufwandes werden nur Komponenten miteinander verglichen, deren Schwerpunkte hinreichend nah beieinander liegen. Der dafür festzulegende Maximalabstand lässt sich unter Berücksichtigung der zu erwartenden Geschwindigkeiten von Objekten in der Szene und der Kamera sowie der Verarbeitungsrate der Bilder abschätzen.

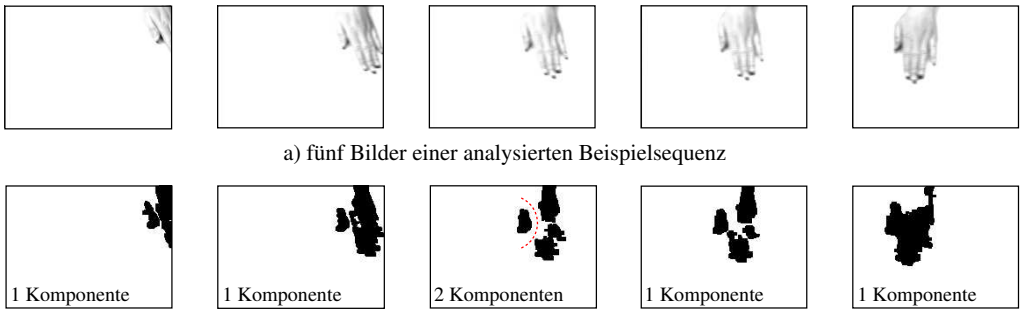



**Abbildung 5.3:** Segmentierte Regionen und Zusammenhangskomponenten bei der Verfolgung eines bewegten Objektes, das über die Zeit jeweils in unterschiedliche Teilregionen zerfällt.

Grundsätzlich können mit der skizzierten Vorgehensweise bewegte Objekte robust verfolgt werden, deren äußeres Erscheinungsbild im Verlauf der Bildfolge lediglich kleinen Änderungen unterworfen ist und die einen ausreichenden Kontrast zum Hintergrund aufweisen. Zerfallen Objekte bei der Segmentierung in mehrere Regionen, garantiert das Tracking der Zusammenhangskomponenten weiterhin eine robuste Verfolgung. Probleme treten erst auf, wenn die zu einem Objekt korrespondierenden, bewegten Regionen in den einzelnen Bildern in unterschiedliche Zusammenhangskomponenten gruppiert werden (z.B. aufgrund von Änderungen der spezifischen Objekteigenschaften innerhalb einer Bildsequenz, so dass kein allgemeingültiger Schwellwert für die maximale Punktentfernung benachbarter Regionen in den Komponenten angegeben werden kann). In solchen Fällen (wie in Abb. 5.3 zwischen den mittleren drei Bildern) lassen sich die Zusammenhangskomponenten zumeist bereits aufgrund signifikanter Größendifferenzen nicht mehr zuordnen und der zeitliche Kontext der Bewegung geht verloren. Um diesem Effekt entgegenzuwirken, werden Zusammenhangskomponenten, die ohne gültige Zuordnung bleiben, gemäß der Potenzmenge enthaltener Regionen in Teilzusammenhangskomponenten zerlegt. Diese werden anschließend erneut paarweise mit den Teilzusammenhangskomponenten des anderen Bildes verglichen und auf Übereinstimmungen untersucht.[4] Auf diese Weise lässt sich auch der Zerfall oder die Verschmelzung von Objekten in einer Szene erfassen und in einen korrekten Kontext einordnen.

Die Verwaltung der Ergebnisse des Trackings erfolgt in einer kontinuierlich aktualisierten Graphdatenstruktur, dem *Korrespondenzgraphen* (vgl. [Coh99]). Dabei handelt es sich um einen zeitlich gerichteten Graphen, dessen Knoten jeweils mit einzelnen Zusammenhangskomponenten assoziiert werden. Kanten symbolisieren detektierte Korrespondenzen, wobei ein Knoten in Abhängigkeit zugeordneter Teilzusammenhangskomponenten auch mehrere ein- und ausgehende Kanten aufweisen kann. Aus zusammenhängenden Pfaden im Graphen lassen sich final die Trajektorien bewegter Objekte rekonstruieren.

[4] Obgleich die Potenzmenge im Allgemeinen deutlich mehr Elemente enthält als die Ursprungsmenge, ist der mit einem solchen Vorgehen verbundene Aufwand in diesem Kontext nur gering, da zumeist verhältnismäßig wenige bewegte Regionen in einer Komponente zusammengefasst werden.

### 5.2.2 Trajektorienanalyse

Die extrahierten Trajektorien charakterisieren die Bewegungsmuster von Objekten innerhalb einer Bildfolge über die Zeit. Gemeinsam mit dem Mosaikbild des statischen Szenenhintergrundes können sie als Eingangsdaten nachfolgender Analysemodule dienen, die unter Hinzuziehung von Weltwissen eine Interpretation der visuellen Daten zu generieren versuchen. Auch ohne die Nutzung weitergehenden Weltwissens erlauben die Trajektorien jedoch oftmals bereits eine rudimentäre Interpretation der Daten. Werden beispielsweise Teilbereiche einer Szene (z.B. der Eingang eines Gebäudes, in dem Personen ein- und ausgehen) mit spezifischen Aktionen oder Ereignissen assoziiert , so kann die räumliche Lage von Trajektorien Hinweise auf entsprechende Aktivitäten in einer Szene geben [Iva99]. Dabei lassen sich durch eine Verknüpfung verschiedener Aktivitäten auch komplexe Szenarien modellieren [Med98, Med01]. In [Bra96] wird ein Ansatz zur Analyse von Manipulationshandlungen in Videosequenzen vorgestellt, der primär auf einer Auswertung von Interaktionen zwischen bewegten Regionen beruht. Aus Aufspaltungen, Verschmelzungen und Berührungen verschiedener bewegter Regionen wird auf Geschehnisse in der aufgenommenen Szene rückgeschlossen, so dass sich eine semantische Interpretation der visuellen Daten herleiten lässt.

Der im vorausgegangenen Abschnitt eingeführte Korrespondenzgraph bietet einen guten Ausgangspunkt für derartige Untersuchungen. Objektspaltungen und -verschmelzungen lassen sich durch eine Analyse des Verhältnisses zwischen ein- und ausgehenden Kanten einzelner Knoten im Korrespondenzgraphen detektieren und unter Einbeziehung der zeitlichen Historie auch interpretieren. Eine solche, prototypische Analyse der Daten wurde auf Basis eines endlichen Automaten realisiert [Möl02]. Insbesondere spezifische Aktivitäten in dem Konstruktionsszenario des SFB 360 [Ric96] konnten dabei auf Basis von Objektbewegungen rudimentär interpretiert werden. Im Hinblick auf die vorrangige Zielsetzung des visuellen Speichers, den statischen Hintergrund einer Szene in Mosaikbildern adäquat zu repräsentieren, ist jedoch eine Überprüfung der Verlässlichkeit der extrahierten Bewegungsdaten von größerer Bedeutung als ihre Interpretation. Auch dafür stellt der Graph mit den repräsentierten Objekttrajektorien eine geeignete Basis bereit.

Fehlklassifizierte Bereiche in der Bewegungsdetektion resultieren oftmals aus Positionsveränderungen von Objekten in einer betrachteten Szene (vgl. Abb. 5.2). Ebenso führt das Hinzufügen und Entfernen von Objekten zu lokalen Differenzen zwischen der Hintergrundrepräsentation und den aktuellen Bilddaten, aus denen als bewegt klassifizierte Regionen resultieren, die über die Zeit verfolgt werden können. In allen drei Fällen sind die vermeintlich bewegten Regionen jedoch tatsächlich statisch, da sie zu keiner realen Objektbewegung korrespondieren. Diese Eigenschaft lässt sich insbesondere anhand der Varianzen innerhalb der Trajektorienpunkte zugehöriger Zusammenhangskomponenten feststellen. Die Punkte weisen im Allgemeinen nur eine geringe Varianz auf, die somit eindeutige Hinweise zur Identifikation der fehlklassifizierten Regionen liefert.

Zur Verifikation der Detektionsresultate wird in den einzelnen Knoten des Graphen zusätzlich die Varianz der Trajektorienpunkte über eine geeignet festzulegende Anzahl zurückliegender Zeitpunkte gespeichert. Bei jeder Erweiterung der Trajektorie um neue

Punkte erfolgt eine Aktualisierung der Daten und gleichzeitige Überprüfung der errechneten Varianz. Liegt sie unterhalb einer gegebenen Schwelle (z.B. $\approx 3$ Pixel), so deutet dies auf eine Fehlklassifikation der Zusammenhangskomponente hin, die somit aus der aktuellen Bewegungskarte zu entfernen ist. Damit werden die fälschlich als bewegt eingeordneten Bildregionen nicht länger bei der Integration maskiert, so dass aktuelle Daten ins Mosaikbild kopiert und die Konsistenz der Repräsentation mit den realen Gegebenheiten der Szene wiederhergestellt werden können.

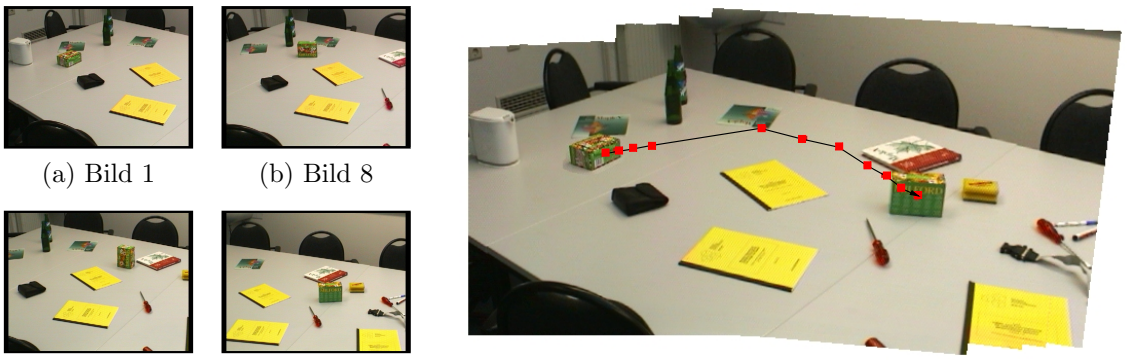
## 5.3 Diskussion

Das Konzept des in dieser Arbeit vorgestellten ikonischen Speichers sieht grundsätzlich die Verarbeitung von Bildfolgen statischer *und* dynamischer Szenen vor. Die vorrangige Zielsetzung der Arbeit besteht dabei in erster Linie in einer adäquaten Modellierung des statischen Szenenhintergrundes und weniger in einer exakten Beschreibung aller Bewegungskomponenten. Aus diesem Grund beruht der gewählte Lösungsansatz primär auf einer Trennung der statischen und dynamischen Anteile und ihrer jeweils gesonderten Repräsentation in einem Mosaikbild bzw. einer parallel dazu verwalteten Graphdatenstruktur. Die in den vorhergehenden Abschnitten skizzierten Verfahren zur Detektion von Residuen stellen dafür eine gute Grundlage bereit, da sie im Allgemeinen eine sichere Identifikation bewegter Objektpixel gewährleisten. Allerdings setzt eine solche Herangehensweise zwingend eine erfolgreiche Registrierung der Bilder voraus. Dies kann jedoch unter der Präsenz von bewegten Objekten in einer Bildfolge nicht immer sichergestellt werden, da eine robuste Parameterschätzung eine konsistente Datenbasis erfordert, diese jedoch insbesondere durch bewegte Objekte korrumpiert wird (vgl. Abschnitt 3.2.1).

Der daraus resultierenden, wechselseitigen Abhängigkeit wird in der vorliegenden Arbeit durch eine im Allgemeinen leicht zu erfüllende Vorbedingung begegnet. Sie erlaubt zu Beginn nur geringe dynamische Anteile in einer Bildfolge, die den Prozess der Parameterschätzung nicht stören und eine eindeutige Identifikation des statischen Hintergrundes ermöglichen. Alternativ lassen aber auch extern bereitgestellte, initiale Parameter eine direkte Detektion bewegter Objekte ohne eine vorherige, fehleranfällige Parameterschätzung zu.

Eine abschließende Schwierigkeit bei der Anwendung residuenbasierter Detektionsverfahren, die nicht übersehen werden darf, besteht in der Festsetzung geeigneter Schwellwerte für die binäre Klassifikation. Diese Werte sind zumeist abhängig von den spezifischen Bildeigenschaften und daher nur schwer automatisch festzulegen. Aus einer Analyse mehrerer Bildfolgen einzelner Anwendungsszenarien lassen sich jedoch zumeist gute Richtwerte ableiten.

Auch wenn eine adäquate Modellierung des statischen Szenenhintergrundes in dieser Arbeit den Schwerpunkt bildet, sollten dynamische Informationen in einer Bildfolge bei der Generierung einer Mosaikbild-basierten Szenenrepräsentation nicht verloren gehen. In der vorliegenden Arbeit wird diesem Aspekt durch eine zeitliche Integration der Klassifikationsresultate Rechnung getragen. Dabei bildet eine Gruppierung bewegter Pixel in



(a) Bild 1 (b) Bild 8

(c) Bild 12 (d) Bild 17

**Abbildung 5.4:** Mosaikbild einer Bildfolge, die ein unabhängig bewegtes Objekt beinhaltet (s. Beispielbilder links). Die extrahierte Trajektorie des Objektes ist im Mosaikbild eingezeichnet.

Regionen und Zusammenhangskomponenten den Ausgangspunkt. Die Zusammenhangskomponenten ermöglichen eine robuste Verfolgung von bewegten Objekten trotz auftretender Varianzen in der Regionensegmentierung, die bei Detektionsansätzen auf Basis von Residuen nicht auszuschließen sind (Abb. 5.3). Darüber hinaus trägt auch das Zuordnungskriterium für die Komponenten auf Basis ihrer Intensitätshistogramme zu einer hohen Robustheit der Verfolgung bei, da selbst Rotationen und moderate Formänderungen ohne erhöhten Aufwand behandelt werden können (Abb. 5.4).

Im Rahmen dieser Arbeit wird die Domänenunabhängigkeit bei der Extraktion von Bewegungsdaten über ein robustes Tracking spezifischer Objekte gestellt. Damit fließt keinerlei Modellwissen in die Algorithmen ein. Grundsätzlich gilt, dass eine Verfolgung von auftretenden Objekten in einer Szene mit zunehmendem Wissen über deren Eigenschaften einfacher wird. Insbesondere eine Modellierung der Bewegungsmuster und Formveränderungen von Objekten über die Zeit kann zu erheblichen Verbesserungen in der Performanz führen. Gleichzeitig wird mit einer stärkeren Berücksichtigung von Modellwissen jedoch die Menge zulässiger Objekte beschnitten. Der gewählte Ansatz erzwingt in dieser Hinsicht keinerlei Einschränkungen und ist damit zwangsläufig mit einem erhöhten Risiko verbunden, in Form und Farbe signifikant variierende Objekte nicht robust verfolgen zu können. Dennoch bleiben die dynamischen Daten auch in diesen Fällen in Form der extrahierten Residuen erhalten und können bei Bedarf zu einem späteren Zeitpunkt mit spezialisierteren Ansätzen weiter untersucht werden.

Die implementierten Verfahren zur Objektverfolgung bilden eine ausreichende Grundlage zur Verifikation extrahierter Bewegungsdaten, obwohl sie nicht immer eine vollständige Extraktion der Objekttrajektorien gewährleisten können. Eine Verifikation der Detektionsresultate ist insbesondere im Hinblick auf eine möglichst exakte Übereinstimmung der realen Szenendaten mit der ikonischen Repräsentation des Szenenhintergrundes unerlässlich. Fehlklassifizierte Regionen werden im Wesentlichen durch eine verschwindende Varianz innerhalb ihrer Trajektorienpunkte charakterisiert. Da sie darüber hinaus keinen Formveränderungen unterworfen sind, erlauben die vorgestellten Detektions- und Verfolgungsalgorithmen in Kombination mit der skizzierten Varianzanalyse eine robuste

**Abbildung 5.5:** Entwicklung eines Mosaikbildes mit dynamischer Adaption des Szenenhintergrundes: Die dem Mosaik zu Grunde liegende Bildfolge ist in Abb. 5.6 zu sehen. Während der Aufnahme wurde die Szene durch das Hinzufügen bzw. Entfernen von Objekten an verschiedenen Stellen verändert.

Erkennung dieser Regionen. Bewegte Objekte lassen sich damit in aller Regel zuverlässig ausblenden. Darüber hinaus werden Objekte, die aufgrund von Änderungen in ihren Bewegungscharakteristika fehlklassifiziert wurden, nach der Identifikation dem Szenenhintergrund wieder zugeordnet bzw. vollständig daraus entfernt. Auf diese Weise kann die Konsistenz der Repräsentation mit der realen Szene wiederhergestellt werden, so dass sich insgesamt eine stabile Repräsentation des statischen Szenenhintergrundes im zeitlichen Verlauf gewährleisten lässt (Abbildungen 5.5 und 5.6).

(a) Bild 1 (b) Bild 11 (c) Bild 26 (d) Bild 33 (e) Bild 42

**Abbildung 5.6:** Fünf Bilder der Folge, aus der das Mosaikbild in Abbildung 5.5 berechnet wurde. Im Verlauf der Bildaufnahme wurden der Szene sowohl Objekte zugefügt wie auch entfernt.

Bewegungsdaten in einer Bildfolge sind ein grundlegender Baustein in der Analyse und Interpretation der visuellen Daten. Mit Blick auf das menschliche Interaktions- und Kommunikationssystem (Kap. 1) stellen sie einen entscheidenden Faktor bei der aufmerksamkeitsbasierten Datenakquisition zur Verhaltensplanung dar. Kapitel 7 der vorliegenden Arbeit befasst sich mit der Verknüpfung der ikonischen Szenenrepräsentation und einer aktiven Akquisition neuer Daten. Auch dafür bilden die mit den Algorithmen dieses Kapitels gewonnenen Bewegungsdaten eine wichtige Grundlage.

# 6 Multi-Mosaikbilder

In der Informatik hat sich in den zurückliegenden Jahren zunehmend die Tendenz herausgebildet, Entwicklungen interaktiver technischer Systeme verstärkt auf einen Einsatz außerhalb der Laboratorien und in der Alltagswelt auszurichten. Dafür stellt nicht zuletzt das jüngst von Honda™ vorgestellte, neue Modell des Roboters „Asimo“ einen Beleg dar.[1] Er ist aufgrund ausgereifter motorischer Fähigkeiten in der Lage, auch in hochgradig dynamischen Umgebungen zu agieren, wo schnelle Reaktionen und Bewegungsabläufe erforderlich sind. Allerdings setzt eine Abkehr von strikten Laborbedingungen nicht nur eine robuste Mechanik voraus, sondern auch flexible sensorische Komponenten und effiziente Analysestrategien für die akquirierten Informationen. Ohne solche Mechanismen ist eine zielgerichtete Verhaltensplanung und Ansteuerung der Motorik, die auch eine *aktive* Auswahl interessanter Informationen einschließt, nicht realisierbar.

Im Hinblick auf die Verarbeitung *visueller* Daten in technischen Systemen hat das Forschungsgebiet der Active Vision große Bedeutung erlangt. Dort wird eine aktive Auswahl relevanter Informationen unter anderem durch eine hardwareseitige Kameraansteuerung und softwarebasierte Fokussierungsmechanismen realisiert. Allerdings entfalten diese Ansätze ihre volle Leistungsfähigkeit erst in Kombination mit geeigneten Strukturen zur internen Repräsentation der Bilddaten. Diese unterstützen sowohl eine Verknüpfung der Daten mit zusätzlichem Weltwissen als auch eine Aufdeckung zeitlicher Zusammenhänge, die die Basis einer stabilen Wahrnehmung der Umgebung bilden.

Grundsätzlich können interne Repräsentationen visueller Daten auf verschiedenen Abstraktionsebenen angesiedelt sein (vgl. z.B. [Jun98, Bau04] bzw. auch Kap. 1, S. 4). Im Rahmen dieser Arbeit wurde ein Konzept entwickelt, dessen Schwerpunkt auf einer signalnahen, ikonischen Repräsentation von Bilddaten liegt. Sie ist nicht auf spezifische Anwendungskontexte ausgerichtet, wie etwa eine Erkennung von Objekten oder eine 3D-Szenenrekonstruktion, sondern unterstützt vielmehr eine direkte Anwendung konventioneller Bildverarbeitungstechniken, so dass eine große Flexibilität bei der Einbindung der Repräsentationsdatenstruktur in verschiedene interaktive Systeme resultiert.

Den Kern der Struktur bilden Mosaikbilder, die eine redundanzfreie, zeitlich integrierte ikonische Darstellung von Bildfolgen aktiver Kameras ermöglichen (Unterkap. 1.1). Der überwiegende Teil bislang veröffentlicher Arbeiten zur Berechnung von Mosaikbildern und deren weiterer Verwendung zielt dabei, im Gegensatz zu dieser Arbeit, weder auf eine dauerhafte Online-Verarbeitung von Bildfolgen noch auf eine direkte Anwendung von Bildverarbeitungsalgorithmen auf die Bilder. Um beiden Anforderungen zu genügen, wurde das neue Konzept der *Multi-Mosaikbilder* entwickelt, das in diesem Kapitel detail-

[1] `http://world.honda.com/ASIMO/`

liert diskutiert wird (s. auch [Möl04]). Das Referenzkoordinatensystem eines Multi-Mosaikbildes ist durch eine Menge verschieden orientierter Ebenen definiert, die gleichmäßig um das optische Zentrum der Kamera angeordnet werden. Die grundsätzliche 3D-Struktur dieser Anordnung orientiert sich dabei an *Polyedern*, so dass eine adäquate Repräsentation des Sichtbereichs einer stationären rotierenden Kamera ermöglicht werden kann. Im Gegensatz zu den zumeist in diesem Kontext verwendeten, sphärischen Koordinatensystemen stellen die stückweise planaren Polyeder euklidische Koordinaten bereit, die eine unerlässliche Grundlage vieler existierender Bildverarbeitungsansätze darstellen. Aus der Struktur dieser Koordinatensysteme, die nicht nur eine einzelne Menge verschiedener Teilflächen umfassen, sondern sich auch über mehrere Auflösungsebenen erstrecken können, leitet sich die Bezeichnung *Multi*-Mosaikbilder ab.

Im nachfolgenden Unterkapitel 6.1 werden zunächst die polyedrischen Koordinatensysteme motiviert und ihre geometrischen Grundlagen sowie die Handhabung in der Praxis beschrieben. Dabei lässt sich durch die Organisation innerhalb einer Auflösungshierarchie auch eine weitgehend verlustfreie Repräsentation unterschiedlich skalierter Bilddaten erreichen (Unterkap. 6.2). Bei der Implementierung des Konzepts ist einerseits zu berücksichtigen, dass die Repräsentation der Bilder, trotz der im Ansatz der Mosaikbilder inhärenten Datenreduktion, mit einem großen Speicheraufwand verbunden sein kann (Abschnitt 6.3). Andererseits ergeben sich auch aus der angestrebten Online-Berechnung zusätzliche Anforderungen (Abschnitt 6.4). Unterkapitel 6.5 stellt Resultate aus der praktischen Anwendung der Mosaikbilder zur Diskussion, wobei insbesondere das mit einer Online-Berechnung verbundene Risiko von Registrierungsfehlern nochmals aufgegriffen wird (vgl. auch Abschnitt 3.5.2).

## 6.1 Koordinatensysteme auf der Basis von Polyedern

Die Berechnung von Mosaikbildern aus Bildfolgen aktiver Kameras gründet auf der Wahl eines geeigneten Referenzkoordinatensystems zur Registrierung und Projektion der Bilddaten (vgl. Unterkap. 1.1). Es leitet sich primär aus den Freiheitsgraden der eingesetzten Kamera und der Struktur der Szene ab, wird aber auch durch die Handhabbarkeit im Hinblick auf die Integration neuer Daten und den Datenzugriff, sowie den implizierten Verwendungszweck der Darstellung maßgeblich beeinflusst. Auf Basis dieser Kriterien wurden den Multi-Mosaikbildern in der vorliegenden Arbeit polyedrische Referenzkoordinatensysteme zu Grunde gelegt. Im nachfolgenden Abschnitt 6.1.1 wird diese Wahl zunächst begründet, bevor im Anschluss die geometrischen Charakteristika der polyedrischen Koordinatensysteme und ihre Handhabung in der Praxis beschrieben werden.

### 6.1.1 Motivation

Zur Aufnahme von Bildfolgen finden in der vorliegenden Arbeit ausschließlich stationäre, rotierende und zoomende Kameras Anwendung, deren Bewegungen durch ein projektives Abbildungsmodell beschrieben werden können (Abschnitt 2.3). Die Bilder einer sol-

chen Sequenz sind damit über Homographien miteinander verknüpft. Im Rahmen einer Mosaikbildberechnung werden die Parameter dieser Homographien aus den Bilddaten rekonstruiert, um die Bilder zueinander in Beziehung setzen und in ein gemeinsames Referenzkoordinatensystem überführen zu können. Das Bewegungsmodell erlaubt dabei grundsätzlich Rotationen um beliebige Winkel in allen drei Raumrichtungen, so dass auch das Referenzkoordinatensystem diesen, in allen Richtungen 360° umfassenden Sichtbereich vollständig repräsentieren muss. Im Hinblick auf die angestrebte Unterstützung bestehender Bildverarbeitungsansätze ist es darüber hinaus unerlässlich, euklidische Koordinaten bereitzustellen und eine ausreichende Qualität der Multi-Mosaikbilder zu garantieren. Letzteres bedingt insbesondere, geometrische Verzerrungen bei der Überführung der Bilddaten in das Referenzkoordinatensystem weitestgehend zu vermeiden.

Geometrische Verzerrungen umfassen in diesem Kontext ausschließlich Veränderungen innerhalb der topologischen 2D-Anordnung von Bildpunkten. Projektive Verzerrungen aufgrund perspektivischer Effekte, wie sie etwa durch Differenzen in der räumlichen Tiefe von 3D-Szenenpunkten bei der Bildaufnahme selbst hervorgerufen werden können, und dafür existierende Korrekturansätze gehen über den Rahmen der vorliegenden Arbeit hinaus (für weitere Details hierzu s. z.B. [Har00], Kap. 1, oder auch [Cri02]).

Im Folgenden werden verschiedene Ansätze zur Wahl von Referenzkoordinatensystemen für Mosaikbilder aus Bildfolgen stationärer, rotierender Kameras vorgestellt. Vorrangig erfolgt dabei ein Vergleich der Eigenschaften planarer, zylindrischer bzw. sphärischer sowie polyedrischer Koordinatensysteme, verbunden mit einer detaillierten Analyse der Eignung dieser Systeme als Grundlage für die Multi-Mosaikbilder in dieser Arbeit.

## Planare Koordinatensysteme

Der einfachste und in der Literatur weit verbreitete Ansatz zur Wahl des Referenzkoordinatensystems eines Mosaikbildes besteht in der Verwendung einer einzelnen euklidischen Bildebene, auf die die registrierten Bilder projiziert werden [Man96, Dav98, Még99, Saw99]. Diese (virtuelle) Bildebene wird zumeist anhand der Bildebene eines ausgewählten Bildes der zu repräsentierenden Sequenz festgelegt, sie lässt sich grundsätzlich aber auch unabhängig von den Bilddaten definieren. Ein solches Koordinatensystem ist einerseits komfortabel zu handhaben und stellt andererseits auch die geforderte Unterstützung konventioneller Bildverarbeitungstechniken zur Verfügung. Verzerrungsfreie Projektionen der Bilddaten können mit diesem Ansatz allerdings nicht in allen Fällen gewährleistet werden. Insbesondere große Kamerarotationen, die signifikante Winkeldifferenzen zwischen der Orientierung der ausgewählten Referenzebene und den Ausrichtungen der Bildebenen einzelner Bilder der Folge bedingen, führen zu deutlichen geometrischen Verzerrungen bei der Projektion, wie etwa das Beispielmosaikbild in Abbildung 6.1 veranschaulicht. Die Verzerrungen nehmen dabei mit einem steigenden Rotationswinkel der Kamera exponentiell zu (vgl. Abb. 6.2). Ferner können Bilder (oder zumindest Teilbereiche von diesen), die mit Rotationswinkeln von mehr als ±90° relativ zur Referenzebene aufgenommen wurden, nur durch das Kamerazentrum hindurch abgebildet

(a) Bild 1 (b) Bild 16 (c) Bild 31 (d) Bild 41

**Abbildung 6.1:** Verzerrungen bei der Projektion eines Mosaikbildes auf eine einzelne Ebene, die durch das erste Bild der Folge festgelegt wurde: die Bildsequenz (41 Bilder, s. Beispiele links) umfasst einen weitwinkligen Szenenausschnitt, so dass die Ebene keine adäquate Projektionsbasis darstellt und die Szene somit lediglich verzerrt wiedergegeben werden kann (rechts).

werden, was neben den Verzerrungen zusätzlich zu einer unerwünschten Spiegelung der Daten führt. Dies bedingt somit eine Beschränkung des zulässigen Sichtbereichs der rotierenden Kamera auf maximal 180°. Eine einzelne Ebene erscheint damit, bezogen auf die Zielsetzungen des visuellen Speichers, die sowohl eine vollständige Repräsentation des Sichtbereichs einer rotierenden Kamera als auch eine verzerrungsfreie Darstellung der Daten umfassen, als Koordinatenreferenz für die Multi-Mosaikbilder ungeeignet.

Als Alternative wird in [Bur94] und [Sze96] eine Erweiterung des Referenzkoordinatensystems für Mosaikbilder von einer einzelnen Ebene zu einer Menge mehrerer Ebenen vorgeschlagen. In beiden Arbeiten erfolgt während der Mosaikbildberechnung eine dynamische Anpassung des Referenzkoordinatensystems an die Orientierung der jeweils aktuellen Bilddaten, so dass eine Zerlegung des Mosaikbildes in kleinere Teilausschnitte

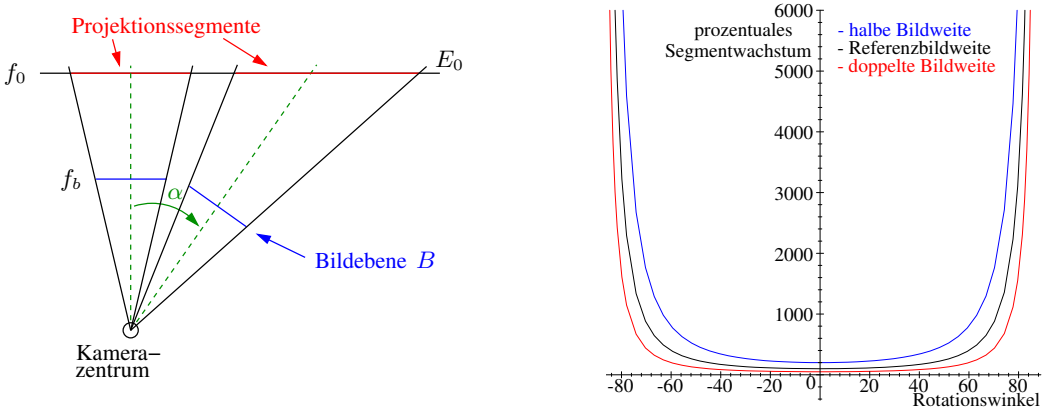



**Abbildung 6.2:** Theoretische Betrachtung auftretender Verzerrungen (zur Vereinfachung in 2D) bei der Projektion von Daten einer Bildebene $B$ auf eine (Referenz-)Ebene $E_0$: je größer der Rotationswinkel $\alpha$ wird und je größer auch die Differenz in den Bildweiten $f_0$ und $f_b$ der beiden Ebenen ist, desto stärkere Verzerrungen resultieren. Die Kurven in der rechten Grafik, die das damit korrespondierende Wachstum der Projektionssegmente skizzieren, implizieren eine signifikante Zunahme der Verzerrungen sobald der Rotationswinkel $\alpha$ etwa 50° überschreitet.

mit jeweils unterschiedlich ausgerichteten Referenzebenen resultiert. Allerdings werden die räumlichen Zusammenhänge zwischen den einzelnen Ebenen bei diesen Ansätzen nur implizit repräsentiert und sind damit nicht abfragbar. Falls jedoch die grundsätzliche räumliche (nicht allein projektive) Orientierung einzelner Bildebenen zueinander bekannt ist (z.B. aufgrund spezifischer Bewegungsmuster der eingesetzten Kamera), so kann dies einerseits etwa bei der Verifikation geschätzter Homographieparameter hilfreich sein, und andererseits eine spätere Extraktion ikonischer Daten von interessanten Szenenausschnitten vereinfachen (vgl. Kap. 8). Eine Rekonstruktion solcher Informationen ist zwar grundsätzlich auch zu einem späteren Zeitpunkt auf Basis der Bilddaten selbst möglich, dies bedingt jedoch zumeist einen hohen Aufwand (z.B. [Saw98]).

### Zylindrische & sphärische Koordinatensysteme

Die vorstehenden Ausführungen implizieren eine direkte Berücksichtigung bekannter räumlicher Zusammenhänge zwischen den Ebenen der einzelnen Bilder einer Folge schon bei der initialen Festlegung eines geeigneten Referenzkoordinatensystems für Mosaikbilder. Eine aufgrund dessen im Kontext stationärer, rotierender Kameras naheliegende Wahl für die Koordinaten stellen damit zylindrische bzw. sphärische Koordinatensysteme dar. Die Ebenen der einzelnen Bilder einer Sequenz, die mit einer solchen Kamera aufgenommen wurden, liegen tangential zu einer um das optische Zentrum der Kamera angeordneten Kugel, deren Radius sich aus der Bildweite der Eingangsbilder ableitet.[2] Eine Projektion der Daten auf diese Kugel konserviert damit die geometrischen Relationen innerhalb der Bilddaten und führt auf eine verzerrungsfreie Darstellung des vollständigen Sichtbereichs einer rotierenden Kamera. Derartige Koordinatensysteme finden insbesondere im Gebiet des Image Based Rendering als Teilbereich der Computergrafik Anwendung (vgl. Abschnitt 1.1, S. 5). Im Hinblick auf eine Verwendung im Rahmen des visuellen Speichers weisen allerdings auch diese Systeme große Nachteile auf.

Grundsätzlich sind euklidische Koordinatensysteme über spezifische Eigenschaften definiert (z.B. die Erhaltung der Kollinearität von Punkten bei einer Abbildung in diese Systeme), die die Grundlage vieler heute gängiger Bildverarbeitungsansätze bilden. Zylindrische und sphärische Koordinatensysteme erfüllen diese Voraussetzungen nicht und ihre Verwendung im Rahmen des visuellen Speichers würde damit eine Entwicklung neuer Analyseverfahren erzwingen. Darüber hinaus ist eine explizite, pixelweise Repräsentation von Mosaikbildern auf Basis zylindrischer oder sphärischer Koordinatensysteme schwierig.

Zylindrische Koordinatensysteme lassen sich zwar verhältnismäßig einfach in planare Darstellungen überführen (z.B. [Bis95]), die sich auch in Form konventioneller Digitalbilder abspeichern lassen und damit im Grundsatz eine Online-Integration neuer Daten und die Anwendung existierender Bildverarbeitungsalgorithmen unterstützen. Derartige Darstellungen beinhalten jedoch im Allgemeinen signifikante Verzerrungen der repräsen-

[2] Zur Vereinfachung wird hier zunächst eine konstante Bildweite innerhalb der Bildfolge vorausgesetzt, Details zum Umgang mit variierenden Bildweiten folgen in Abschnitt 6.2.

tierten Bildinformationen, die einer weiteren Verarbeitung der Daten entgegenstehen. Außerdem beschränken Zylinder die zulässigen Bewegungen der Kamera auf Rotationen um eine einzelne Achse, so dass sie den Anforderungen der vorliegenden Arbeit nicht gerecht werden können. Eine Darstellung von Mosaikbildern auf Basis sphärischer Koordinaten umfasst im Gegensatz dazu zwar den vollständigen Sichtbereich einer rotierenden Kamera, ihr Einsatz ist jedoch mit zusätzlichen Schwierigkeiten verbunden. Sphärische Koordinatensysteme lassen sich nicht direkt in planare Darstellungen überführen, sondern erfordern alternative Herangehensweisen. Insbesondere in der Computergrafik gibt es dazu verschiedene Ansätze.

Ein wichtiges Ziel in der Computergrafik besteht in der Entwicklung von Verfahren zur Berechnung und Darstellung (Rendering) neuer Ansichten aus Bilddaten, die an verschiedenen Standorten in einer Szene aufgenommen wurden. Die Daten eines einzelnes Standortes kodieren dabei eine Funktion, die spezifische Blickrichtungen auf zugehörige Farbwerte abbildet (vgl. auch *Light Fields*, z.B. in [Sla02], S. 511ff.). Zur adäquaten Repräsentation dieser Funktionen sind Kugeln als Projektionsziel für die ikonischen Informationen am besten geeignet. Aufgrund der mit einer expliziten Darstellung von Kugeln verbundenen Schwierigkeiten liegen einer solchen Modellierung jedoch zumeist stückweise planare 3D-Körper zu Grunde. Sie approximieren die Kugeln lediglich, bedingen aber eine bessere Handhabbarkeit und einen vereinfachten Datenzugriff [Bis95].

Greene schlug bereits 1986 die Verwendung von Würfeln zur Darstellung dieser so genannten „Environment Maps“ vor, die eine gute Ausgangsbasis zum Rendern beliebiger Szenen bilden [Gre86]. Die Maps werden zumeist in zwei Schritten erzeugt, wobei zunächst eine Offline-Registrierung der zu Grunde liegenden Bildsequenzen erfolgt, bevor die Daten direkt auf die Projektionskörper übertragen werden. Eine spätere Integration neuer Daten ist dabei im Allgemeinen nicht vorgesehen, da in der Computergrafik primär eine zeitlich konstante Modellierung statischer Szenen angestrebt wird.

Auch Shum und Szeliski beschreiben ein Verfahren zur Generierung von Environment Maps auf der Grundlage von Mosaikbildern, das ebenfalls auf einer zweistufigen Strategie basiert [Shu00]. In einem ersten Schritt werden dabei die Mosaikbilder berechnet, die anschließend auf konvexe, stückweise planare 3D-Grundkörper projiziert werden können. Die Berechnung der Mosaikbilder gründet dabei auf einer impliziten Repräsentation sphärischer Koordinaten. Für jedes Bild einer Folge wird dazu eine Transformation ermittelt, die seine Position auf einer um das optische Kamerazentrum angeordneten Kugel kodiert, und die den Ausgangspunkt für eine spätere Erzeugung neuer Ansichten bildet. Diese Darstellung impliziert jedoch, dass die vollständigen Bildfolgen und alle zur Projektion der einzelnen Bilder geschätzten Transformationen gespeichert werden müssen. Eine solche Vorgehensweise steht damit in direktem Widerspruch zur der in dieser Arbeit angestrebten, speichereffizienten Berechnung der Mosaikbilder, die insbesondere ihre Verwendung in interaktiven Systemen mit beschränkten Ressourcen ermöglichen soll.

Neben dem Datenvolumen der Repräsentation ist in dem skizzierten Ansatz auch die Online-Verarbeitung von Bilddaten mit Schwierigkeiten verbunden. Shum und Szeliski stellen zwar neben einer globalen Registrierung der Bildsequenzen zusätzlich einen Algo-

rithmus zur Online-Generierung der Mosaikbilder vor, dabei wird jedoch in jedem Schritt zunächst explizit ein herkömmliches Bild als Referenz für die Registrierung neuer Daten erzeugt. Darüber hinaus kann durch die Online-Berechnung zwar die Menge der registrierten Bilder sukzessive erweitert werden, eine Übernahme der neuen Informationen in eine bereits zuvor berechnete Environment Map, d.h. eine Online-Aktualisierung der eigentlichen Mosaikbilder, erfordert jedoch deren vollständige Neuberechnung. Daraus lässt sich ersehen, dass eine vereinheitlichte Handhabung von Mosaikbilddaten bei der Registrierung und Repräsentation insbesondere für Online-Verfahren von hoher Bedeutung ist. Euklidische Koordinaten erlauben dabei, im Gegensatz zu den vorstehenden Ansätzen, auch eine direkte Anwendung existierender Bildverarbeitungsalgorithmen.

### Polyedrische Koordinatensysteme

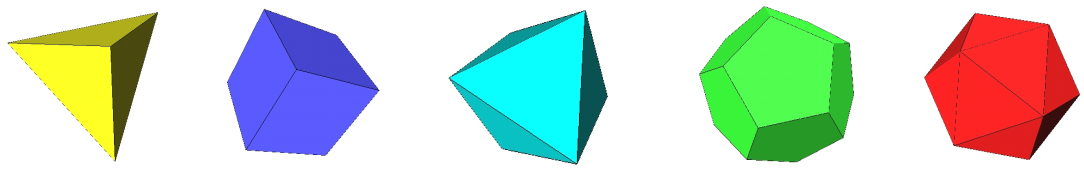
Eine Analyse der spezifischen Vor- und Nachteile planarer und sphärischer Koordinatensysteme im Hinblick auf eine adäquate Repräsentation der Bilddaten stationärer, rotierender Kameras legt für die Wahl geeigneter Referenzkoordinatensysteme für die Multi-Mosaikbilder einen Kompromiss zwischen beiden Ansätzen nahe. Unter anderem motiviert durch vergleichbare Vorgehensweisen zur Repräsentation von Bilddaten in der Computergrafik (s. vorhergehender Abschnitt) basieren die Referenzkoordinatensysteme in dieser Arbeit daher auf einer Menge euklidischer Ebenen, die regelmäßig um das optische Zentrum der Kamera angeordnet werden und eine Kugel stückweise planar approximieren. Auf diese Weise stehen trotz einer weitgehenden Vermeidung geometrischer Verzerrungen euklidische Koordinaten für eine direkte Anwendung existierender Bildanalyseverfahren auf die Mosaikdaten zur Verfügung.

Die globale Anordnung der Ebenen orientiert sich an *Polyedern*, wobei die einzelnen Ebenen über Homographien miteinander verknüpft sind (Unterkap. 2.3). Hieraus ergeben sich insbesondere Vorteile im Hinblick auf eine effiziente, einheitliche Verwaltung und Aktualisierung der repräsentierten Daten, wie sie in Abschnitt 6.1.3 beschrieben wird. Der gewählte Ansatz eröffnet darüber hinaus auch Perspektiven für eine flexiblere, dynamische Online-Modellierung von Environment Maps in der Computergrafik. Im nachfolgenden Abschnitt findet sich zunächst eine kurze Einführung in die Geometrie von Polyedern, auf deren Basis ein geeigneter 3D-Körper als Ausgangspunkt für das Referenzkoordinatensystem eines Multi-Mosaikbildes ausgewählt wird.

## 6.1.2 Geometrie von Polyedern

Ein Polyeder[3] ist formal durch einen dreidimensionalen Körper gegeben, der durch eine spezifische Anzahl aneinandergrenzender und zusammenhängender Teilflächen definiert wird. Als Ausgangspunkt zur Festlegung des Referenzkoordinatensystems eines Multi-Mosaikbildes lassen sich grundsätzlich beliebige Polyeder auswählen (Übersichten über die große Vielzahl existierender Polyeder finden sich beispielsweise in [Pea78] oder [Mai03]). Im Hinblick auf die gewünschte, möglichst gute Approximation einer Kugel

[3] auch „Vielflächner“, abgeleitet von den griechischen Worten „poly“ (viel) und „hedra“ (Sitz, Fläche)



(a) Tetraeder (b) Hexaeder (c) Oktaeder (d) Dodekaeder (e) Ikosaeder

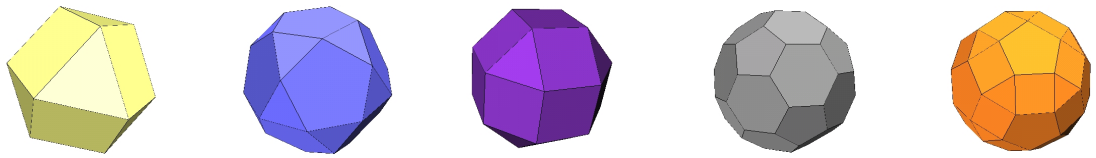
**Abbildung 6.3:** Die fünf existierenden, regelmäßigen, konvexen Polyeder (Platonische Körper).

liegt jedoch eine Beschränkung der Menge potenziell geeigneter Polyeder auf ausschließlich konvexe Körper nahe, da sich eine Kugel damit exakter approximieren lässt als es durch beliebige (nicht-konvexe) Anordnungen von Ebenen möglich wäre.

Die Beurteilung der Eignung spezifischer Polyeder als Basis für ein Referenzkoordinatensystem beruht vorrangig auf ihren geometrischen Eigenschaften, die sich nach [Mai03] unter anderem durch verschiedene Merkmale mathematisch beschreiben lassen. Im Wesentlichen sind dabei die Anzahl und Form der auftretenden Flächen (woraus direkt die Anzahl der vorhandenen Kanten folgt), die Winkel zwischen aneinandergrenzenden Flächen („Flächenwinkel“) und die Struktur der so genannten „Körperecken“ (Abb. 6.5) von Bedeutung. Als Körperecke wird ein Teilelement des Polyeders bezeichnet, das sich aus einer Ecke, in der mehrere Kanten aufeinandertreffen, und Teilstücken der angrenzenden Flächen zusammensetzt. Anhand dieser Merkmale (vgl. auch die Tabellen B.1 und B.2 im Anhang) lassen sich sechs Eigenschaften für Polyeder definieren, die für eine Einordnung in einzelne Klassen herangezogen werden können. Neben der Form und Anzahl der Grundflächen innerhalb einer Polyederfamilie dient dabei insbesondere die Anordnung der Flächen auf dem Körper als charakterisierende Eigenschaft.

Die Polyeder in der vorliegenden Arbeit sollten eine möglichst regelmäßige und tangentiale Anordnung der einzelnen Flächen relativ zu einer Kugel aufweisen, d.h. mit weitgehend identischen senkrechten Abständen zum Mittelpunkt des Körpers und damit zum Kamerazentrum. Diese Bedingung wird unter anderem von den „Platonischen Körpern“ (Abb. 6.3) und einer Teilmenge der „Archimedischen Körper“ (Abb. 6.4) erfüllt, die nach dem griech. Philosophen Platon ($427-348$ v. Chr.) bzw. dem Mathematiker und Ingenieur Archimedes ($285-212$ v. Chr.) benannt sind. Beide Mengen von Körpern lassen sich unter anderem in die Klasse der „gleicheckigen“ Polyeder einordnen, deren Elemente jeweils durch einen *eindeutigen* Typ von Körperecken charakterisiert sind. Platonische Körper, von denen nur fünf verschiedene Typen existieren, weisen zusätzlich nur eine Form vorkommender Teilflächen auf. Im Gegensatz dazu können Archimedische Körper aus verschiedenen Grundflächen bestehen, die jedoch an jeder Körperecke in derselben Sortierreihenfolge auftreten müssen.

Neben den vorstehenden Überlegungen spielen bei der Auswahl eines polyedrischen Grundkörpers für die Multi-Mosaikbilder auch die Anzahl gegebener Teilflächen, ihre Größe und die Flächenwinkel zwischen ihnen eine entscheidende Rolle. Grundsätzlich gilt, dass eine Kugel mit einer steigenden Anzahl von tangentialen Teilflächen zunehmend besser approximiert wird (Abb. 6.6). Die Flächenwinkel zwischen verschiedenen Ebenen nehmen dabei im Mittel zu (vgl. Tabelle B.2), so dass sich die Rotationswinkel zwischen



(a) Kuboktaeder (b) Ikosidodekaeder (c) Rhombenkuboktaeder (d) Ikosaederstumpf (e) Rhombenikosidodekaeder

**Abbildung 6.4:** Eine Auswahl Archimedischer Körper.

der Bildebene eines neuen Bildes und der jeweils am ähnlichsten ausgerichteten Teilfläche verringern. Als Folge wird der Einfluss von Verzerrungen bei der Projektion vermindert und die Qualität des Mosaikbildes damit erhöht (vgl. auch Abb. 6.2).

Allerdings steigt mit der Anzahl von Teilflächen auch die Anzahl von Unstetigkeiten innerhalb des Koordinatensystems, die aus den Verbindungskanten zwischen benachbarten Flächen resultieren. Sie müssen sowohl bei der Registrierung und Integration neuer Daten als auch beim Zugriff auf repräsentierte Informationen in geeigneter Weise berücksichtigt werden. Es erscheint offensichtlich, dass der Aufwand beim Umgang mit diesen Unstetigkeitsstellen direkt mit ihrer Anzahl korreliert ist und eine zu große Zahl an Teilflächen den Aufwand für ihre Verwaltung im Verhältnis zur erzielbaren Reduktion von Verzerrungen nicht mehr rechtfertigt. Darüber hinaus bedingt eine zunehmende Anzahl von Teilflächen bei einer weitgehend regelmäßigen Anordnung eine Verkleinerung der Einzelflächen. Auch dieser Effekt erschwert den Umgang mit dem Koordinatensystem und erhöht zudem das Risiko, dass Objekte innerhalb der Repräsentation auf verschiedene Teilflächen projiziert werden.

Unter Berücksichtigung aller vorstehend diskutierten Kriterien bilden Kuboktaeder, Dodekaeder und insbesondere Rhombenkuboktaeder eine gute Grundlage zur Festlegung des polyedrischen Referenzkoordinatensystems eines Multi-Mosaikbildes in dieser Arbeit. Alle drei Körper besitzen eine Anzahl von je 12 bis 26 Teilflächen, wobei die Einzelflächen eine adäquate Größe aufweisen und über geeignete Flächenwinkel miteinander verbunden sind. Das Rhombenkuboktaeder verfügt darüber hinaus über einen hohen Anteil quadratischer Grundflächen, die in der Praxis zwar keine zwingende Voraussetzung für eine Repräsentation von Bilddaten sind, im Hinblick auf eine effiziente Speicherverwaltung jedoch Vorteile bieten können (vgl. Abschnitt 6.3). Aus diesem Grund bildet es den Ausgangspunkt zur Festlegung der Referenzkoordinatensysteme für die Multi-Mosaikbilder.

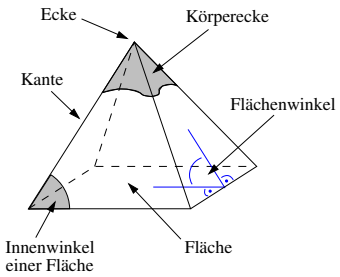



**Abbildung 6.5:** Geom. Merkmale eines Polyeders [Mai03].

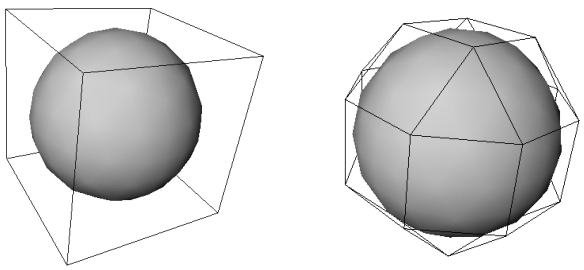


**Abbildung 6.6:** Approximation einer Kugel durch Polyeder mit unterschiedlichen Flächenanzahlen.

### 6.1.3 Praktische Handhabung der Koordinatensysteme

Bei einer Berechnung von Mosaikbildern unter ausschließlicher Betrachtung projektiver Abbildungszusammenhänge und der Verwendung einer einzelnen Ebene als Koordinatenreferenz ist eine explizite Festsetzung der räumlichen Orientierung der Ebene von untergeordneter Bedeutung. Die Qualität der Mosaikbilder und der Grad auftretender Verzerrungen hängt direkt von der gewählten Referenzebene ab (vgl. Abschnitt 6.1.1). Allerdings führt die weit verbreitete Vorgehensweise, das Referenzkoordinatensystem eines Mosaikbildes automatisch durch ein beliebiges Bild der zu repräsentierenden Sequenz festzulegen und nur einen beschränkten Szenenausschnitt darzustellen, zumeist implizit auf korrekte metrische Eigenschaften der Projektionsebene.

Ein polyedrisches Koordinatensystem bietet dagegen größere Spielräume bei der exakten Festlegung seiner geometrischen Parameter. Einzelne Werte üben einen maßgeblichen Einfluss auf die Qualität der späteren Repräsentation aus und müssen daher geeignet gewählt werden. Neben der grundsätzlichen, räumlichen Orientierung der Ebenen relativ zum Zentrum der Kamera sind dabei insbesondere das Pixelraster auf den Teilflächen des Polyeders sowie die damit unmittelbar in Zusammenhang stehende Skalierung des Körpers, d.h. der Abstand der einzelnen Flächen zum Kamerazentrum, von hoher Bedeutung. Eine fehlerhafte Skalierung etwa führt zu einer verzerrten Projektion der Daten, die beispielsweise bei der Berechnung eines 360°-Panoramas deutlich hervortreten kann (vgl. Abschnitt 6.5.2).

#### Ausrichtung und Skalierung

Die grundlegende Position des Referenzkoordinatensystems eines Multi-Mosaikbildes, die durch den geometrischen Mittelpunkt des ausgewählten 3D-Körpers spezifiziert wird, ist durch das optische Zentrum $O_{cam}$ der Kamera eindeutig festgelegt. Für die Orientierung des Körpers relativ zum Kamerazentrum, die vor Beginn der Mosaikbildberechnung explizit angegeben werden muss, folgen daraus jedoch zunächst keine Vorgaben. Sie ist prinzipiell frei wählbar. Im Hinblick auf eine komfortable Handhabung des Koordinatensystems dient daher im Rahmen der vorliegenden Arbeit das erste Bild $I_0$ einer Sequenz als Anhaltspunkt zur Festlegung der Orientierung und damit auch zur Initialisierung des Multi-Mosaikbildes insgesamt. Der 3D-Körper wird dabei so ausgerichtet, dass $I_0$ parallel zu deren Kanten auf eine der quadratischen Teilflächen des Rhombenkuboktaeders projiziert werden kann, auf die so genannte *Basisebene*. Diese Ebene muss damit achsenparallel zur xy-Ebene des 3D-Kamerakoordinatensystems bei der Aufnahme des ersten Bildes ausgerichtet werden und senkrecht auf dessen z-Achse (optische Achse) stehen (Abb. 6.7). Basierend auf dieser Ausrichtung lassen sich dann alle weiteren Bilder, die mit veränderten Kameraparametern aufgenommen werden, d.h. unter Rotationen des 3D-Kamerakoordinatensystems und Verschiebungen der Bildebene entlang der optischen Achse, direkt auf die jeweils korrespondierenden Teilflächen des Polyeders projizieren.

Die Festlegung einer geeigneten Skalierung des Polyeders hängt direkt von der Pixelrasterung der lokalen Bildebenen des polyedrischen Koordinatensystems ab. Auch dabei

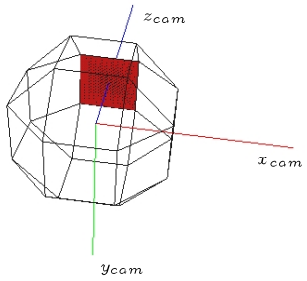



**Abbildung 6.7:** Ausrichtung eines Polyeders relativ zum 3D-Kamerakoordinatensystem des ersten Bildes einer Folge. Die gewählte Basisebene ist rot markiert.

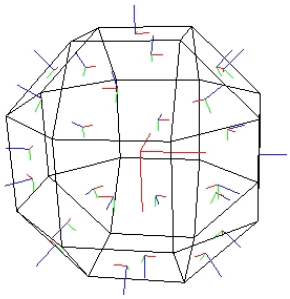


**Abbildung 6.8:** Festlegung der lokalen Bildkoordinaten auf den Flächen eines Rhombenkuboktaeders. Das initiale Kamerakoordinatensystem ist rot eingezeichnet.

gilt, dass die Größe der Pixel und damit die Skalierung des Polyeders im Grundsatz beliebig gewählt werden können. Sofern die Bildweite der Eingangsbilder und die Größe der Pixel auf den Teilflächen des Polyeders relativ zu den Bildpunkten der Eingangsbilder bekannt sind, lassen sich ohne Schwierigkeiten geeignete Homographien bestimmen, die eine Überführung der Bilder in das definierte Koordinatensystem ermöglichen. Im Hinblick auf eine möglichst adäquate Darstellung der Eingangsdaten empfiehlt es sich jedoch, die Pixelrasterung und Skalierung des 3D-Basiskörpers in Abhängigkeit von den konkreten Bilddaten festzulegen. Aus diesem Grund bildet auch hier wiederum das erste Bild einer Folge den Ausgangspunkt. Für die Teilflächen des Polyeders werden dabei Pixel zu Grunde gelegt, die in ihrer Größe den Bildpunkten der Eingangsbilder entsprechen. Darüber hinaus soll das erste Bild einer Sequenz ohne Skalierungen direkt auf die Basisebene des Grundkörpers projiziert werden können. Daraus folgt, dass der Abstand der Basisebene zum Kamerazentrum der Bildweite $f_0$ des ersten Bildes $I_0$ entsprechen muss, die folglich als Referenzbildweite interpretiert wird. Aus ihr leiten sich auch die Größen der einzelnen Teilflächen des polyedrischen Koordinatensystems ab.

Abschließend sei an dieser Stelle darauf verwiesen, dass die globale Lage der Kamera und damit auch die Positionen und Ausrichtungen der polyedrischen Ebenen innerhalb eines gegebenenfalls definierten 3D-Weltkoordinatensystems zunächst vernachlässigt werden können. Solange nur ein einzelnes Multi-Mosaikbild erzeugt und verarbeitet werden soll, gehen keine 3D-Weltkoordinaten in die Berechnungen ein. Die absolute räumliche Lage und Orientierung eines Multi-Mosaikbildes erlangt erst Bedeutung, wenn die Daten mehrerer Mosaikbilder, die an verschiedenen Positionen im Raum aufgenommen wurden, zueinander in Beziehung gesetzt werden sollen, wie dies im Rahmen eines Einsatzes des visuellen Speichers in mobilen Systemen sinnvoll sein kann (Kap. 8).

## Lokale 2D-Bildkoordinatensysteme

Mit jeder einzelnen Teilfläche des wie vorstehend definierten Referenzkoordinatensystems ist ein lokales, euklidisches 2D-Bildkoordinatensystem verknüpft. Analog zur 3D-Orientierung des vollständigen Systems kann auch die Festlegung dieser lokalen Koordinatensysteme im Grundsatz beliebig erfolgen. Zwischen den einzelnen Ebenen liegen projektive

Abbildungen vor (vgl. Unterkap. 2.3), die sich an die jeweils gewählte Orientierung der lokalen Koordinatenachsen anpassen lassen. Die oben skizzierte Wahl der Basisebene parallel zur xy-Ebene des 3D-Kamerakoordinatensystems bei der Aufnahme des ersten Bildes legt allerdings zumindest für die Basisebene die Definition eines zu diesem Koordinatensystem achsenparallelen 2D-Bildkoordinatensystems nahe. Es lässt sich leicht durch eine senkrechte Projektion der 3D-Achsen des initialen Kamerakoordinatensystems erzeugen, wobei die „upper left"-Konvention eingehalten wird (vgl. S. 14).

Die Koordinatensysteme der übrigen Teilflächen können wahlweise unabhängig von dem Koordinatensystem der Basisebene definiert, oder aber auch direkt aus diesem abgeleitet werden. Durch die bekannte dreidimensionale Geometrie des Körpers sind die räumlichen Anordnungen aller Teilflächen zueinander bekannt. Insbesondere sind 3D-Rotationsmatrizen gegeben, die die einzelnen Ebenen zueinander in Beziehung setzen. Mit ihrer Hilfe können die Koordinatenachsen des lokalen Bildkoordinatensystems der Basisebene über 3D-Rotationen geeignet transformiert und auf die anderen Teilflächen übertragen werden. Die aus dieser Vorgehensweise resultierenden, lokalen Bildkoordinatensysteme für das Rhombenkuboktaeder sind in Abbildung 6.8 skizziert. Der Ursprung der Koordinatensysteme liegt jeweils im geometrischen Mittelpunkt einer Fläche, der bei dem ausgewählten Polyeder auch den Auftreffpunkt des Abstandsvektors zum Kamerazentrum kennzeichnet (die Flächennormalen sind in Abb. 6.8 blau angedeutet).

## Initialisierung und interne Repräsentation

Die Festlegung der Orientierung und Skalierung des polyedrischen Grundkörpers sowie der lokalen Koordinatensysteme der einzelnen Flächen erfolgt im Rahmen einer Initialisierungsphase vor Beginn der eigentlichen Mosaikbildberechnung. Die einzelnen Teilflächen werden anschließend mit ihren spezifischen Parametern in einen ungerichteten Graphen eingetragen, der die 3D-Anordnung der Flächen zueinander im Raum *implizit* kodiert und in der Online-Phase einen effizienten Zugriff auf die Daten gewährleistet.

Jeder Knoten des Graphen entspricht einer Teilfläche, während die Kanten räumliche Nachbarschaften widerspiegeln (Abb. 6.9). Als Nachbarschaftskriterium dienen dabei gemeinsame Kanten zwischen aneinandergrenzenden Teilflächen. Jede Kante des Graphen trägt als Markierung die Homographie zwischen den betreffenden Flächen, die im Vorfeld der Mosaikberechnung über die 4-Punkt-Methode (vgl. S. 31) ermittelt werden kann. Projektive Abbildungen zwischen zwei beliebigen Flächen lassen sich später durch eine Konkatenation aller Homographien rekonstruieren, die auf einem frei wählbaren Pfad liegen, der die beiden Flächen miteinander verbindet (Abb. 6.9).

Die vorstehend skizzierte, implizite interne Repräsentation der polyedrischen Koordinatensysteme bildet die Basis für eine flexible Verwaltung der Koordinatensysteme hinsichtlich der angestrebten Online-Berechnung der Mosaikbilder. Darüber hinaus vereinfacht sie aber auch die praktische Handhabung der Koordinatensysteme insgesamt, in deren Rahmen insbesondere die unvermeidlichen Unstetigkeitsstellen an Übergängen zwischen einzelnen Flächen (vgl. S. 87) Probleme bereiten können.

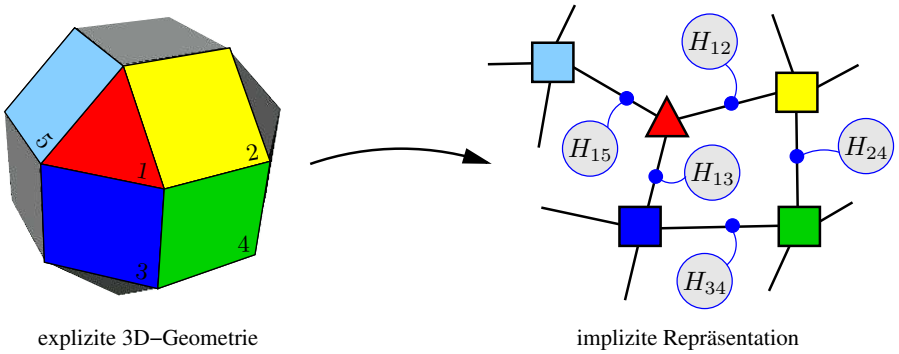
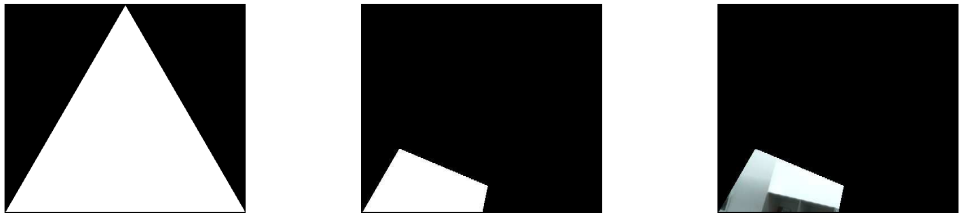



**Abbildung 6.9:** Skizze des Nachbarschaftsgraphen zur impliziten Repräsentation der 3D-Anordnung der einzelnen Teilflächen eines polyedrischen Koordinatensystems. Die Kanten des Graphen tragen als Markierung die Homographien $H_{ij}$ zwischen den jeweils benachbarten Flächen $E_i$ und $E_j$.

Werden wichtige Informationen aus einer Szene, etwa spezifische Objektdaten, direkt auf eine Unstetigkeitsstelle projiziert, so ist eine direkte Analyse der Daten mit Bildverarbeitungsalgorithmen schwierig. Die implizite Darstellung der Polyedergeometrie ermöglicht jedoch eine Verminderung dieses Effektes durch eine gezielte (Re-)Skalierung der einzelnen Flächen. Dazu werden die aus der gewählten Pixelrasterung auf den Teilflächen des Polyeders sowie der initialen Bildweite $f_0$ resultierenden Größen $A_i^{f_0}$ der einzelnen Flächen $E_i^{f_0}$ lediglich als untere Richtwerte interpretiert und die exakten Ausmaße der Flächen stattdessen in Abhängigkeit vom jeweiligen Anwendungskontext um einen spezifischen Prozentsatz (zumeist etwa $5-10\%$) größer gewählt. Diese Vorgehensweise führt zu einer partiellen Überlappung der einzelnen Flächen an Übergangsstellen, so dass Bilddaten in diesen Regionen jeweils auf beiden angrenzenden Flächen repräsentiert und damit im Allgemeinen vollständiger dargestellt werden können. Zwar resultieren dabei auch Redundanzen innerhalb der repräsentierten Informationen, die Verminderung des Einflusses der Unstetigkeitsstellen und die damit verbundene, verbesserte praktische Handhabbarkeit des Koordinatensystems rechtfertigen jedoch diesen Ansatz.

Jede Teilfläche kann zusammen mit ihrem lokalen 2D-Bildkoordinatensystem und den ihr zugeordneten Bilddaten als eigenständiges (Mosaik-)Bild gemäß der konventionellen Definition von Digitalbildern mit euklidischen Koordinaten und quadratischen Bildpunkten aufgefasst werden. Allerdings ist dabei zu berücksichtigen, dass nicht alle Teilflächen eines Multi-Mosaikbildes rechteckig sind, sondern vielmehr durch beliebige, konvexe Polygonzüge beschrieben werden können. Da die Bilddaten sich damit nicht unmittelbar als einzelne 2D-Matrix von Farbwerten interpretieren lassen, werden die einzelnen Bildebenen in der vorliegenden Arbeit durch ein *3-Tupel* von Matrizen repräsentiert. Alle drei Matrizen bilden das achsenparallele, umschließende Rechteck der Fläche ab. Ihre Dimensionen resultieren aus den maximal zulässigen Ausdehnungen der Bilddaten auf der Fläche entlang der Achsen des lokalen Bildkoordinatensystems. Die erste Matrix definiert eine boolesche *Validitätsmaske* (Abb. 6.10(a)), die nicht-definierte Teilbereiche einer Fläche kennzeichnet und beim Zugriff auf die Daten explizit berücksichtigt werden muss. In der zweiten Matrix werden undefinierte Pixel innerhalb des gültigen Definitionsbereichs der Teilfläche markiert, für die noch keine Bildinformationen vorliegen (Abb.



(a) Validitätsmaske (b) aktueller Datenbereich (c) gegebene Bilddaten

**Abbildung 6.10:** 3-Tupel aus Matrizen zur Repräsentation einer einzelnen Teilfläche eines Multi-Mosaikbildes: Alle drei Matrizen bilden das umschließende Rechteck der Fläche ab. Die Validitätsmaske (a) definiert dabei den gültigen Datenbereich auf der Fläche (weiss), während die beiden anderen Matrizen (b) und (c) die aktuell gegebenen Bilddaten kodieren.

6.10(b)), während die dritte (ggf. mehrschichtige[4]) Matrix die Bilddaten selbst vorhält (Abb. 6.10(c)).

Durch die skizzierte Vorgehensweise wird in Abhängigkeit von der Flächenform und der Wahl der lokalen Koordinatenachsen unter Umständen mehr Speicher für eine Teilfläche alloziert als die ausschließliche Repräsentation der Bilddaten erfordern würde. Bei Repräsentationsdatenstrukturen, die eine exaktere Beschreibung der realen Formen der Flächen ermöglichen (z.B. Quadtrees, Binary Space Partitioning Trees [Nay90]), treten diese Probleme nicht auf. Im Allgemeinen ist ihr Einsatz jedoch mit einem hohen Verwaltungsaufwand verbunden, der insbesondere im Hinblick auf den Datenzugriff deutlich mehr Nachteile mit sich bringt als sie durch den erhöhten Speicherverbrauch der angewandten Heuristik entstehen. Darüber hinaus lässt sich der Speicherbedarf des gewählten Ansatzes durch geeignete Mechanismen noch weiter reduzieren (Abschnitt 6.3).

## 6.2 Repräsentation verschiedener Auflösungsebenen

Im Allgemeinen sind bei der Betrachtung einer realen Szene nicht alle Ausschnitte von gleich hoher Bedeutung. In einzelnen Teilbereichen genügt oftmals ein grober Blick, um die wesentlichen Strukturen und relevanten Daten zu erfassen, während an anderen Stellen eine genauere Exploration notwendig ist, um auch feine, lokale Details entschlüsseln zu können, die für die korrekte Interpretation der visuellen Informationen unerlässlich sind. Das Wahrnehmungssystem des Menschen begegnet diesem variierenden Informationsgehalt in den visuellen Daten, die aus der Umwelt auf den Menschen einströmen, unter anderem durch eine gezielte Fokussierung auf aktuell relevante Informationen (Details s. Kap. 7). Auch für technische Systeme wird damit der Einsatz von aktiven Sensoren zur Akquisition visueller Daten mit variierender Granularität impliziert.

Bei der Bildaufnahme einer Szene mit handelsüblichen Kameras lassen sich unterschiedliche Detailgrade durch Variation der Zoomeinstellungen und damit der Bildweiten erzielen. Insbesondere in Kombination mit einer gezielten, algorithmischen Steuerung

[4]Zur Darstellung von Grauwertbildern genügt eine einzelne Schicht in dieser Matrix, für eine Speicherung von Farbbildern muss jedoch jeder Farbkanal (etwa R, G und B) separat repräsentiert werden.

des Kamerazooms lässt sich eine große Flexibilität bei der Lösung verschiedenster Fragestellungen erreichen (vgl. z.B. [Tor04]). Eine Repräsentation von Bildsequenzen, die aus derartigen Ansätzen resultieren, in einem ikonischen Speicher erfordert allerdings eine geeignete Berücksichtigung der unterschiedlichen Bildauflösungen.

Ein Mosaikbild wird üblicherweise mit einer festgewählten Auflösung assoziiert. Obwohl das projektive Abbildungsmodell (Unterkap. 2.3) grundsätzlich auch eine Registrierung von Bilddaten mit unterschiedlichen Skalierungen erlaubt, bringt eine Integration solcher Daten in ein einzelnes Mosaikbild mit fester Auflösung große Nachteile mit sich. Visuelle Daten, die eine höhere Auflösung aufweisen als das Mosaik, können nur unter einer künstlichen Verringerung ihrer Auflösung und damit verlustbehaftet integriert werden (Abb. 6.11(a)). Zur Integration niedriger aufgelöster Bilder ist dagegen eine Interpolation der Daten und eine Vervielfachung ihres Volumens erforderlich, die jedoch zu keinem zusätzlichen Informationsgewinn führt (Abb. 6.11(c)).


a) halbierte Auflösung b) Originalauflösung c) doppelte Auflösung durch Interpolation

**Abbildung 6.11:** Vergleich verschiedener Auflösungsstufen eines Bildes (b), die aus einer Reskalierung resultieren: Bei einer Auflösungsreduktion (a) gehen unter Umständen Detailinformationen verloren, während das Datenvolumen durch die Skalierung in eine höhere Auflösung zunimmt (c), ohne jedoch einen Informationsgewinn zu bewirken.

Das Konzept der Multi-Mosaikbilder begegnet diesen Effekten durch eine Hierarchie verschiedener Auflösungsebenen, wie sie konzeptuell auch in [Ira95] vorgeschlagen wurde. Ein Multi-Mosaikbild wird dazu aus mehreren, jeweils unterschiedlich skalierten Mengen von Projektionsebenen *(Projektionsinstanzen)* zusammengesetzt, die ineinander geschachtelt werden und sich damit schalenweise um das optische Zentrum der Kamera legen (Abb. 6.12). Die konkreten Skalierungen, zwischen denen jeweils einheitliche Abstände $\Delta f$ vorliegen (Abb. 6.12, rechts), korrespondieren dabei zu verschiedenen Bildweiten. Bei der Projektion von Bilddaten wird in Abhängigkeit von der Bildweite $f$ der aktuellen Eingangsdaten jeweils diejenige Projektionsinstanz $S^*_{f_j}$ als Basis ausgewählt, die die geringste Skalierung und damit Verfälschung der zu repräsentierenden Daten bedingt, d.h. deren Skalierung $f_j$ der Bildweite $f$ am nächsten kommt:

$$S^*_{f_j} = \underset{S_{f_i}}{\operatorname{argmin}} \; |f - f_i|.$$

Daraus folgt direkt, dass sich unerwünschte Reskalierungen der Eingangsdaten mit einer steigenden Anzahl von Projektionsinstanzen zunehmend besser vermeiden lassen. Allerdings korreliert eine hohe Anzahl von Instanzen auch mit einem hohen Verwaltungs- und Speicheraufwand, der zu der Exaktheit der Darstellung ins Verhältnis zu setzen ist.

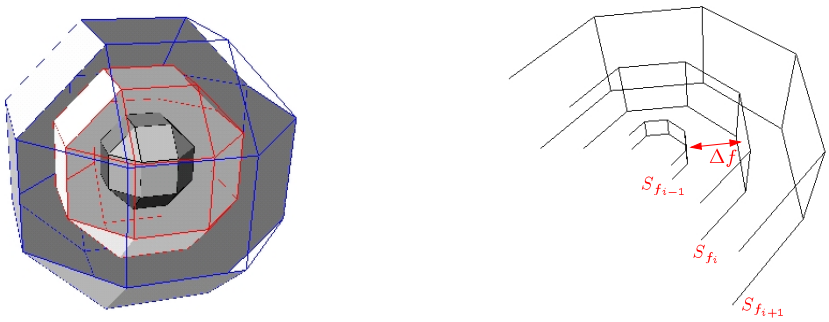



**Abbildung 6.12:** Repräsentation verschiedener Auflösungsebenen in Multi-Mosaikbildern durch eine Schachtelung verschiedener Projektionsinstanzen, zwischen denen jeweils ein Abstand von $\Delta f$ vorliegt.

Da der konkrete Wert für den Abstand $\Delta f$ zwischen den einzelnen Auflösungsebenen des Weiteren stark durch den gegebenen Anwendungskontext beeinflusst wird, kann er in der vorliegenden Implementierung extern vorgegeben werden, so dass sich eine flexible Anpassung an die Eigenschaften der jeweils zu repräsentierenden Daten realisieren lässt.[5]

Die skizzierte Herangehensweise zur adäquaten Repräsentation von Bilddaten mit unterschiedlichen Auflösungen steht einer Vielzahl von Ansätzen zur Repräsentation ikonischer Daten in verschiedenen Skalierungen in der Literatur gegenüber (s. etwa [Bur83a]). Zumeist besteht das Ziel der Verfahren dabei in einer Vereinfachung der Handhabung von Daten, die in einer *einheitlichen* Auflösungsstufe vorliegen, jedoch mit einem lokal variierenden Detailgrad bearbeitet werden sollen. Neben einer Vielzahl von Bildanalysealgorithmen, die innerhalb von Auflösungspyramiden arbeiten, profitieren dabei insbesondere Anwendungen im Bereich der interaktiven Editierung von Videosequenzen von solchen Darstellungen [Ber94, Fin96]. Die Daten werden in derartigen Ansätzen bei ihrer Überführung in die Repräsentationen zumeist durch eine *explizite* Skalierung in verschiedene Auflösungsstufen transformiert, so dass allen Ebenen dieser Darstellungen jeweils dieselbe Datenbasis zu Grunde liegt.

Im Kontext der vorliegenden Arbeit sind die in den einzelnen Auflösungsstufen repräsentierten Daten grundsätzlich unabhängig voneinander. Zwischen ihnen bestehen zwar projektive Abbildungszusammenhänge, mit deren Hilfe sich bei Bedarf Korrespondenzen zwischen den verschiedenen Instanzen etablieren lassen, die Daten selbst sind im Regelfall jedoch nicht identisch. Die entwickelte Repräsentationsstruktur zielt auf eine flexible, an die jeweiligen lokalen Gegebenheiten in einer Szene angepasste Darstellung der Daten, bei der eine Szene nicht vollständig in allen Auflösungen gespeichert wird. Vielmehr können für jeden Ausschnitt einer Szene gezielt die lokal relevanten Auflösungsebenen selektiert werden. Auf diese Weise wird es insbesondere möglich, verschieden aufgelöste Bilddaten eines einzelnen Szenenausschnittes zu speichern, die zu unterschiedlichen Zeitpunkten im Verlauf der Mosaikbildgenerierung aufgenommen wurden und damit unter Umständen auch unterschiedliche Hintergrundstrukturen (vgl. Kap. 5) des Szenenausschnittes repräsentieren können.

[5] Perspektivisch ist es auch denkbar, die einzelnen Abstände individuell und unabhängig voneinander festzulegen, um so die spezifischen Anforderungen verschiedener Anwendungsszenarien noch besser berücksichtigen zu können.

Die an die lokalen Gegebenheiten in einer Szene angepasste Darstellung der Daten bedingt, dass für einzelne Ausschnitte nur in spezifischen Auflösungsebenen auch Daten vorliegen, da keine explizite Reskalierung zur Übernahme der Daten in mehrere Auflösungsebenen erfolgt. Damit muss die konkrete Verfügbarkeit von Daten beim Zugriff jeweils explizit überprüft werden. Eine (verlustbehaftete) Übertragung neuer Daten auch in niedrigere Auflösungsstufen könnte zwar zur Verringerung dieser „Definitionslücken“ beitragen, im Rahmen der derzeitigen Implementierung wurde darauf jedoch verzichtet. Ein solches Vorgehen bedingt vorrangig eine Erhöhung des zu speichernden Datenvolumens, was jedoch unnötig erscheint, da spezifische Daten in niedrigeren Auflösungen bei Bedarf auch noch zu einem späteren Zeitpunkt online generiert werden können.

## 6.3 Speicherorganisation

Die Entwicklung von Algorithmen zur Bildregistrierung und damit auch zur Berechnung von Mosaikbildern hat ihren Ursprung einerseits in der Computergrafik, wo Mosaikbilder zu einer Erweiterung des Sichtfeldes einer Kamera und damit einer Vereinfachung des Umgangs mit realen Texturen beitragen können. Andererseits ist auch der Fortschritt in der Nachrichtentechnik und der daraus resultierende Bedarf an effizienten Kompressionsverfahren als Motivation für den Ansatz anzuführen. Dabei zeigt insbesondere das zweite Anwendungsfeld die Leistungsfähigkeit von Mosaikbildern im Hinblick auf eine effiziente Repräsentation ikonischer Daten auf. Die dem Ansatz inhärente Eliminierung redundanter Informationen erlaubt eine signifikante Reduktion des Datenvolumens von Bildsequenzen, so dass sich ihre Handhabung und Analyse stark vereinfacht. Allerdings darf trotz dieser Eigenschaften das Datenvolumen von Mosaikbildern im Verlauf ihrer Generierung nicht gänzlich außer Acht gelassen werden. Insbesondere das mit den Multi-Mosaikbildern in dieser Arbeit verbundene Ziel, Bildfolgen mit großen Sichtbereichen und variierenden Bildweiten zu repräsentieren, erfordert eine geeignete Speicherorganisation.

### Motivation und Konzept

Der Speicherbedarf eines konventionellen Mosaikbildes, dem eine einzelne Ebene als Projektionsbasis zu Grunde liegt, ist näherungsweise proportional zur Anzahl der enthaltenen Bildpunkte. Im Allgemeinen wird die Bildebene dabei nur in den Bereichen explizit repräsentiert, in denen auch Bilddaten verfügbar sind (zumeist anhand des umschließenden Rechtecks festgelegt). Bei der Integration neuer Daten erfolgt dann bei Bedarf eine dynamische Erweiterung der Projektionsfläche durch eine Allokation zusätzlichen Speichers, wobei sich die finalen Maße der Fläche erst nach Verarbeitung aller Bilder eindeutig bestimmen lassen. Im Gegensatz dazu kann der maximale Speicherbedarf von Multi-Mosaikbildern direkt aus den geometrischen Eigenschaften der zu Grunde liegenden Polyeder abgeschätzt werden. Er ist näherungsweise proportional zu der Gesamtsumme der Flächeninhalte aller Teilflächen und hängt damit direkt von deren Skalierung ab. Diese wiederum wird durch die Bildweite $f$ der zu speichernden Daten vorgegeben.

Die Repräsentation von ikonischen Daten in Multi-Mosaikbildern ist aufgrund des gegenüber einzelnen Projektionsebenen deutlich erweiterten Sichtfeldes mit einem erhöhten Speicheraufwand verbunden. Während konventionelle Mosaikbilder zumeist nur kleine Ausschnitte einer Szene geeignet widerspiegeln können (vgl. Abb. 6.1), erlauben Multi-Mosaikbilder die Darstellung des vollständigen Sichtbereichs einer stationären, rotierenden Kamera. Darüber hinaus bedingt auch die im vorhergehenden Abschnitt vorgestellte Repräsentation verschiedener Auflösungsstufen durch eine Schachtelung unterschiedlich skalierter Projektionsinstanzen einen weiteren, überproportionalen Anstieg des Datenvolumens der Mosaikbilder. Insgesamt empfiehlt sich damit für die Multi-Mosaikbilder eine partielle Repräsentation, in deren Rahmen nur für die Teilbereiche physikalischer Speicher alloziert wird, für die auch tatsächlich Daten gegeben sind. Ohne eine solche, effiziente Speichernutzung lassen sich die Multi-Mosaikbilder beispielsweise nicht in interaktiven mobilen Systemen einsetzen, die nur beschränkte Kapazitäten zur Verarbeitung und Speicherung von Daten zur Verfügung stellen können (vgl. auch Kap. 8).

Neben den rein technischen Aspekten wird eine derartige Vorgehensweise auch durch wahrnehmungspsychologische Beobachtungen gestützt. Wie zu Beginn von Abschnitt 6.2 bereits angedeutet wurde, wecken verschiedene Teilbereiche einer realen Szene im Allgemeinen unterschiedliches Interesse beim Betrachter. Die einzelnen Bereiche werden in der Regel in einer durch den lokalen „Interessantheitsgrad“ definierten Reihenfolge exploriert (vgl. auch Kap. 7). Dabei genügt in vielen Bereichen ein grober Blick, um alle wichtigen Informationen zu erfassen, und nur an wenigen Stellen ist eine detailliertere Analyse erforderlich. Hochaufgelöste visuelle Daten werden damit sehr selektiv und lokal begrenzt aufgenommen. Im Hinblick auf eine hierarchische Repräsentation von Bilddaten folgt daraus, dass zumeist keine Notwendigkeit besteht, den vollständigen Sichtbereich einer Kamera in der höchsten zur Verfügung stehenden Auflösung vorzuhalten. Vielmehr genügt für große Teile einer Szene eine grobe Darstellung in niedriger Auflösung, die an einzelnen Stellen gezielt um Daten in einer höheren Auflösung erweitert werden kann. Eine speichereffiziente, partielle Repräsentation von Multi-Mosaikbildern spiegelt damit auch die bei der Verarbeitung visueller Daten zu beobachtende, selektive Akquisition und zielgerichtete Fokussierung auf relevante Informationsquellen wider.

## Praktische Umsetzung

Die praktische Realisierung der partiellen Repräsentation von Multi-Mosaikbildern erfolgt durch die Implementierung einer dynamischen Speicherverwaltung. Sie erlaubt eine an die jeweils aktuelle Datenlage angepasste Speichernutzung und damit eine effiziente und ressourcenschonende Verwaltung der Mosaikbilder. Während sich die Implementierung derzeit auf eine dynamische *Allokation* von Speicher beschränkt, eröffnet der realisierte Ansatz perspektivisch Möglichkeiten, auch Mechanismen zur *Deallokation* von Ressourcen ohne Aufwand einzubinden. Diese Option ist insbesondere im Hinblick auf eine explizite Modellierung von Vergessen von hoher Bedeutung. Im Folgenden wird nun die implementierte Speicherverwaltung genauer beschrieben.

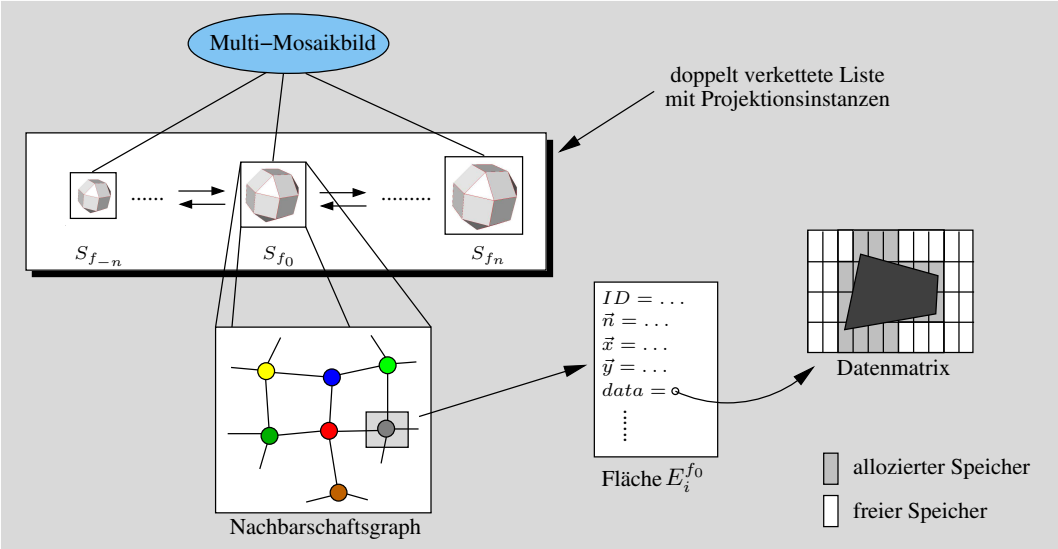



**Abbildung 6.13:** Skizze der implementierten Datenstrukturen zur speichereffizienten Verwaltung von Multi-Mosaikbildern: Sowohl für die Projektionsinstanzen $S_{f_i}$ wie auch für einzelne Flächen einer Instanz bzw. Teilbereiche von diesen wird physikalischer Speicher lediglich inkrementell und nach Bedarf alloziert, so dass der gesamte Speicherverbrauch des Multi-Mosaikbildes in direkter Abhängigkeit von den tatsächlich zu repräsentierenden Daten wächst.

Die grundlegende Struktur eines Multi-Mosaikbildes und aller enthaltenen Komponenten ist in Abbildung 6.13 skizziert. Zu Beginn einer Berechnung werden zunächst im Rahmen einer Systeminitialisierung die Teilkomponenten instanziiert, die zur Repräsentation von Bilddaten in der initial angenommenen Bildweite $f_0$ notwendig sind. Dazu wird eine Projektionsinstanz $S_{f_0}$ erzeugt, die durch den zuvor bereits skizzierten Nachbarschaftsgraphen repräsentiert wird (vgl. Abschnitt 6.1.3). Im Rahmen seiner Initialisierung werden sowohl die $N$ Flächen $E_i^{f_0}, i = 1 \ldots N$, der Instanz selbst wie auch die zwischen ihnen vorliegenden Homographien berechnet. Die Instanziierung der Flächen beschränkt sich dabei auf die korrekte Bestimmung ihrer geometrischen Parameter (z.B. Ausrichtung, Größe, Flächenform oder Definitionsbereich), während physikalischer Speicher für Bilddaten zu diesem Zeitpunkt noch nicht alloziert wird. Dies geschieht erst, wenn im weiteren Verlauf der Berechnungen Daten für den korrespondierenden Szenenausschnitt aufgenommen werden. Die Projektionsinstanz $S_{f_0}$ bildet anschließend das erste Element einer doppelt verketteten, sortierten Liste, die im weiteren Verlauf der Mosaikbildberechnungen sukzessive um zusätzliche Projektionsinstanzen $S_{f_s}$ ergänzt werden kann, sobald Bilddaten für von der initialen Bildweite $f_0$ signifikant abweichende Skalierungen $f_s$ anfallen. Durch diese inkrementelle Speicherallokation lässt sich der zur Repräsentation eines Multi-Mosaikbildes notwendige Speicher signifikant reduzieren, da das Datenvolumen der Mosaikbilder nun in direkter Abhängigkeit vom Umfang der tatsächlich zu repräsentierenden Daten wächst (Beispielabbildungen partiell repräsentierter Mosaikbilder finden sich in Abschnitt 6.5).

Auch mit dem vorstehend skizzierten Ansatz zur effizienten Speicherallokation kann in bestimmten Fällen noch immer ein signifikanter Anteil von Speicher reserviert werden,

für den keine Bilddaten vorliegen. Dieser Effekt ist insbesondere dann zu beobachten, wenn auf einzelnen Teilflächen nur für kleine Bereiche Bilddaten gegeben sind, während für den überwiegenden Teil der Fläche noch keine Daten vorliegen. Zwar werden in niedrigen Auflösungsstufen einzelne Teilflächen bei der Integration von Bilddaten zumeist zum Großteil ausgefüllt, in höheren Auflösungsebenen empfiehlt es sich jedoch, eine weitere Partitionierung vorzunehmen und nicht nur für vollständige Teilflächen Speicher zu allozieren. Dies resultiert aus der Tatsache, dass der Flächeninhalt der einzelnen Teilflächen mit einer höheren Auflösung zunimmt, während die aufgenommenen Bilder eine konstante Größe aufweisen. Als Folge nimmt das Verhältnis der Flächeninhalte von eingehenden Bildern zu denen der Projektionsflächen ab, so dass der Anteil überschüssigen Speichers bei einer Allokation vollständiger Teilflächen quadratisch wächst (vgl. Abb. 6.14). Hinzukommt, dass grundsätzlich große Teile einer Szene nicht in höheren Auflösungen aufgenommen werden und damit komplette Teilflächen in diesen Auflösungsstufen auch langfristig nur in seltenen Fällen vollständig mit Daten ausgefüllt werden.

Um der mit diesen Effekten verbundenen Allokation ungenutzten Speichers entgegenzuwirken, wird in der vorliegenden Arbeit eine zusätzliche Partitionierung der einzelnen Teilflächen bzw. der zugehörigen Bildmatrizen in Subzellen anhand eines regelmäßigen Rasters vorgenommen (s. Datenmatrix in Abb. 6.13). Für jede Zelle kann dabei unabhängig Speicher alloziert werden. Der Zugriff auf die Daten erfolgt durch Routinen, die die Partitionierung der Bildebene kapseln, wobei die Regelmäßigkeit des Rasters die direkte Adressierung jedes einzelnen Pixels innerhalb der Routinen erlaubt.

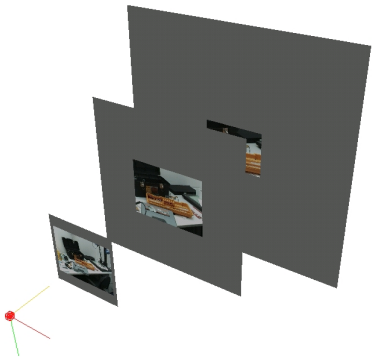


**Abbildung 6.14:** Verhältnis von alloziertem zu tatsächlich belegtem Speicher in den einzelnen Ebenen einer Auflösungspyramide bei der Integration von Bildern konstanter Größe: Die grauen Bereiche markieren noch undefinierte Ausschnitte der Flächen, die mit steigender Bildweite anteilmäßig zunehmen.

Bei der Auswahl eines Rasters für eine spezifische Teilfläche ist zu berücksichtigen, dass mit einer steigenden Anzahl von Subzellen der Mehrverbrauch von Speicher zunehmend besser reduziert werden kann, dies allerdings gleichzeitig einen erhöhten Verwaltungsaufwand bedingt. Die Anzahlen der Subzellen in horizontaler und vertikaler Richtung $s_w$ und $s_h$ für eine spezische Teilfläche mit einer Breite von $P_w$ und einer Höhe von $P_h$ Pixeln werden daher jeweils in Abhängigkeit von der erwarteten Größe $W_e \times H_e$ der zu integrierenden Bilder festgelegt. Im Detail findet für die Berechnung der beiden Werte dabei die nachfolgende Heuristik Anwendung:

$$s_w = \frac{P_w}{W_e} + 2 \quad \text{und} \quad s_h = \frac{P_h}{H_e} + 2.$$

Sie realisiert einen geeigneten Kompromiss zwischen einer Reduktion des Speicherbedarfs einerseits und einer in einem akzeptablen Verhältnis zu dieser Reduktion stehenden Handhabbarkeit der Datenstrukturen andererseits.

## 6.4 Online-Berechnung von Multi-Mosaikbildern

In der Einleitung der vorliegenden Arbeit wurden im Wesentlichen zwei Anforderungen an den zu entwickelnden, visuellen Speicher gestellt, die eine einfache Einbettung des Konzepts in interaktive Systeme ermöglichen sollen. Einerseits wird eine direkte Unterstützung konventioneller Bildverarbeitungsalgorithmen durch die Bereitstellung euklidischer Koordinatensysteme angestrebt, so dass die Daten des visuellen Speichers direkt in existierende Analyse- und Interpretationsmodule eingebunden werden können. Darüber hinaus sollen die entwickelten Datenstrukturen einen effizienten Zugriff auf die bis zum jeweiligen Zeitpunkt akquirierten Daten gewährleisten. Insbesondere dieser zweite Aspekt ist dabei im Hinblick auf die Einbindung des visuellen Speichers in interaktive Systeme von grundlegender Bedeutung. Da in derartigen Systemen unerwartete, zur Mosaikbildberechnung in Konkurrenz stehende Ereignisse eintreten können, die unmittelbare Reaktionen des Systems erforderlich machen (z.B. Nutzerinteraktionen, vgl. auch Kap. 8), darf die Mosaikbildgenerierung die Interaktivität der Systeme nicht beschneiden. Vor dem Hintergrund der oftmals beschränkten, und auch zumeist nur unter großen Schwierigkeiten erweiterbaren Speicher- und Rechenkapazitäten mobiler Systeme lässt sich daraus direkt ersehen, dass beispielsweise eine Zwischenpufferung mehrerer Bilder und ihre anschließende, simultane Bearbeitung im Widerspruch zu diesen Anforderungen steht. Der visuelle Speicher muss damit zwingend und vorrangig eine Online-Verarbeitung aufgenommener Bilddaten unterstützen.

In den vorausgegangenen Abschnitten wurde zunächst das grundlegende Konzept der Multi-Mosaikbilder eingeführt, das auf stückweise planaren Referenzkoordinatensystemen aus polyedrisch angeordneten Projektionsebenen basiert. Die daraus resultierende Repräsentationsdatenstruktur stellt einerseits die gewünschte Schnittstelle zu gängigen Bildverarbeitungsmodulen bereit. Darüber hinaus unterstützt sie aber auch – beispielsweise im Gegensatz zu sphärischen oder zylindrischen Koordinatensystemen (Abschnitt 6.1.1) – eine effiziente Online-Verarbeitung von Bilddaten, die im Folgenden vorgestellt wird. Das implementierte Verfahren durchläuft wiederholt die in den Kapiteln 3 und 4 dargestellten Phasen der Bildregistrierung und Integration, wobei im Kontext der Multi-Mosaikbilder insbesondere geeignete Referenzdaten für beide Schritte zu wählen sind. Darüber hinaus unterscheidet sich das Referenzkoordinatensystem eines Multi-Mosaikbildes von dem eines herkömmlichen Mosaiks dadurch, dass mehrere lokale 2D-Bildkoordinatensysteme vorhanden sind, die zwar über Homographien miteinander verknüpft sind, jedoch nicht als Einheit betrachtet werden können. Somit erfordert sowohl die Registrierung und Integration neuer Daten wie auch der Zugriff auf die Informationen eine explizite Berücksichtigung der zwischen den verschiedenen Teilflächen vorliegenden Unstetigkeitsstellen.

Der nachfolgende Abschnitt beschreibt zunächst den hierfür gewählten Ansatz, der im Wesentlichen die Verwendung einer zusätzlichen Bildebene umfasst, die so genannte *Fokus-Bildebene*. Sie ermöglicht einen effizienten Umgang mit der Multi-Mosaik-Datenstruktur, indem sie die Unstetigkeiten des Koordinatensystems maskiert. In Abschnitt 6.4.2 werden die Integration von neuen Bilddaten in die Multi-Mosaikbilder und

damit verbundene Veränderungen der Verfahren aus Kapitel 4 skizziert, bevor das Unterkapitel 6.4.3 abschließend auf Besonderheiten der Bildregistrierung verweist, die aus der Struktur der Multi-Mosaikbilder resultieren.

### 6.4.1 Fokus-Bildebene

Eine Grundvoraussetzung zur Registrierung eines Bildes relativ zu einem vorgegebenen Referenzkoordinatensystem ist durch das Vorhandensein geeigneter Referenzdaten gegeben, auf deren Basis sich die Parameter des zur Transformation benötigten Bewegungsmodells schätzen lassen. Dabei empfiehlt es sich, gemäß der Ausführungen in Unterkapitel 3.3, diese Referenzdaten aus dem bis zum aktuellen Zeitpunkt berechneten Mosaikbild zu gewinnen, um den Einfluss sich akkumulierender Registrierungsfehler zu vermindern. Bei der Extraktion entsprechender Referenzdaten aus einem Multi-Mosaikbild muss dabei die Topologie des zu Grunde liegenden Referenzkoordinatensystems berücksichtigt werden, so dass sich zunächst zwei verschiedene Vorgehensweisen anbieten.

Einerseits ist es möglich, die Daten einer einzelnen Teilfläche des Multi-Mosaikbildes als Grundlage zur Registrierung neuer Bilder zu verwenden. Die Orientierung dieser Teilfläche sollte dabei eine minimale Differenz zur Ausrichtung der Bildebene des neuen Bildes aufweisen, um Verzerrungen im Rahmen der Parameterschätzung weitestgehend auszuschließen. Allerdings bleiben bei dieser Vorgehensweise relevante, ikonische Daten auf benachbarten Teilflächen unberücksichtigt. Die zur Vereinfachung des Datenzugriffs vorgeschlagene, partielle Überlappung benachbarter Teilflächen (Abschnitt 6.1.3) vermindert zwar diesen Effekt, alle relevanten Daten werden aber dennoch nur in wenigen Ausnahmefällen vollständig auf einer einzelnen Teilfläche repräsentiert sein.

Ein zweiter möglicher Ansatz zur Generierung geeigneter Referenzdaten besteht daher darin, die Daten aller Teilflächen eines Multi-Mosaikbildes zu fusionieren, die für die aktuelle Registrierung von Bedeutung sind. Dabei kann beispielsweise die Teilfläche, deren Orientierung zu der des zu registrierenden Bildes am ähnlichsten ist, als Zielebene dienen, auf die ergänzend die ikonischen Daten benachbarter Teilflächen projiziert werden. Darüber hinaus besteht aber auch die Möglichkeit, explizit eine neue, der geschätzten Orientierung des aktuellen Bildes besser entsprechende Ebene zu erzeugen. Diese Vorgehensweise gewährleistet, dass die zur Registrierung des neuen Bildes zur Verfügung stehende Basis an ikonischen Daten vollständig ausgeschöpft wird. Allerdings ist sie mit einem hohen Aufwand verbunden, da für jedes neue Bild Referenzdaten durch Transformation und Projektion erzeugt, und dabei jeweils mehrere Teilflächen des Multi-Mosaikbildes im Hinblick auf das Vorhandensein relevanter Daten überprüft werden müssen.

Die bei der Aufnahme einer Szene durchgeführten Kamerabewegungen sind im Allgemeinen stetig und beinhalten zumeist keine großen Sprünge zwischen verschiedenen Punkten im Raum. Damit weisen aufeinander folgende Bilder einer Sequenz einen hohen Überlappungsgrad auf, so dass auch innerhalb der für eine Bildregistrierung aus den Mosaikbildern zu extrahierenden Referenzdaten große Überschneidungen vorliegen. Um die dadurch notwendige, wiederholte Extraktion nahezu identischer Referenzdaten

aus dem Multi-Mosaikbild zu vermeiden, wird im Kontext dieser Arbeit eine zusätzliche Bildebene *(Fokus-Bildebene)* eingeführt, die als Zwischenspeicher für aktuell relevante Bilddaten dient und damit gleichsam eine Schnittstelle zwischen eingehenden Bilddaten und dem Multi-Mosaikbild bildet. Neue Bilddaten werden relativ zu dieser Ebene registriert und dort auch zunächst integriert, so dass die eigentliche Struktur des Referenzkoordinatensystems maskiert wird und nicht in jedem Schritt ein direkter Zugriff auf einzelne Teilflächen des Multi-Mosaikbildes erforderlich ist. Auf diese Weise ergibt sich eine zweistufige Repräsentationsstruktur, die die Daten implizit nach ihrer aktuellen Relevanz „klassifiziert“. Wichtige Daten lassen sich direkt über die Fokus-Bildebene abfragen, während nicht unmittelbar interessante Informationen für spätere Zugriffe in dem darunter liegenden Multi-Mosaikbild gespeichert werden.

Die Fokus-Bildebene ist durch eine tangential an das polyedrische Koordinatensystem angelegte Bildebene gegeben, die die Unstetigkeitsstellen zwischen einzelnen Teilflächen des Referenzkoordinatensystems überdeckt (Abb. 6.15). Ihre Ausrichtung orientiert sich dabei an der jeweils aktuellen Orientierung des Kamerakoordinatensystems und ist damit unabhängig von den geometrischen Parametern der einzelnen Teilflächen des Polyeders. Die initiale Lage und Orientierung leitet sich aus der anfänglichen Ausrichtung der Kamera ab, wobei der Abstand zum Kamerazentrum der Bildweite $f_0$ des ersten Bildes und somit auch der Skalierung der zu Beginn ausgewählten Projektionsinstanz $S_{f_0}$ des Multi-Mosaikbildes entspricht.

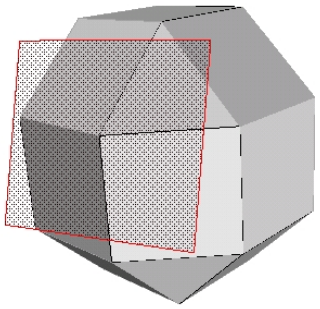


**Abbildung 6.15:** Skizze der Fokus-Bildebene: die Ebene wird tangential an den polyedrischen Grundkörper angelegt, so dass sie die Unstetigkeiten zwischen benachbarten Teilflächen innerhalb des Referenzkoordinatensystems überdeckt.

Je weiter sich die Kamera durch anschließende Bewegungen von dieser Grundorientierung entfernt, desto stärkere Verzerrungen sind in Abhängigkeit von der festgewählten Größe der Fokus-Bildebene (sie wird zumeist rund zweieinhalb mal größer gewählt als die Eingangsbilder) bei der Registrierung und Integration neuer Daten zu beobachten (vgl. Abschnitt 6.1.1). Ebenso steigt dabei auch das Risiko, dass neue Daten außerhalb des gültigen Projektionsbereichs der Fokus-Bildebene liegen. Um diesen Effekten entgegenzuwirken, erfolgt regelmäßig eine Neuausrichtung der Fokus-Bildebene gemäß den veränderten intrinsischen und extrinsischen Kameraparametern, d.h. insbesondere eine Anpassung ihrer 3D-Position im Raum sowie der Orientierung und Entfernung relativ zum Kamerazentrum.

Die Zeitpunkte dieser Aktualisierungen werden aufgrund einer stetigen Überwachung der Bewegungsparameter der Kamera nach jeder Integration eines neuen Bildes bestimmt. Durch eine Analyse des Integrationsbereichs des letzten Bildes auf der Fokus-Bildebene lassen sich signifikante Veränderungen in der Kameraposition und -orientierung feststellen. Sobald das umschließende Rechteck des aktuellen Integrationsbereichs den durch einen spezifischen Schwellwert $\theta_d$ festgelegten Mindestabstand zum Rand der Fokus-Bildebene unterschreitet, wird eine Aktualisierung der Ebenenposition

und -orientierung initiiert. Geeignete Werte für $\theta_d$ lassen sich dabei beispielsweise durch eine Schätzung des zu erwartenden Wachstums des Datenbereichs pro Integrationsschritt ermitteln. In der Praxis hat sich ein Wert von etwa $10-15\%$ der Flächenbreite bzw. -höhe bewährt. Die Überprüfung der Veränderungen in den Bildweiten erfolgt durch eine Auswertung der paarweisen Differenzen aufeinander folgender Bilder, wobei die Daten der in Unterkapitel 2.4 vorgestellten Verfahren zur Kamerakalibrierung die Grundlage bilden. Die Basis zur Aktualisierung der Fokus-Bildebene bildet ein dreistufiges Verfahren, das in Abbildung 6.16 skizziert ist. Dabei werden zunächst die aktuell auf der Fokus-Bildebene repräsentierten ikonischen Daten auf die korrespondierenden Teilflächen des Multi-Mosaikbildes übertragen (Details hierzu finden sich im nachfolgenden Abschnitt 6.4.2). In einem zweiten Schritt erfolgt anschließend die Korrektur der Lage und Orientierung der Fokus-Bildebene und/oder ihres Abstandes zum Kamerazentrum. Die neue Position der Ebene in 3D-Raumkoordinaten resultiert dabei aus einer quadratischen Extrapolation der letzten drei Kamerapositionen auf der aktuellen Fokus-Bildebene, wobei durch diese prädiktive Vorgehensweise die Anzahl notwendiger Aktualisierungen minimal gehalten werden soll.

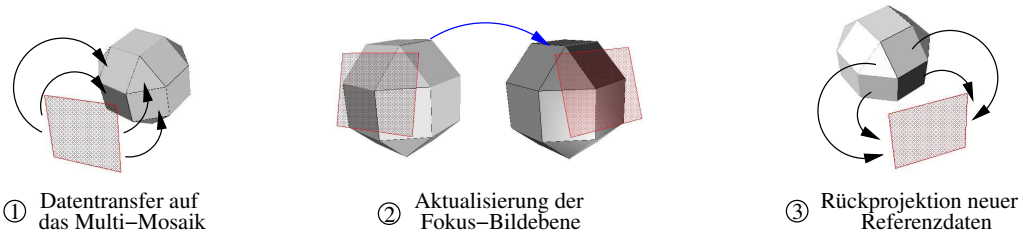



**Abbildung 6.16:** Dreistufiges Verfahren zur Aktualisierung der Fokus-Bildebene im Verlauf der Online-Mosaikberechnung: 1) Transfer der aktuellen Daten ins Multi-Mosaikbild, 2) Neuausrichtung der Ebene und 3) Rückprojektion ikonischer Referenzdaten vom Multi-Mosaikbild auf die neue Ebene.

Eine Korrektur des Abstandes der Fokus-Bildebene zum optischen Kamerazentrum wird durchgeführt, wenn die Bildweite des aktuellen Bildes um mehr als $\Delta f/2$ von der aktuellen Referenzbildweite abweicht (s. Abschnitt 6.2). Die neue Position der Ebene lässt sich in diesem Fall durch eine Verschiebung der aktuellen Ebene entlang ihres Normalenvektors in der durch das Vorzeichen der Bildweitenänderung vorgegebenen Richtung berechnen. Zusätzlich zur reinen Neuausrichtung der Fokus-Bildebene findet in diesem Aktualisierungsschritt auch ein Wechsel der Projektionsinstanz statt. Dazu wird in Abhängigkeit vom Vorzeichen der Bildweitenänderung der Vorgänger bzw. Nachfolger der aktuellen Projektionsinstanz in der doppelt verketteten Liste aller Instanzen gesucht. Falls dieser noch nicht existiert (beispielsweise, weil noch keine entsprechend skalierten Bilddaten zu früheren Zeitpunkten aufgenommen wurden), erfolgt an dieser Stelle dessen Instanziierung. Im letzten Schritt der Aktualisierung werden schließlich die Daten der alten Fokus-Bildebene auf die neue Ebene projiziert. Zusätzlich können dabei auch Bilddaten berücksichtigt werden, die bereits auf den Teilflächen des Multi-Mosaikbildes vorhanden waren und für die neue Fokus-Bildebene ebenfalls relevant sind.

### 6.4.2 Datenintegration

Bei der Integration neuer Bilddaten in ein Multi-Mosaikbild finden grundsätzlich die in Kapitel 4 vorgestellten Algorithmen Anwendung. Jedes neue Bild wird zunächst in die Fokus-Bildebene integriert, wobei die bereits vorhandenen Daten mit den neuen überschrieben werden. Neben einer Angleichung der aus dieser Vorgehensweise resultierenden Regionen durch eine Überblendung in Grenzbereichen erfolgt dabei auch eine Maskierung bewegter Pixel. Die Integration der Daten ins Multi-Mosaikbild selbst wird, wie im vorangegangenen Abschnitt skizziert, erst zu den aus dem Verlauf der Kamerabewegung abgeleiteten Zeitpunkten vollzogen, an denen eine Neuausrichtung der Fokus-Bildebene notwendig ist. In diesem Fall werden die Bilddaten der Fokus-Bildebene mit Hilfe projektiver Transformationen auf die jeweils korrespondierenden Teilflächen des polyedrischen Referenzkoordinatensystems kopiert und dort für den späteren Datenzugriff gespeichert. Da zumeist nur ein kleiner Teil aller vorhandenen Teilflächen eines Multi-Mosaikbildes von einem Datenaustausch betroffen ist und diese Flächen zudem unmittelbar benachbart sind, folgt das Kopieren der Daten auf die einzelnen Flächen einem rekursiven Ablaufschema, das in Abbildung 6.17 skizziert ist.

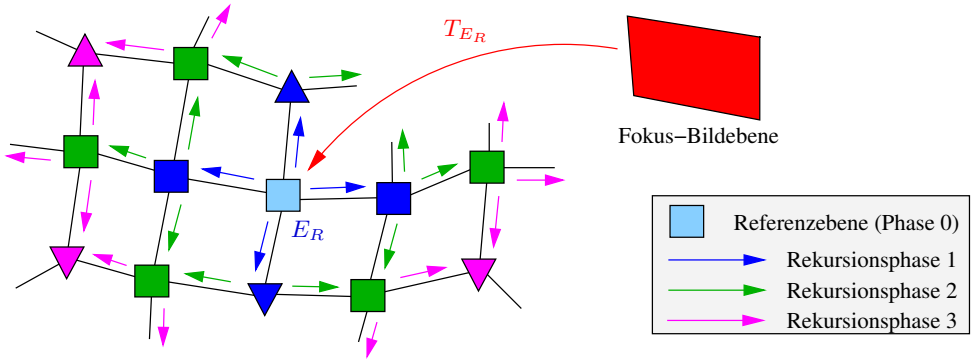



**Abbildung 6.17:** Schema des Datenaustauschs zwischen Fokus-Bildebene und Multi-Mosaikbild im Rahmen einer Aktualisierung der Position und Orientierung der Fokus-Bildebene: Die Bilddaten der Fokusebene werden zunächst mit Hilfe der Transformation $T_{E_R}$ auf die Referenzfläche $E_R$ des Multi-Mosaiks kopiert und von dort aus bei Bedarf rekursiv weiter auf benachbarte Teilflächen verteilt.

Im Verlauf der Rekursion werden sukzessive die Bereiche auf den einzelnen Teilflächen ermittelt, die zu Teilstücken des auf der Fokus-Bildebene aktuell mit Bilddaten gefüllten Bereichs (Datenbereich) korrespondieren und daher im Rahmen des Datenaustausches zu aktualisieren sind. Zur Bestimmung dieser Ausschnitte auf einzelnen Teilflächen erfolgt eine Projektion des umschließenden Rechtecks des Datenbereichs[6] auf die jeweiligen Teilflächen. Die Ausschnitte sind dann durch die Schnittflächen des projizierten Rechtecks mit den umschreibenden Polygonzügen der einzelnen Flächen gegeben, die den gültigen Projektionsbereich für Bilddaten auf den Flächen definieren.

[6]Der Datenbereich auf der Fokus-Bildebene ist zwar im Allgemeinen nicht rechteckig, eine exakte Beschreibung durch ein Polygon ist aber mit einem unverhältnismäßig hohen Aufwand verbunden.

Die Rekursion beginnt in der Phase 0 mit der Projektion des Rechtecks auf diejenige Teilfläche, die die geringste Abweichung in der Orientierung gegenüber der Fokus-Bildebene aufweist (Referenzfläche $E_R$). Sie wird in der Multi-Mosaikdatenstruktur jeweils durch einen Zeiger referenziert. Die Grundlage der Projektion bildet dabei eine zuvor mit Hilfe der 4-Punkt-Methode ermittelte projektive Transformation $T_{E_R}$ zwischen der aktuellen Fokus-Bildebene und der Referenzfläche $E_R$. Aus der Berechnung der Schnittfläche des Definitionsbereichs der Teilfläche mit dem projizierten Rechteck kann schließlich der Teilbereich der Fläche bestimmt werden, der im Rahmen des Datenaustausches zu aktualisieren ist.

Die Notwendigkeit zur rekursiven Fortführung des Datenaustausches auch auf benachbarte Teilflächen lässt sich durch eine erweiterte Auswertung des Ergebnisses der Schnittflächenberechnung feststellen. Liegt das projizierte Rechteck der Fokus-Bildebene vollständig innerhalb des validen Bereichs der Referenzfläche, so können alle Daten direkt auf diese Fläche projiziert werden und die Rekursion endet. Falls dies jedoch nicht der Fall ist und das projizierte Rechteck die Fläche nur partiell überdeckt, befinden sich auch auf benachbarten Flächen Teilbereiche, die von der Datenintegration betroffen sind. Die Projektion des Rechtecks wird in diesem Fall rekursiv für alle direkt benachbarten Flächen wiederholt (Rekursionsphasen $i, i \geq 1$), wobei die dabei anzuwendenden Homographien jeweils aus einer Konkatenation der zuvor bestimmten Abbildung $T_{E_R}$ mit den Homographien an den Kanten des Nachbarschaftsgraphen resultieren. Der Abbruch der Rekursion erfolgt, wenn eine Überprüfung bzw. Aktualisierung aller Nachbarn einer Fläche bereits stattgefunden hat.

Zur Reduktion des Aufwandes werden im Verlauf der Aktualisierung grundsätzlich nur die Flächen untersucht, deren Grundorientierungen um nicht mehr als 80° von der der Fokus-Bildebene abweichen. Der im Vergleich zu Abbildung 6.2 recht groß gewählte Wert für die maximal zulässige Winkeldifferenz motiviert sich dabei durch die Zielsetzung, insbesondere bei großen Fokus-Bildebenen möglichst viele Daten auf das Multi-Mosaik zu übertragen. Auch bei großen Winkeln zwischen den Ebenen lassen sich oftmals zumindest Teilmengen der Daten noch kopieren und damit in die eigentliche Speicherstruktur übernehmen. Allerdings können dabei auch singuläre Konfigurationen auftreten (Abb. 6.18), die geeignet zu behandeln sind. In derartigen Situationen werden einzelne Eckpunkte des Datenbereichs durch das Kamerazentrum hindurch abgebildet (für Details, s. auch [Har00], Kap. 20), so dass sich die Projektion des Rechtecks nicht mehr durch ein konvexes Polygon beschreiben lässt und Spiegelungen der Daten bei der Projektion resultieren. In der vorliegenden Arbeit werden diese Fälle durch ein „Clipping“ des Rechtecks behandelt: Punkte, die außerhalb der für den gegebenen Anwendungskontext gültigen Grenzen der projektiven Transformation liegen, werden entfernt. Dies führt im Allgemeinen zu einer Degeneration des Rechtecks zu einem Dreieck (s. Abb. 6.18), so dass nicht alle Bilddaten großer Fokus-Bildebenen vollständig kopiert werden können. Eine alternative, vollständige Berechnung der exakten Grenzlinie, die eine Kopie aller nicht singulär abgebildeten Daten erlaubt, ist jedoch als bedeutend aufwändiger einzuschätzen und steht in keinem Verhältnis zu den geringen Nachteilen der derzeit verwendeten Heuristik. Darüber hinaus sind die Kamerabewegungen bei einer Bildaufnahme zumeist aber

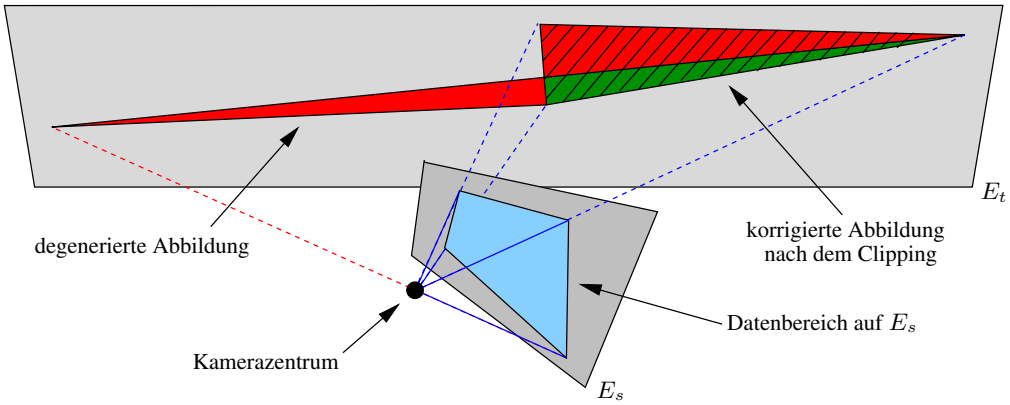



**Abbildung 6.18:** Künstliches Beispiel zur Veranschaulichung möglicher Singularitäten bei der Projektion von Daten einer Ebene $E_s$ auf eine Ebene $E_t$: Obwohl der Winkel zwischen beiden Ebenen klein genug erscheint, um eine robuste Transformation zu gewährleisten, können in Teilbereichen der Ebene $E_s$ Singularitäten bei der Projektion auftreten. Der abzubildende Datenbereich wird in diesen Fällen durch ein gezieltes Clipping modifiziert, um eine konvexe Abbildungsfläche (schraffiert) zu erzeugen.

auch hinreichend klein, so dass moderate Größen für die Fokus-Bildebene angemessen sind, bei denen derartige Singularitäten nicht auftreten.

Der Integration der Daten ins Multi-Mosaikbild selbst liegen dieselben Heuristiken zu Grunde, die auch bei der Integration aktueller Daten in die Fokus-Bildebene Anwendung finden. Dies bedeutet insbesondere, dass die bereits im Multi-Mosaikbild gespeicherten Informationen jeweils durch die aktuelleren Daten der Fokus-Bildebene überschrieben werden, wobei auch dort in den Randbereichen eine Überblendung stattfindet. Allerdings ist hierbei zu berücksichtigen, dass der tatsächlich mit Daten gefüllte Bereich auf der Fokus-Bildebene (der vorstehend zur Vereinfachung durch sein umschließendes Rechteck approximiert wurde) im Allgemeinen durch einen beliebigen, geschlossen Polygonzug spezifiziert wird (Abb. 6.19). Dessen explizite Repräsentation und stetige Aktualisierung im Verlauf der Mosaikbildberechung bedingt jedoch einen hohen Aufwand, so dass darauf im Rahmen der aktuellen Implementierung verzichtet wurde. Die Abstände einzelner Pixel zum Rand des Datenbereichs lassen sich daher nicht mehr analytisch bestimmen (vgl. Abschnitt 4.3), sondern resultieren aus einer vor dem Datenaustausch berechneten Distanztransformation der Fokus-Bildebene (Abb. 6.19).

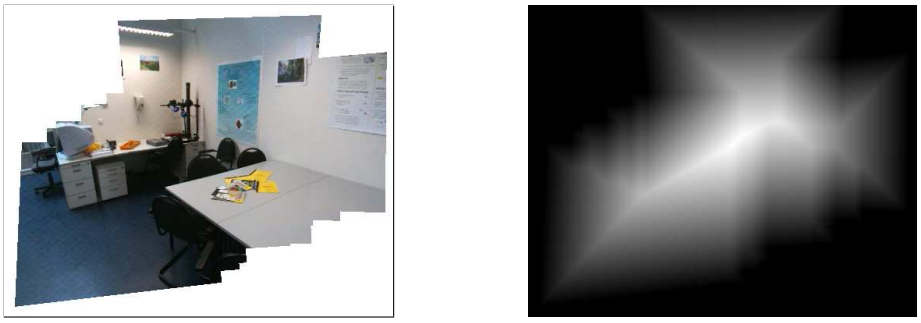


**Abbildung 6.19:** Ergebnis der Distanztransformation (rechts) einer exemplarischen Fokus-Bildebene (links) zur Bestimmung des Abstandes der einzelnen Pixel zum Rand des gültigen Datenbereichs. Hohe Intensitätswerte kennzeichnen große Abstände.

Die im Rahmen des letzten Schrittes des Aktualisierungsschemas (Abb. 6.16) stattfindende Rückprojektion zusätzlicher, relevanter Bilddaten des Multi-Mosaikbildes auf die neue Fokus-Bildebene folgt demselben rekursiven Schema, das zur Datenübertragung von der Fokus-Bildebene in das Referenzkoordinatensystem des Multi-Mosaikbildes genutzt wird. Der einzige Unterschied besteht darin, dass die Richtung der Datenprojektion umgekehrt verläuft und der Zielbereich der Daten auf der neuen Fokus-Bildebene sukzessive mit relevanten Bilddaten von den einzelnen Teilflächen gefüllt wird.

### 6.4.3 Besonderheiten der Bildregistrierung

Durch die Verwendung der Fokus-Bildebene zur vereinfachten Handhabung des polyedrischen Koordinatensystems eines Multi-Mosaikbildes wird eine effiziente Registrierung und Integration neuer Daten in den visuellen Speicher möglich. Die Fokus-Bildebene bietet einerseits einen direkten Zugriff auf die zuletzt integrierten Daten und erlaubt andererseits die Verwendung gängiger Algorithmen zur Bildregistrierung, die Bilder mit euklidischen Koordinaten voraussetzen. Damit kann insbesondere der in Abschnitt 3.3.1 vorgestellte Ansatz des Frame-to-Mosaic ohne Schwierigkeiten eingebunden werden, der bei einer Integration neuer Daten im Online-Modus eine Reduktion von Akkumulationsfehlern verspricht. Da sich die Fokus-Bildebene grundsätzlich nicht von einem konventionellen Mosaikbild mit nur einer einzelnen Projektionsebene unterscheidet, lässt sich der dazu jeweils benötigte Referenzausschnitt des aktuellen Mosaikbildes mit den zuvor beschriebenen Mechanismen extrahieren.

Zur Lösung des mit der Anwendung des Projective Flow zur Parameterschätzung verbundenen Problems zu großer Kamerabewegungen zwischen aufeinander folgenden Bildern einer Sequenz wurde in Abschnitt 3.2.2 die Möglichkeit vorgestellt, Wissen über die Kamerabewegung bei der Initialisierung des Schätzprozesses im Frame-To-Mosaic-Modus auf der untersten Ebene der Auflösungspyramide zu berücksichtigen. Auch dieser Ansatz lässt sich direkt in die Parameterschätzung für ein Multi-Mosaikbild bzw. für die zugehörige Fokus-Bildebene einbinden. Während bei Verwendung einer einzelnen Ebene durch deren Geometrie obere Schranken für die maximal zulässigen Rotationswinkel der Kamera festgelegt sind (vgl. Abschnitt 6.1.1), bietet ein Multi-Mosaikbild durch die uneingeschränkte Repräsentation des Sichtbereichs einer stationären, rotierenden und auch zoomenden Kamera hier zusätzlich eine größere Flexibilität.

Die Grundidee des Frame-To-Mosaic-Modus besteht darin, die Referenzdaten zur Registrierung eines neuen Bildes aus dem bislang erzeugten Mosaikbild zu extrahieren und damit alle darin repräsentierten Bilder früherer Zeitpunkte implizit in die Parameterschätzung einzubeziehen. Wenn dabei keine Annahmen über die zwischen den beiden zuletzt aufgenommenen Bildern $I_{t-1}$ und $I_t$ durchgeführten Kamerabewegungen vorliegen, so dient die inverse, projektive Transformation $T^{-1}_{\vec{p}_{t-1}}$ zur Integration des zuletzt verarbeiteten Bildes $I_{t-1}$ als Grundlage zur Extraktion eines geeigneten Ausschnittes des Mosaikbildes. Die Registrierung gelingt in diesem Fall nur dann, wenn die Abbildung zwischen beiden Bildern nahe der Identität liegt. Sind dagegen die Bewegungsparame-

ter der Kamera bekannt, so kann durch eine Konkatenation des letzten Parametersatzes $\vec{p}_{t-1}$ mit den durch die Kamerabewegung induzierten Parametern $\Delta\vec{p}_r$ direkt ein besserer Referenzausschnitt generiert werden. Auf einem Multi-Mosaikbild sind dabei für die Kamerabewegungen keinerlei Einschränkungen gegeben. Unter der Voraussetzung, dass in dem durch die aktuelle Kamerabewegung angesteuerten Szenenbereich bereits zu einem früheren Zeitpunkt Bilddaten aufgenommen wurden, lassen sich selbst dann geeignete Referenzdaten gewinnen, wenn zwischen den letzten beiden Bildern kein Überlapp gegeben ist und die Bilder vollständig disjunkte Szenenausschnitte abbilden.

In einem solchen Fall wird die Fokus-Bildebene zunächst an die geschätzte Position des neuen Bildes angepasst, bevor die eigentliche Parameterschätzung erfolgt (Abb. 6.20). Neben großen Rotationswinkeln können auf diese Weise auch signifikante Änderungen in der Bildweite innerhalb einer Sequenz behandelt werden. Während eine Registrierung zweier Bilder, die zwar denselben Szenenausschnitt zeigen, jedoch stark divergierende Bildweiten aufweisen, im Allgemeinen große Probleme aufwirft (vgl. z.B. [Zog97] oder auch Abschnitt 3.5.2), lassen sich im Rahmen des vorliegenden Kontextes auch solche Daten bei vorhandenen Referenzdaten in der entsprechenden Auflösung registrieren.

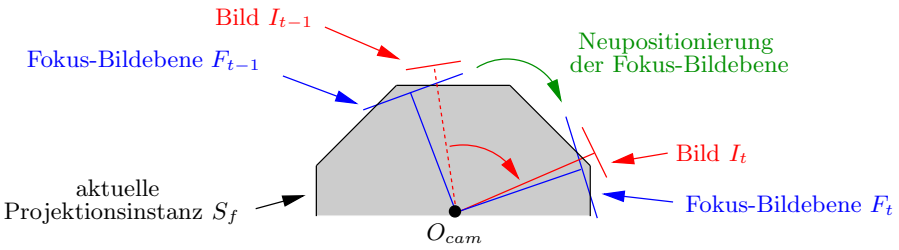



**Abbildung 6.20:** Erweiterte Möglichkeiten einer Parameterinitialisierung auf Basis von Multi-Mosaikbildern: Die Fokus-Bildebene wird vor der Schätzung gemäß der angenommenen Kamerabewegung neu positioniert, so dass prinzipiell beliebige Kamerabewegungen behandelt werden können.

Die vorstehend skizzierten, erweiterten Möglichkeiten einer Registrierung von Bildern auf Basis von Multi-Mosaikbildern unterstützen im Grundsatz uneingeschränkte Kamerabewegungen. Allerdings muss in der Praxis der Einfluss von Registrierungsfehlern im Verlauf einer Online-Mosaikbildberechnung berücksichtigt werden. Während sich die aktuellen Bewegungsparameter der Kamera aus deren Hardwaredaten errechnen lassen, resultieren die Integrationsbereiche aufgenommener Bilder im Multi-Mosaikbild allein aus den geschätzten projektiven Transformationen und im Vorfeld ermittelten Bildweiten. Fehler in diesen Daten können zu deutlichen Differenzen zwischen den realen Positionen von Bildern auf dem Grundkörper und der aufgrund der Transformationsparameter angenommenen Lage führen. Bei einer Extraktion von Referenzdaten aus dem Multi-Mosaikbild, deren Ursprung wiederum aus den Hardwaredaten der Kamera hergeleitet wird, ist damit nicht zwangsläufig gewährleistet, dass die extrahierten Bilddaten auch tatsächlich den implizierten Szenenbereich abbilden. Die praktischen Auswirkungen dieser Einflussfaktoren werden in Abschnitt 6.5.1 genauer analysiert.

## 6.5 Ergebnisse & Diskussion

Das in den vorangegangenen Abschnitten skizzierte Konzept der Multi-Mosaikbilder wurde als Erweiterung eines integrierten Systems (und Softwarepaketes) zur Berechnung von Mosaikbildern implementiert und evaluiert [Wil03]. Das Gesamtsystem umfasst dabei neben den zur Generierung eines Mosaiks notwendigen Grundfunktionen der Bildregistrierung und Integration auch die in Kapitel 5 vorgestellten Verfahren zur Detektion von Bewegungen und einer erweiterten Analyse der daraus resultierenden Daten. Damit konnten diese Algorithmen direkt in die Generierung der Multi-Mosaikbilder eingebunden werden (vgl. Abb. C.1). In diesem Abschnitt werden Ergebnisse aus der praktischen Anwendung der Multi-Mosaikbilder vorgestellt und diskutiert. Der nachfolgende Unterabschnitt zeigt dabei zunächst die grundlegende Leistungsfähigkeit des Ansatzes anhand ausgewählter Beispiele auf. Da bei einer Online-Berechnung von Multi-Mosaikbildern aufgrund der gegebenen Datengrundlage eine fehlerfreie Registrierung der Bilder nicht in allen Fällen gewährleistet werden kann (vgl. Unterkap. 3.3), geht Abschnitt 6.5.2 dabei gezielt auf die Auswirkungen lokaler Registrierungsfehler und mögliche Korrekturansätze ein. In Abschnitt 6.5.3 schließlich wird die derzeitige Performanz des Systems analysiert, die hinsichtlich eines Einsatzes der Bilder in interaktiven Systemen bedeutsam ist.

### 6.5.1 Multi-Mosaikbilder in der Praxis

Die nachfolgend gezeigten Multi-Mosaikbilder wurden gemäß den Ausführungen in den vorherigen Abschnitten unter Verwendung einer aktiv gesteuerten Kamera im Online-Modus berechnet. Innerhalb des Systems lag den Algorithmen dabei ausschließlich die eingeführte Graphdatenstruktur zur *impliziten* Repräsentation der 3D-Topologie der Multi-Mosaikbilder zu Grunde (Abschnitt 6.1.3). Zur besseren Veranschaulichung der geometrischen Struktur der Multi-Mosaikbilder erfolgte daher im Rahmen der nachfolgenden Auswertungen eine *explizite* Generierung von 3D-Visualisierungen mit Hilfe der unter der „GNU Lesser General Public License“ frei verfügbaren 3D-Bibliothek „Open Inventor“[7]. Die Visualisierungen der Beispiele sowie die verarbeiteten Bildfolgen finden sich auch im Internet unter `http://www.informatik.uni-halle.de/~moeller/phd/`.

In Abbildung 6.21 ist zunächst der grundlegende Aufbau eines Multi-Mosaikbildes anhand eines Beispiels veranschaulicht. Dargestellt sind dort alle Teilflächen des zu Grunde liegenden Rhombenkuboktaeders, auf die im Verlauf der Berechnungen Bilddaten projiziert wurden. Schwarze Zonen auf den einzelnen Flächen markieren dabei Teilbereiche, für die ungenutzter Speicher alloziert wurde. Die grauen Bereiche entsprechen undefinierten Regionen der Teilflächen und dienen vorrangig einer besseren Darstellung der grundlegenden, dreidimensionalen Geometrie des Körpers. Die im Rahmen dieses ersten Beispiels verarbeitete Bildfolge (Abb. A.4) umfasst einen horizontalen Rotationswinkel der Kamera von rund 70° und einen vertikalen Winkel von etwa 15°. Wie aus der Abbildung deutlich hervorgeht, können die Bilddaten auf den Teilflächen des Polyeders trotz der großen Rotationswinkel adäquat repräsentiert werden.

[7] SGI - Open Inventor™, `http://oss.sgi.com/projects/inventor/`

Diese Eigenschaft der Multi-Mosaikbilder wird auch durch das zweite Beispiel in Abbildung 6.22 untermauert. Dort ist der Ausschnitt eines Mosaikbildes gezeigt, das durch einen horizontalen Kamera-Scan erzeugt wurde. Das Mosaikbild weist insgesamt eine gute Qualität auf, wobei insbesondere keine geometrische Verzerrungen zu beobachten sind. Es bildet damit eine gute Ausgangsbasis, um interaktiven Systemen eine effiziente Speicherung der Bildfolge durch ein signifikant vermindertes Datenvolumen (die gesamte Sequenz umfasst in diesem Fall rund 162 MB, während ein vollständiges Mosaikbild nur etwa 18 MB aufweist[8]) und eine einfache Weiterverarbeitung der akquirierten Daten zu ermöglichen.

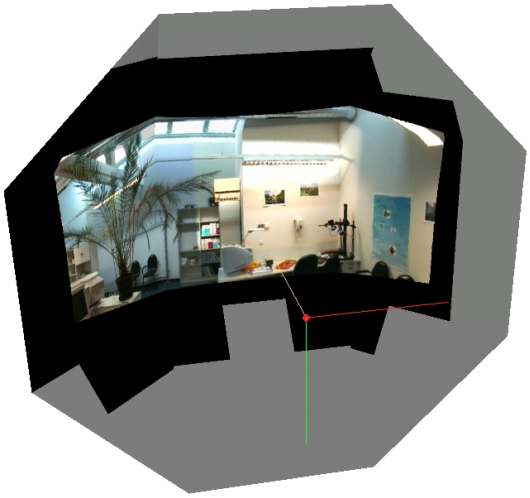


**Abbildung 6.21:** Ein exemplarisches Multi-Mosaikbild, berechnet aus 32 Einzelbildern (Abb. A.4).

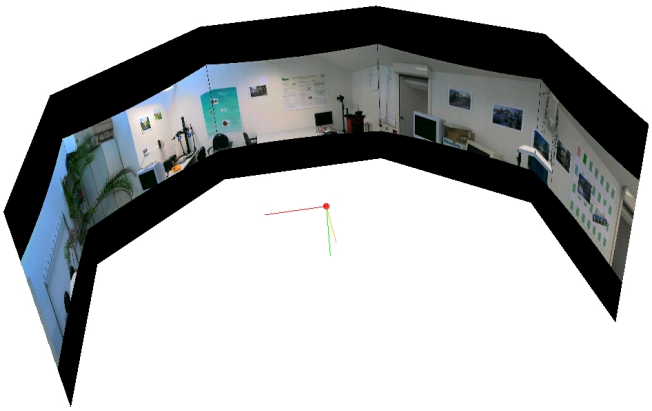


**Abbildung 6.22:** Ein Multi-Mosaikbild, das einen horizontalen Kamerascan umfasst. Die dabei zu Grunde gelegten Bilder entstammen der Sequenz, die in Abbildung A.5 zu sehen ist.

Beiden vorstehend diskutierten Multi-Mosaikbildern liegen Bildfolgen zu Grunde, die nur eine einzelne Auflösungsebene umfassen. Das Konzept der Multi-Mosaikbilder erlaubt jedoch auch eine adäquate Handhabung verschiedener Auflösungsstufen innerhalb einer Bildfolge. Die finale Repräsentation einer solchen, mehrere Auflösungsebenen umfassenden Bildsequenz ist in Abbildung 6.23 zu sehen. Innerhalb der aufgenommenen Szene (s. Bilder in Abb. A.6) wurden das Regal auf der rechten Seite der Szene, sowie die Tafel in der Mitte mit großen Bildweiten aufgenommen. Die resultierenden Ausschnitte in Abbildung 6.24 veranschaulichen, dass spezifische Informationen der Szene, etwa die Schrift auf der Tafel (Abb. 6.24(b)) oder Details der Titelseite des Buches im Regal

[8] Der Umfang der Datenreduktion skaliert im Allgemeinen mit der Redundanz in den Eingabedaten. Je seltener einzelne Szenenausschnitte wiederholt aufgenommen werden, desto kleiner ist auch die mit den Multi-Mosaikbildern erzielbare Kompressionsrate der Daten.

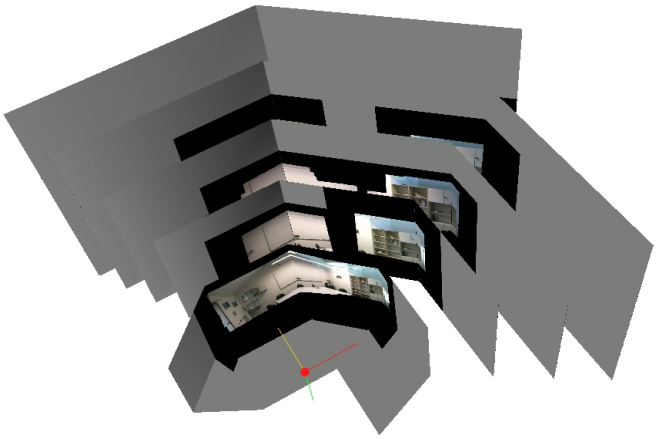


**Abbildung 6.23:** Das Multi-Mosaikbild einer Bildfolge, die mehrere Auflösungsstufen umfasst. Die lokal variierenden Bildweiten wurden jeweils in Abhängigkeit vom lokalen Detailgrad in der Szene ausgewählt, so dass eine flexible, gut an die Daten angepasste, ikonische Szenenrepräsentation resultiert.

(Abb. 6.24(a)) nur in hohen Auflösungen zugänglich sind und eine gezielte Steuerung des Kamerazooms damit für viele Anwendungsfelder unerlässlich ist.

Ein Vergleich der auf den einzelnen Ebenen des Multi-Mosaikbildes in Abbildung 6.23 repräsentierten Daten veranschaulicht zusätzlich die partielle Datenrepräsentation innerhalb des Mosaikbildes. Es ist deutlich zu sehen, dass verschiedene Ausschnitte der Szene nicht in allen Auflösungsebenen gleichermaßen gespeichert sind, sondern sich die Darstellung vielmehr an den lokalen Gegebenheiten in der Szene orientiert. Der visuelle Speicher bietet somit eine hohe Flexibilität bei der Darstellung lokal variierender Detailgrade in den ikonischen Daten einer Szene.

(a) (Skalierte) Ausschnitte der niedrigsten und höchsten Stufen im direkten Vergleich.

(b) (Kontrastverstärkte) Detailaufnahme der Tafel.

**Abbildung 6.24:** Exemplarische Ausschnitte des Multi-Mosaikbildes aus Abbildung 6.23.

Im Mittelpunkt der Entwicklungen in dieser Arbeit steht der Entwurf eines adäquaten Konzeptes zur Repräsentation ikonischer Daten, die den vollständigen Sichtbereich einer stationären, rotierenden Kamera umfassen. Neben der Bereitstellung euklidischer Koordinaten für eine direkte Weiterverarbeitung der gespeicherten Daten, konnte dabei auch eine Online-Berechnung der Multi-Mosaikbilder realisiert werden. Im Hinblick auf eine robuste Online-Parameterschätzung profitiert dabei insbesondere der Frame-To-Mosaic-Modus in Kombination mit einer expliziten Initialisierung der Schätzung aus Hardwaredaten von dem erweiterten Sichtfeld der Multi-Mosaikbilder. Während bei der

Berechnung von Mosaikbildern auf Basis einer einzelnen Ebene nur die auf der Ebene gespeicherten Daten als Referenz in der Parameterschätzung verwendet werden können, repräsentieren Multi-Mosaikbilder den vollständigen Sichtbereich der Kamera. Sie unterstützen damit durch eine gezielte Aktualisierung der für die Registrierung neuer Bilder explizit eingeführten Fokus-Bildebene (Abschnitt 6.4.1) vor der Schätzung im Grundsatz beliebige Kamerabewegungen.

In Abbildung 6.25 sind zwei Beispiele zu sehen, in denen eine explizite Initialisierung der Fokus-Bildebene eine Behandlung großer Rotationen (linke Spalte) und signifikanter Zoomveränderungen (rechte Spalte) ermöglicht hat. Im linken Beispiel erfolgte eine große Kamerarotation, durch die der Überlapp des aktuellen Bildes (linke Spalte Mitte) mit den Daten der Fokus-Bildebene (oben) auf wenige Prozentpunkte reduziert wurde. Erst durch eine explizite Aktualisierung der Bildebene (unten) wurde eine Registrierung möglich. Das rechte Beispiel zeigt denselben Effekt bei einer starken Bildweitenänderung. Ohne eine explizite Adaption der aktuellen Fokus-Bildebene (oben) ist für eine Registrierung des neuen Bildes nur eine signifikant verminderte Datengrundlage gegeben, die zudem gravierende Skalierungen bedingt (vgl. auch Abschnitt 3.5.2). Auch hier erlaubt erst eine Adaption der Fokusebene (rechts unten) eine Schätzung von Transformationsparametern.

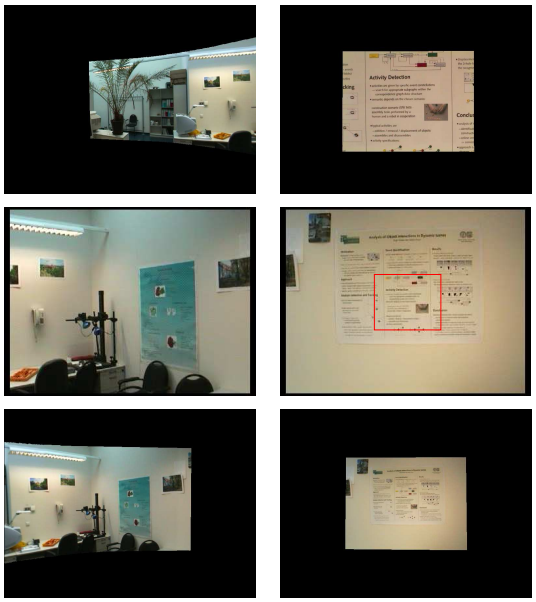


**Abbildung 6.25:** Zwei Beispiele für eine explizite Adaption der Fokus-Bildebene bei großen Kamerabewegungen (links Rotation, rechts Bildweite): Die obere Zeile zeigt die Fokusebenen vor der Aktualisierung, mittig sind die zu registrierenden Bilder zu sehen, und die untere Zeile enthält die aktualisierten Ebenen.

In der praktischen Anwendung stehen der Flexibilität des Ansatzes, der theoretisch beliebige Kamerasprünge zulässt, allerdings oftmals Inkonsistenzen in der Repräsentation gegenüber, die eine Beschränkung der Bewegungen bedingen. Die angenommene Position eines neuen Bildes relativ zum Referenzkoordinatensystem, die als Ausgangspunkt zur Adaption der Fokus-Bildebene dient, resultiert aus den Hardwareparametern der Kamera. Die Positionen der Bilddaten auf dem polyedrischen Grundkörper werden jedoch allein aus den geschätzten Homographien abgeleitet. Damit wirken sich Registrierungsfehler direkt auf die explizite Initialisierung der Fokus-Bildebene aus. Das Risiko für das Auftreten derartiger Fehler ist dabei umso höher, je größer die Kamerasprünge sind und je höher die Anzahl zwischenzeitlich registrierter Bilder ist. Darüber hinaus übt auch das Zeitintervall zwischen der Aufnahme des akuellen Bildes und der zugehörigen Referenzdaten sowie die Bildweite der Eingangsbilder (vgl. nachfolgenden Abschnitt) einen Einfluss aus. Obgleich es zwar oftmals mit Hilfe des vorgestellten Ansatzes gelungen ist, auch Rotationswinkel von mehr als 90° zu behandeln und bei einer Änderung des Zooms gleich mehrere Auflösungsstufen zu überspringen, so bergen sehr große Ka-

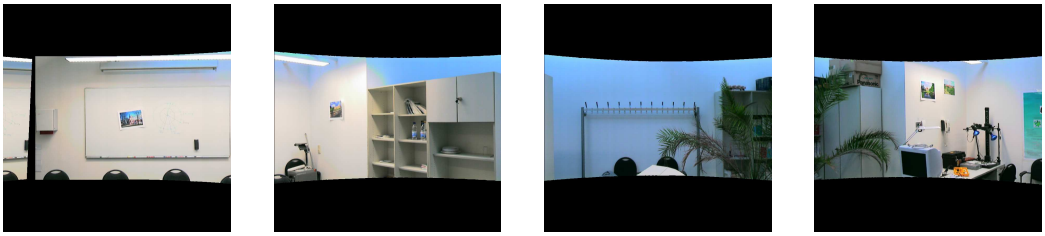


**Abbildung 6.26:** Die ersten vier Teilflächen eines Multi-Mosaikbildes, das einen 360°-Scan umfasst.

merabewegungen trotzdem ein hohes Risiko für eine fehlschlagende Registrierung. Sie sollten somit vor dem Hintergrund einer robusten Langzeitregistrierung nur in Ausnahmefällen durchgeführt werden. Die direkte Abhängigkeit der Robustheit einer expliziten Initialisierung der Fokus-Bildebene von der aktuellen Qualität der Parameter und damit der Repräsentation insgesamt erschwert dabei auch die Spezifikation allgemeingültiger, sinnvoller oberer Schranken für zulässige Bewegungen.

### 6.5.2 Datenkonsistenz und Fehlerkorrektur

Eine Registrierung von Bilddaten im Online-Modus erlaubt keine Bestimmung global optimaler Parameter für eine vollständige Bildsequenz, da zu einem Zeitpunkt jeweils nur Teilmengen dieser Daten für die Parameterschätzung gegeben sind. Daraus folgt, dass die Qualität einer Multi-Mosaikrepräsentation direkt mit dem Auftreten lokaler Registrierungsfehler verknüpft ist und damit zwischenzeitlich in Teilbereichen auch inkonsistente Daten enthalten kann. Diese Effekte treten insbesondere bei Kamerapfaden deutlich hervor, die verschiedene Ausschnitte einer Szene wiederholt zu verschiedenen Zeitpunkten anfahren. Ein Beispiel für solche Bewegungen sind etwa horizontale 360°-Scans, wie sie zur Analyse der globalen Datenkonsistenz hier durchgeführt wurden.

In den Abbildungen 6.26 und 6.27 ist das exemplarische Ergebnis eines solchen Scans zu sehen. Die Bildqualität ist im Allgemeinen gut und es sind insgesamt keine signifikanten geometrischen Verzerrungen zu beobachten, die auf Registrierungsfehler schließen lassen. Dennoch zeigen sich gegen Ende der Drehung (vgl. erstes Bild in Abb. 6.26 links mit dem Bild rechts in Abb. 6.27) deutlich die Auswirkungen einer Akkumulation von geringen Registrierungsfehlern. Innerhalb der zu Grunde liegenden Bildfolge (Abb. A.5) überlappen sich das erste und das letzte Bild, in dem finalen Multi-Mosaikbild werden diese jedoch auf verschiedene Teilflächen projiziert, so dass eine markante „Lücke“ im Mosaikbild von etwa 20° verbleibt. In mehreren Experimenten hat sich dabei ein deutlicher Zusammenhang zwischen der Bildweite der Kamera bei der Aufnahme der Sequenzen und der Größe der Lücke gezeigt. Mit einer zunehmenden Bildweite verringert sich die Lücke fortwährend. Diese Beobachtung legt die Vermutung nahe, dass Linsenverzerrungen, deren Einfluss im Allgemeinen mit einer zunehmenden Bildweite abnimmt, die Schätzung korrekter Parameter bei kleinen Bildweiten maßgeblich stören und somit als Hauptursache für die verbleibenden Unstimmigkeiten anzunehmen sind (vgl. auch Abschnitt 2.4.1). Dabei ist allerdings festzuhalten, dass sich diese Effekte vorrangig auf die metrischen Relationen innerhalb des Mosaikbildes auswirken. Während Bilddaten

**Abbildung 6.27:** Die zweiten vier Teilflächen eines Multi-Mosaikbildes, das einen 360°-Scan umfasst.

damit zwar unter Umständen auf eine falsche Position innerhalb des Multi-Mosaikbildes projiziert werden, ist die Bildqualität auch in diesen Fällen gut, so dass sich daraus insgesamt kaum Einschränkungen für eine nachfolgende Analyse ergeben.

Perspektivisch sind dennoch weitere Verbesserungen durch eine Integration von expliziten Korrekturmechanismen denkbar. Dabei ist zunächst anzumerken, dass geringe Abweichungen in den metrischen Relationen und Registrierungsfehler bei einem moderaten Versatz in den Aufnahmezeitpunkten der betroffenen Bildregionen oftmals schon durch den Frame-To-Mosaic-Modus reduziert werden. Durch die Integration der neuen Daten erfolgt im Prinzip eine „automatische“ Korrektur des Mosaiks. Die Berechnung eines vollständigen 360°-Scans kann aufgrund des insgesamt geringen Überlapps zwischen nicht direkt aufeinander folgenden Bildern der Sequenz jedoch kaum von einer solchen Vorgehensweise profitieren (vgl. Abb. 6.26/6.27). Die auftretenden Unstimmigkeiten sind dort somit weitaus größer als bei gängigeren Kamerabewegungen, die sich öfter kreuzen.

Auch eine Online-Korrektur derartiger Fehler ist aufgrund der geometrischen Zusammenhänge bei der Projektion von Bilddaten, die mit einer Bildweite $f$ aufgenommen wurden, jedoch auf ein Polyeder mit einer Referenzskalierung von $f + \epsilon_f$ projiziert werden, komplex. Insbesondere bewirkt eine explizite Reskalierung des Grundkörpers lediglich eine lokale Stauchung bzw. Dehnung der Bilddaten auf dem Körper, ohne jedoch die Lücke selbst signifikant zu verändern (vgl. hierzu auch die Zusatzbeispiele im Internet).

In der Literatur finden sich verschiedene Ansätze zur Behandlung solcher Fehler in Offline-Verfahren. Die Grundidee besteht zumeist darin, ausgehend von einem Vorliegen der kompletten Bildfolge zunächst Parameter zu schätzen und diese nachfolgend direkt gemäß der initial verbleibenden Lücke anzupassen [Sze97, Gon98]. Eine Übertragung derartiger Ansätze auf den vorliegenden Kontext ist jedoch schwierig, da insbesondere die Originalbilder einer Sequenz bei einer Online-Berechnung nicht mehr für eine erneute Anwendung nachträglich modifizierter Parameter zur Verfügung stehen. Eine Korrektur könnte in diesem Fall somit eher anhand einer zusätzlichen Transformation der Bilddaten des Mosaikbildes selbst realisiert werden, die jedoch nicht mehr in unmittelbarem Zusammenhang zu den zuvor bestimmten Homographien steht. Zudem wirft auch die vor einer solchen Korrektur notwendige, automatische Detektion von Registrierungsfehlern und Abweichungen in den metrischen Relationen eine Reihe neuer Probleme auf, die an dieser Stelle über den Rahmen der Arbeit hinausgehen.

Die hier und in den vorangegangenen Abschnitten vorgestellten Ergebnisse zeigen insgesamt, dass es mit dem Konzept der Multi-Mosaikbilder gelungen ist, einen visuellen Speicher zu entwickeln, der eine effiziente Verarbeitung und Analyse von Bildfolgen aktiver Kameras in interaktiven Systemen optimal unterstützen kann. Den Abschluss der

Evaluation bildet nun eine Auswertung der derzeitigen Performanz des Gesamtsystems, die im Hinblick auf einen Einsatz in interaktiven Systemen bedeutsam ist.

### 6.5.3 Performanz des Gesamtsystems

Wie eingangs skizziert, ist die Implementierung der Multi-Mosaikbilder Teil eines umfangreichen Systems zur Berechnung von Mosaikbildern. Die darin eingebettete Hauptroutine zur Berechnung der Multi-Mosaikbilder (s. auch Abb. C.1) orientiert sich grundsätzlich an dem allgemeinen Schema zur Mosaikbildberechnung, das in Abbildung 1.3 der Einleitung vorgestellt wurde. Die Performanz der entwickelten Algorithmen setzt sich damit aus den Laufzeiten der Basismodule zur Bildregistrierung, Bewegungsdetektion/ -analyse und zur Integration zusammen, sowie aus denen der zur Berechnung der Multi-Mosaikbilder notwendigen, zusätzlichen Komponenten (Funktionen zur Ein-/Ausgabe wurden an dieser Stelle nicht berücksichtigt). Zur Evaluation erfolgte eine Berechnung verschiedener (Grauwert-) Multi-Mosaikbilder (mit insgesamt 496 Einzelbildern der Größe $320 \times 240$) auf dem Testsystem (vgl. S. 56), wobei die mittleren Laufzeiten aller Teilmodule pro Bild ermittelt wurden. Die Ergebnisse sind in Tabelle 6.1 zusammengefasst.

Das System benötigt demnach bei Bildern der Größe $320 \times 240$ derzeit knapp 1,3 $s$ Verarbeitungszeit pro Bild. Der Großteil der Laufzeit wird dabei für die Parameterschätzung beansprucht. Die verhältnismäßig große Standardabweichung in ihren Zeiten ist auf einzelne Schätzungen mit einer deutlich erhöhten Iterationsanzahl zurückzuführen. Beim Multi-Mosaik-Modul resultiert die große Abweichung dagegen aus der Tatsache, dass Aktualisierungen der Fokus-Bildebene nicht in jedem Schritt durchgeführt werden müssen. Eine reine Überprüfung der aktuellen Lage schlägt sich in den Laufzeiten kaum nieder, während eine Aktualisierung aufgrund der dabei notwendigen Datenkopien im Mittel etwa 1,25 Sekunden erfordert. Die Gesamtlaufzeit hängt somit maßgeblich von der Anzahl notwendiger Aktualisierungen der Fokus-Bildebene ab.

| Modul | $t_{run}[ms]$ | $\sigma[ms]$ |
|---|---|---|
| Schätzung | 786,1 | 293,2 |
| Bewegung[9] | 81,8 | 5,2 |
| Integration | 224,9 | 38,1 |
| Multi-Mosaik | 141,5 | 367,0 |
| Gesamtzeit | 1234,3 | — |

**Tabelle 6.1:** Mittlere Laufzeiten $t_{run}$ der einzelnen Systemkomponenten pro Grauwertbild (bei insgesamt 496 Bildern) im Überblick.

Insgesamt zeigt sich, dass die derzeitige Implementierung eine in vielen Anwendungskontexten hinreichende Verarbeitungsrate gewährleisten kann. Insbesondere vor dem Hintergrund, dass oftmals auch niedriger aufgelöste Bilder bereits für die Lösung spezifischer Probleme ausreichend sind, sollte sich das System gut in die Architektur interaktiver Systeme einpassen lassen. Nichtsdestotrotz ist perspektivisch eine weitere Verbesserung der Performanz empfehlenswert. Insbesondere effizientere Ansätze für die Parameterschätzung und eine weitere Optimierung der realisierten Datenstrukturen sowie der Speicherverwaltung können dazu gute Einstiegspunkte bilden.

[9] Der tatsächliche Aufwand der Bewegungsanalyse skaliert mit der Anzahl und Größe bewegter Objekte in einer Bildfolge und kann daher im Einzelfall von den hier gemessenen Werten abweichen.

# 7 Aktive Datenakquisition und Szenenexploration

Spezifische Verhaltensmuster eines Menschen in Alltagssituationen sind das Ergebnis eines komplexen Zusammenspiels verschiedenster Komponenten seines kognitiven Systems unter Auswertung unterschiedlichster Einflussfaktoren. Neben dem Wissens- und Erfahrungsschatz des Menschen kommt dabei insbesondere Informationen über die aktuelle Situation und den Zustand der unmittelbaren Umgebung eine hohe Bedeutung zu. Sie werden durch eine Vielzahl multimodaler Sensoren (Augen, Ohren, etc.) gewonnen. Der Umfang der daraus resultierenden Datenmenge kann dabei aufgrund der großen Flexibilität in der Datenakquisition und der Komplexität der Alltagswelt, in der sich Menschen im Allgemeinen bewegen, beträchtlich sein. Zwar weisen die menschlichen Fähigkeiten zur Verarbeitung und Speicherung aufgenommener Informationen, die eine unabdingbare Existenzgrundlage für den Menschen bilden, eine bedeutende Leistungsfähigkeit auf, ihre Kapazitäten sind jedoch begrenzt (s. z.B. [Cow97], S. 8ff. bzw. S. 77ff.). Nichtsdestotrotz können Menschen auch unter diesen Voraussetzungen eine effiziente Informationsverarbeitung durchführen und in verschiedenen Situationen adäquat (re)agieren. Der Schlüssel zu diesen Fähigkeiten liegt darin, nicht *alle* verfügbaren Informationen auch tatsächlich zu verarbeiten, sondern durch eine gezielte Selektion das zu analysierende Datenvolumen signifikant zu reduzieren. Eine solche Selektion umfasst unter anderem eine Fokussierung der Informationsaufnahme und -verarbeitung auf ausgewählte Informationsquellen, sowie eine Klassifikation der Datenströme im Hinblick auf ihre aktuelle Relevanz. Die Mechanismen des menschlichen Kommunikations- und Interaktionssystems (Abb. 1.1), mit deren Hilfe diese Selektion erfolgt, werden in der Psychologie im Allgemeinen unter dem Begriff der *Aufmerksamkeit* zusammengefasst.

Eine Verarbeitung von multimodalen Informationen in technischen Systemen bringt vergleichbare Schwierigkeiten mit sich, vor die auch der Mensch im alltäglichen Leben gestellt wird. Zur Lösung einer konkreten Aufgabe muss ein System mit einer geeigneten Sensorik ausgestattet sein, die eine Akquisition aller notwendigen Informationen gewährleistet. Die relevanten Informationen sind dabei auch in diesem Kontext zumeist in die sehr viel umfangreichere Menge *aller* aktuell verfügbaren Daten aus der Umgebung eingebettet. Allerdings steht die Leistungsfähigkeit technischer Systeme bei der Informationsverarbeitung deutlich hinter der menschlichen Kognition zurück. Somit lässt sich leicht ersehen, dass eine Selektion relevanter Informationen und eine Beschränkung der Informationsverarbeitung auf ein Minimum notwendiger Daten für technische Systeme umso wichtiger ist. Aus diesen Erkenntnissen hat sich im Bereich der digitalen Bild-

verarbeitung das Forschungsfeld der Active Vision entwickelt, dessen Zielsetzung in der Entwicklung von Mechanismen zur gezielten Fokussierung der Datenanalyse auf relevante Teilmengen aller jeweils zur Verfügung stehenden Informationen besteht. Eine solche Fokussierung lässt sich sowohl durch hardwareseitige Ansätze, d.h. insbesondere aktive Sensoren, wie auch durch softwarebasierte Auswahlheuristiken realisieren. Ein Einsatz aktiver Sensoren ist dabei im Wesentlichen mit der Entwicklung geeigneter Konzepte für eine sinnvolle Ansteuerung verknüpft.

In der Literatur finden sich zahlreiche Arbeiten, die sich mit dem Entwurf von Mechanismen zur Steuerung von aktiven Sensoren befassen. Im Kontext der vorliegenden Arbeit ist insbesondere eine aktive Kamerasteuerung von hohem Interesse (vgl. auch Abschnitt 7.1.2). Im Mittelpunkt der Veröffentlichungen in diesem Bereich steht zumeist die technische Umsetzung von *visueller Aufmerksamkeit* und einer *aktiven Szenenexploration*. Viele Ansätze orientieren sich dabei an den Eigenschaften menschlicher Aufmerksamkeit, wie sie beispielsweise in psychologischen Untersuchungen beobachtet werden können. Den meisten Arbeiten ist dabei gemein, dass die Nachbildung von visueller Aufmerksamkeit auf einer ausschließlichen Analyse der jeweils aktuellsten Daten beruht und Gedächtniseffekte kaum Berücksichtigung finden. Allerdings zeigen psychologische Untersuchungen, dass beim Menschen auch Wechselwirkungen zwischen der visuellen Aufmerksamkeit und im Gedächtnis gespeicherten Informationen existieren, die bei der Lösung spezifischer Aufgabenstellungen Einfluss haben (Abschnitt 7.1.1). Dies legt eine Erweiterung der Datenanalyse auf zeitlich integrierte Daten nahe, wobei insbesondere eine Auswertung vollständiger Bildsequenzen eine größere Flexibilität verspricht.

Der visuelle Speicher auf Basis von Multi-Mosaikbildern, der im Rahmen dieser Arbeit vorgestellt wird, bietet aufgrund seines in Raum und Zeit erweiterten Sichtfeldes eine gute Grundlage für einen solchen Ansatz. In diesem Kapitel wird die prototypische Implementierung einer aktiven Szenenexploration vorgestellt, die auf Multi-Mosaikbildern aufsetzt und deren Eignung für den skizzierten Anwendungskontext illustriert. Das nachfolgende Unterkapitel 7.1 gibt dabei zunächst einen kurzen Einblick in die psychologischen Hintergründe menschlicher visueller Aufmerksamkeit und einige ausgewählte Ansätze zur Nachbildung derartiger Effekte in technischen Systemen. Unterkapitel 7.2 beschreibt anschließend die Verwendung der Mosaikbilder zur aktiven Szenenexploration, bevor Unterkapitel 7.3 auf verschiedene Explorationsstrategien eingeht, die bei der Implementierung umgesetzt wurden. Den Abschluss des Kapitels bildet eine Darstellung von Ergebnissen der implementierten Algorithmen sowie deren Diskussion (Unterkap. 7.4).

## 7.1 Visuelle Aufmerksamkeit

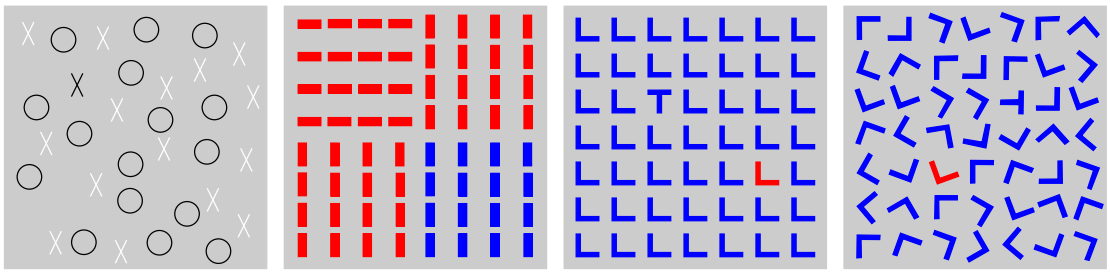
Die Beschränkung der menschlichen Informationsverarbeitung auf ausgewählte Teilmengen aller zu einem bestimmten Zeitpunkt gegebenen Sensordaten ist eine grundlegende Eigenschaft der menschlichen Kognition. Die Basis dieser Selektion bilden Prozesse der Aufmerksamkeit. In Anlehnung an die Multimodalität der zur Verfügung stehenden Sensoren lassen sich grundsätzlich verschiedene Typen von Aufmerksamkeit unterscheiden

(z.B. auditive, taktile), wobei im Kontext dieser Arbeit insbesondere *visuelle* Aufmerksamkeit von hohem Interesse ist. Nach [Wol00a] und [Itt03] bezeichnet sie allgemein alle Mechanismen, die eine Beschränkung bzw. Lenkung der visuellen Informationsverarbeitung auf ausgewählte Objekte oder Punkte der aktuellen Umgebung unterstützen. Im Hinblick auf eine Analyse der internen Informationsverarbeitung beim Menschen ermöglichen Prozesse der visuellen Aufmerksamkeit damit unter anderem, komplexe Aufgabenstellungen in kleinere Teilprobleme zu zerlegen, deren Lösungen mit geringerem Aufwand und einer stärkeren lokalen Fokussierung verbunden sind.

Eine solche Vorgehensweise kann auch in technischen Systemen bei der Verarbeitung visueller Sensordaten und einer Nachbildung menschlicher Wahrnehmungsleistungen hilfreich sein. Die große Komplexität durchzuführender Analysen und der Umfang der zu verarbeitenden Eingangsdaten bedingen oftmals einen hohen Rechenaufwand. Dieser lässt sich jedoch unter anderem durch eine gezielte Beschränkung des zu analysierenden Datenraumes beträchtlich reduzieren. Aufgrund der hohen Leistungsfähigkeit des menschlichen Aufmerksamkeitssystems liegt es dabei nahe, technische Realisierungen derartiger Ansätze an den Eigenschaften des biologischen Vorbildes zu orientieren. Der nachfolgende Abschnitt enthält daher zunächst einen groben Einblick in die psychologischen Grundlagen visueller Aufmerksamkeit. Die Basis der Ausführungen bilden einige wenige ausgewählte Veröffentlichungen, die einen Eindruck des weitläufigen Forschungsgebietes vermitteln und auf einzelne Aspekte hindeuten können, ohne jedoch auch nur annähernd einen Anspruch auf Vollständigkeit zu erheben. Weiterführende Details finden sich z.B. in [Zan96], [Sch98b] oder auch [vdH04].

### 7.1.1 Psychologische Grundlagen

Die Untersuchung von Phänomenen der visuellen Aufmerksamkeit beim Menschen erfolgt oftmals durch die Auswertung des menschlichen Verhaltens bei der Lösung vorgegebener Probleme, beispielsweise der Suche nach spezifischen Objekten in einer Szene. Die Ergebnisse derartiger Versuche deuten nach [Itt03] grundsätzlich daraufhin, dass visuelle Aufmerksamkeit einerseits durch bildbasierte Faktoren geleitet und beeinflusst wird *(bottom-up)*, andererseits aber auch direkt von den gestellten Aufgaben abhängen kann *(top-down)*. Während spezifische Stimuli in einem Bild zumeist einen nicht bewußt wahrgenommenen Einfluss auf die Steuerung der Aufmerksamkeit ausüben, resultiert eine aufgabenabhängige Lenkung primär aus bewusst initiierten, kognitiven Prozessen, die unter anderem durch Wissen über das aktuell zu lösende Problem moduliert werden. Der anteilige Einfluss beider Faktoren auf die visuelle Aufmerksamkeit in konkreten Situationen hängt unter anderem von den spezifischen Eigenschaften und der Komplexität der präsentierten Stimuli ab, sowie von der jeweiligen Aufgabenstellung und dem Wissen darüber (vgl. [Wol03]). Grundsätzlich erscheint es beispielsweise leichter und weniger (zeit-)aufwändig, die Anwesenheit eines Objektes in einer Szene zu prüfen als das Nicht-Vorhandensein einer Objektinstanz zu verifizieren (z.B. [Wol04]).



(a) (b) (c) (d)

**Abbildung 7.1:** Vier Bilder, die den Einfluss und die Wechselwirkungen verschiedener Merkmale von Stimuli bei der Lösung von Suchaufgaben veranschaulichen (nach [Wol04]): Das schwarze „X“ in (a) kann leicht gefunden werden, und auch die Farben und Orientierungen der Geradenstücke in (b) erlauben eine einfache Segmentierung der einzelnen Regionen. Dagegen wird in (c) und (d) deutlich, dass Farbe zur Lokalisation des roten „L“ in beiden Fällen hinreichend ist, die Struktur und Orientierung des „T“ jedoch nur in (c) hervorsticht.

Visuelle Aufmerksamkeit auf Basis von Bilddaten wird nach [Wol00a] vorrangig durch spezifische Merkmale gesteuert, wie beispielsweise die Farbe, Orientierung oder auch Größe von einzelnen Objekten, sowie deren Bewegungsparametern oder möglicherweise auch Form- bzw. Struktureigenschaften. Der individuelle Einfluss spezifischer Merkmale ist dabei sehr unterschiedlich, wie die Ausführungen in [Wol04] im Hinblick auf eine klassifizierende Einordnung zeigen. Die diskriminierende Wirkung einzelner Stimuli kann in Abhängigkeit vom jeweiligen Kontext stark variieren, wobei insbesondere wechselseitige Abhängigkeiten zu anderen Merkmalen eine eindeutige Bewertung der tatsächlichen Wirksamkeit erschweren. In Abbildung 7.1 sind verschiedene Beispielszenen gezeigt, die so oder in ähnlicher Form in psychologischen Untersuchungen zur visuellen Aufmerksamkeit eingesetzt werden. Sie veranschaulichen die unterschiedlichen Einflüsse spezifischer Merkmale bei der Lenkung visueller Aufmerksamkeit. Während in den Teilbildern 7.1(a) und 7.1(b) einzelne Eigenschaften von Objekten genügen (z.B. Form bzw. Farbe oder Orientierung), um die intendierten Zielobjekte und -regionen zu lokalisieren, zeigt ein Vergleich der beiden rechten Bilder 7.1(c) und 7.1(d), dass identische Objekteigenschaften in verschiedenen Kontexten sehr unterschiedliche Einflüsse auf die visuelle Aufmerksamkeit ausüben können.

Mit wachsender Komplexität der Objekteigenschaften und Fragestellungen nimmt der Einfluss kognitiver Prozesse zu, die eine detailliertere Analyse der Einzelobjekte bedingen. Während bildbasierte Aufmerksamkeit hauptsächlich die Erfassung grober Informationen einer Szene und der darin enthaltenen Objekte erlaubt, lassen sich beispielsweise spezifischere, semantische Informationen oftmals nur durch eine explizite, bewusst gesteuerte und vorranging sequenzielle Fokussierung gewinnen. Itti schlussfolgert diesbezüglich in der Einleitung seines Artikels [Itt03]:

> „Vision thus appears to rely on sophisticated interactions between coarse, massively parallel, full-field pre-attentive analysis systems and the more detailed, circumscribed and sequential attentional analysis system.“

Neben dem direkten Einfluss, den bottom-up- und top-down-Faktoren auf die visuelle Aufmerksamkeit bei der Lösung spezifischer Fragestellungen ausüben, werden in der Psychologie auch Effekte von Aufmerksamkeit untersucht, die sich über längere Zeiträume erstrecken. Insbesondere der Einfluss einer gezielten Fokussierung einzelner Objekte auf deren spätere Wahrnehmung und daraus resultierende Konsequenzen für die Steuerung der Aufmerksamkeit sind von hohem Interesse. Derartige Verknüpfungen, die unter anderem auch Gedächtniseffekte einschließen, lassen sich beispielsweise anhand einer wiederholten Suche in mehr oder weniger stark veränderten Szenen analysieren.

Nach [Hor98] hat sich dabei in verschiedenen Experimenten gezeigt, dass die Fokussierung eines Objektes keinen direkten Einfluss auf nachfolgende Suchprozesse zu haben scheint, sich also aus einer früheren Fokussierung keine direkten Vorteile für eine spätere Suche ergeben. Insbesondere scheint visuelle Aufmerksamkeit die Wahrnehmung eines Objektes nur unmittelbar während der Fokussierung zu beeinflussen und keine direkten (inkrementellen) Veränderungen der internen, visuellen Repräsentationen zu bewirken [Wol00b]. Wolfe und Kollegen betonen jedoch explizit, dass trotz dieser Erkenntnisse *Gedächtniseffekte* sehr wohl einen Einfluss auf die wiederholte Suche von Objekten in einer Szene haben können. Es erscheint beispielsweise weitgehend unbestritten und auch durch alltägliche Erfahrungen untermauert, dass die exakte Kenntnis einer Szene oder Umgebung die Lösung von Suchaufgaben und die damit verbundene Steuerung der Aufmerksamkeit beeinflusst [Wol00b]. Aus dem Gedächtnis abgerufene Positionen von Objekten, z.B. die eines Autoschlüssels in einer Wohnung, bewirken, dass die Aufmerksamkeit bei einer Suche nach den Objekten gezielt auf wenige, potenzielle Fundorte gelenkt und die Suchaufgabe damit effizient gelöst werden kann. Die Grundlage bilden dabei vorrangig längerfristige Lern- und Trainingseffekte, die zu veränderten Herangehensweisen bei der Lösung einer Aufgabe und verbesserten Suchstrategien führen. Es sollte jedoch berücksichtigt werden, dass Aufmerksamkeit auch Einfluss auf die im Gedächtnis gespeicherten Daten ausübt [Cow97] und somit auch an diesen Prozessen indirekt beteiligt ist.

### 7.1.2 Nachbildung in technischen Systemen

Die Nachbildung von Mechanismen zur visuellen Aufmerksamkeit in künstlichen, technischen Systemen dient einerseits der Entwicklung effizienter Kontrollstrategien für aktive Sensoren, und andererseits auch einem besseren Verständnis der menschlichen visuellen Aufmerksamkeit durch Simulation. In der Literatur finden sich zahlreiche Ansätze zur Realisierung visueller Aufmerksamkeit und verschiedenste Algorithmen für eine aktive Szenenexploration [Kat94, Bou98b, Rae01]. Den Hauptansatzpunkt der meisten Arbeiten bildet dabei eine bildbasierte (bottom-up) Berechnung von (zumeist biologisch motivierten) Merkmalen auf einzelnen Bildern. Die Bildmerkmale sind dabei beispielsweise durch Orientierungen von Kanten- oder Konturelementen (z.B. [Mil94, Bou98a]), Farb- oder Intensitätsinformationen (z.B. [Che02]) sowie Bewegungsdaten (z.B. [Bol97]) gegeben. Sie werden zumeist unabhängig voneinander auf den Eingangsdaten berechnet und anschließend im Rahmen von so genannten „Aufmerksamkeitskarten“ *(saliency*

*maps)* miteinander verknüpft. Die Kombination der einzelnen Merkmalskanäle erfolgt dabei in vielen Fällen unter Verwendung von neuronalen Ansätzen, die den Aktivitäten im menschlichen Gehirn nachempfunden sind (z.B. [Ahm91, Egn97, Itt01]). Die finale Entscheidung, welcher Punkt im nachfolgenden Zeitschritt von der Kamera fixiert werden soll *(Fokuspunkt)*[1], wird abschließend auf Basis der berechneten Merkmalskarten getroffen, wobei zumeist das Maximum in der Karte den Ausschlag gibt [Itt01, Rae01].

Die Beschränkung auf eine reine Auswertung von Bildinformationen zur Lenkung visueller Aufmerksamkeit spiegelt die biologischen Zusammenhänge beim Menschen nur unzureichend wider (vgl. vorhergehenden Abschnitt). Auch die einer Szenenexploration zu Grunde liegende Aufgabe hat zumeist einen signifikanten Einfluss auf die Auswahl von Fokuspunkten in einer Szene. Eine Berücksichtigung derartiger Effekte in technischen Systemen basiert oftmals auf einer gezielten Steuerung der Merkmalsberechnung. Diese kann einerseits durch eine spezifische Gewichtung einzelner Merkmalskarten realisiert werden, oder auch durch die Integration zusätzlicher Karten, in denen die für die Lösung der Aufgabe wichtigen Informationen explizit kodiert werden [Rae01, Nav02].

Wie einleitend bereits angedeutet wurde, beschränkt sich die Merkmalsberechnung in nahezu allen bisherigen Arbeiten auf die Daten des aktuellen Eingangsbildes (und ggf. seines direkten Vorgängers bei einer Bewegungsdetektion mittels Differenzbildern), so dass Bildinformationen früherer Zeitpunkte keinen direkten Einfluss auf die Auswahl der Fokuspunkte ausüben können. Zwar sind oftmals zusätzlich Mechanismen zur gezielten Ausblendung von Bildbereichen integriert, die beispielsweise eine kontinuierliche Fokussierung einzelner Positionen verhindern sollen und damit zeitliche Zusammenhänge einbeziehen, sie sind jedoch zumeist unabhängig von den konkreten Bilddaten. Auch wenn psychologische Untersuchungen zeigen, dass die Auswirkungen visueller Aufmerksamkeit tatsächlich zeitlich nur sehr begrenzt sind, stellt eine Beschränkung der Merkmalsberechnung und damit auch der Fokuspunkt-Selektion auf ausschließlich aktuelle Daten eine unnötige Beschneidung der Flexibilität interaktiver Systeme dar. Zumindest im Gedächtnis gespeicherte, visuelle Daten früherer Zeitpunkte üben offenbar einen Einfluss auf aktuelle Aufmerksamkeitsstrategien aus. Im Hinblick auf interaktive Systeme liegt es damit (und nicht zuletzt auch aufgrund der daraus resultierenden, technischen Vorteile) nahe, die Selektion von Fokuspunkten auf komplette Bildfolgen auszudehnen und die in den Sequenzen vorhandenen Daten besser zu nutzen.

Die vorliegende Arbeit beschreibt das Konzept eines visuellen Speichers, der eine effiziente, ikonische Repräsentation visueller Daten ermöglicht, die mit aktiven Kameras aufgenommen wurden. Durch die Integration der Bildsequenzen in Mosaikbilder stehen interaktiven Systemen visuelle Informationen zur Verfügung, die sich über einen längeren Zeitraum erstrecken und insbesondere auch das Sichtfeld der Kamera im Raum erweitern. Eine Berechnung von Fokuspunkten auf den Daten eines solchen, visuellen Speichers führt damit zu der zuvor aufgezeigten, größeren Flexibilität bei der Datenauswertung.

In den nachfolgenden Abschnitten wird die Implementierung eines derartigen Konzepts zur Realisierung von visueller Aufmerksamkeit für interaktive Systeme auf Basis

[1] Die Suche nach Fokuspunkten wird oft auch als „Where-to-look-next“-Problem bezeichnet.

der neu entwickelten Multi-Mosaikbilder vorgestellt. Zuvor sei jedoch explizit darauf verwiesen, dass diese Implementierung nicht die in psychologischen Untersuchungen zu beobachtenden Zusammenhänge zwischen Aufmerksamkeit, Gedächtnis und der Effizienz von Problemlösungen widerspiegeln kann und soll. Die Grundidee findet zwar durchaus Parallelen im biologischen Vorbild, die reale Umsetzung orientiert sich jedoch primär an den Anforderungen technischer Systeme. Dort verhilft eine Ausweitung der Selektion von Fokuspunkten auf zeitlich integrierte Daten vorranging zu einer deutlichen Effizienzsteigerung bei der Verarbeitung visueller Informationen. So kann auf diese Weise beispielsweise der mit einem expliziten Einsatz von Hardware (z.B. Kamerabewegungen, Fahrmanöver mobiler Roboter) unter Umständen verbundene Aufwand bei Such- und Explorationsaufgaben signifikant vermindert werden (vgl. auch Kap. 8).

## 7.2 Szenenexploration mit Multi-Mosaikbildern

Die in diesem Kapitel vorgestellte Realisierung von Algorithmen zur aktiven Szenenexploration und zur Nachbildung visueller Aufmerksamkeit auf Basis von Multi-Mosaikbildern verfolgt im Wesentlichen zwei Zielsetzungen. Vorrangig soll sie die grundlegenden Vorteile einer flexibleren Auswahl von Fokuspunkten auf Basis zeitlich integrierter ikonischer Daten aufzeigen, die in dieser Form bislang noch nicht publiziert wurde. Als Nebeneffekt resultiert dabei gleichzeitig aber auch eine effiziente Strategie zur Akquisition visueller Informationen einer Szene als Basis für den Aufbau einer ikonischen Repräsentation. Durch den der Datenaufnahme zu Grunde liegenden Explorationsalgorithmus werden die visuellen Informationen in einer Szene implizit nach ihrer Interessantheit klassifiziert und damit in eine definierte Reihenfolge gebracht. Die vermeintlich wichtigsten Informationen stehen interaktiven Systemen somit zuerst zur Verfügung, so dass sie frühzeitig auf potenziell interessante Teilbereiche in der Szene hingewiesen werden. Insbesondere ein Einsatz von Zoom (Abschnitt 7.3.3) ermöglicht in diesem Zusammenhang die Aufnahme sehr spezifischer Daten, die beispielsweise im Rahmen einer späteren Objekterkennung vorteilhaft sein können (Kap. 8).

Den Ausgangspunkt zur Umsetzung der aktiven Szenenexploration auf Basis der Multi-Mosaikbilder bildet eine Erweiterung des entwickelten Konzeptes (Kap. 6) um zusätzliche Strukturen für Daten, die zur Selektion von Fokuspunkten in einer Szene herangezogen werden können (vgl. auch [Möl05]). Die Repräsentation dieser Zusatzinformationen erstreckt sich dabei, analog zu der des ursprünglichen Speichers, ebenfalls über den vollständigen Sichtbereich einer stationären, rotierenden Kamera. Auf diese Weise lassen sich im Rahmen der Fokuspunkt-Selektion effiziente Analysen aller im zeitlichen Verlauf akquirierten und in die Mosaikbilder integrierten, ikonischen Daten durchführen. Insbesondere wiederholte Zugriffe auf die Hardware können somit vermieden werden, die beispielsweise im Kontext mobiler Roboter oftmals einen nicht zu unterschätzenden Aufwand mit sich bringen. Im nachfolgenden Abschnitt wird zunächst die vorgenommene Erweiterung der Datenstrukturen skizziert, bevor in Abschnitt 7.2.2 Beschreibungen der verwendeten Interessantheitsmaße folgen.

### 7.2.1 Grundidee und Datenstrukturen

Die Repräsentation von Informationen zur Auswahl interessanter Fokuspunkte in Multi-Mosaikbildern orientiert sich grundsätzlich an einer in der Literatur weit verbreiteten und auch biologisch motivierten Vorgehensweise. Die Speicherung verschiedener Merkmale erfolgt in einzelnen Karten, die für die Selektion eines Fokuspunktes miteinander kombiniert werden. Da zur Aufnahme der Bilddaten in der vorliegenden Arbeit eine stationäre, rotierende Kamera Anwendung findet, erstrecken sich die Merkmale dabei prinzipiell über den vollständigen Sichtbereich dieser Kamera. Zur Speicherung der Daten dienen daher, in Anlehnung an die grundlegende Struktur der Multi-Mosaikbilder, polyedrisch angeordnete Mengen von Ebenen für die verschiedenen Merkmale. Für jedes Merkmal, das bei der Selektion von Fokuspunkten im Verlauf einer Szenenexploration berücksichtigt werden soll, wird der Multi-Mosaik-Datenstruktur im Grundsatz eine solche Menge, eine so genannte „Merkmalsinstanz“, zugefügt. Die Merkmalsinstanzen $M^j_{f_M}, j = 1 \ldots m$, werden parallel zu den Projektionsinstanzen $S_{f_i}$ (vgl. Abschnitt 6.2 auf S. 93) um das optische Zentrum der Kamera angeordnet, wobei sie jedoch nicht alle durch die Projektionsinstanzen vorgegebenen Auflösungen abdecken. Sie beschränken sich vielmehr auf eine einzelne grobe Auflösung $f_M$. Die daraus resultierende, vorrangig speichereffiziente Darstellung hat sich als ausreichend erwiesen, um einzelne Regionen in einer Szene im Hinblick auf ihre Interessantheit zu bewerten (vgl. auch Unterkap. 7.4). Eine hochauflösendere Beurteilung jedes einzelnen Szenenpunktes ist im Wesentlichen lediglich mit einem größeren Aufwand bei der Berechnung verbunden und bedingt ein größeres Datenvolumen, das es bei der Fokuspunkt-Suche zu analysieren gilt.

Im nachfolgenden Abschnitt werden die in der derzeitigen Implementierung verwendeten Interessantheitsmaße im Detail vorgestellt. Die Hauptzielsetzung der Implementierung besteht dabei in der Demonstration der Eignung des visuellen Speichers zur Realisierung von visueller Aufmerksamkeit unter Berücksichtigung zeitlich integrierter Daten. Der Entwurf neuer Interessantheitsmaße, etwa zur exakteren Nachbildung der menschlichen visuellen Aufmerksamkeit, steht dagegen nicht im Mittelpunkt dieser Arbeit, so dass die Implementierung auf verhältnismäßig einfachen Selektionskriterien beruht. Diese sind dennoch gut geeignet, um die Leistungsfähigkeit des Ansatzes und Perspektiven für zukünftige Weiterentwicklungen aufzuzeigen (vgl. Unterkap. 7.4). Darüber hinaus lässt sich die bisherige, modulare Implementierung ohne Schwierigkeiten durch Hinzufügen neuer Merkmalsinstanzen um zusätzliche Bewertungskriterien erweitern.

### 7.2.2 Interessantheitsmaße

Der Fokuspunkt-Selektion liegen in der derzeitigen Implementierung im Wesentlichen vier Bewertungskriterien zu Grunde. Neben der lokalen Entropie finden Bewegungsdaten sowie der zeitliche und räumliche Kontext bei der Auswahl Berücksichtigung (s. auch Abschnitt 7.3). Zur Speicherung der lokalen Zeitstempel und Entropien dienen dabei die Merkmalsinstanzen $M^1_{f_M}$ und $M^2_{f_M}$, in denen die entsprechenden Daten abgelegt werden. Informationen über unabhängig bewegte Objekte in dem aktuell betrachteten Szenenaus-

schnitt sind dagegen bereits im Korrespondenzgraphen repräsentiert, der in Abschnitt 5.2.1 eingeführt wurde. Sie werden daher direkt aus diesem entnommen und nicht zusätzlich nochmals explizit in den Multi-Mosaikbildern gespeichert. Auch der räumliche Kontext geht in Form einer spezifischen Gewichtungsfunktion in die Fokuspunkt-Selektion ein, ohne die explizite Verwaltung einer zugehörigen Merkmalsinstanz erforderlich zu machen.

### Entropie

Die Entropie $H$ entstammt ursprünglich der Informationstheorie (z.B. [Mac03], S. 32ff.). Sie quantifiziert den mittleren Informationsgehalt (Shannon Information) einer gesendeten Nachricht, die aus einer Zeichenfolge $S = s_1 s_2 \dots s_n$ mit $n$ Zeichen besteht. Die Zeichen gehören dabei einem endlichen Alphabet $A = \{a_1, a_2, \dots, a_{n_A}\}$ mit $n_A$ Elementen an. $H$ wird auf dieser Basis in Abhängigkeit von der Auftrittswahrscheinlichkeit $p_S(a_i)$ der einzelnen Zeichen $a_i$ in $S$ und deren Informationsgehalt $J_S(a_i)$ wie folgt berechnet[2]:

$$H = \sum_{i=1}^{n_A} p_S(a_i) \cdot J_S(a_i) = \sum_{i=1}^{n_A} -p_S(a_i) \log p_S(a_i). \tag{7.1}$$

Der Informationsgehalt $J_S(a_i)$ eines Zeichens ist dabei als negativer Logarithmus der Auftrittswahrscheinlichkeit definiert. Diese Definition basiert auf einem Modell, nach dem der Informationsgehalt eines Zeichens mit einer abnehmenden „Zufälligkeit“ seines Auftretens, d.h. einer zunehmenden Auftrittswahrscheinlichkeit, sinkt. Wenn beispielsweise bei einer binären Datenquelle jeweils ausschließlich Nullen oder ausschließlich Einsen emittiert werden, so ist deren Auftrittswahrscheinlichkeit jeweils 1. Der Informationsgehalt der Nachrichtenquelle ist in diesen Fällen jedoch gleich 0, da mit der Emission neuer Zeichen kein zusätzlicher Informationsgewinn verbunden ist. Analog dazu erreicht die Entropie maximale Werte, wenn alle emittierbaren Zeichen mit derselben Wahrscheinlichkeit beobachtet werden können.

In der digitalen Bildverarbeitung kann die Entropie zur Bewertung des Informationsgehaltes lokaler Bildregionen $R$ herangezogen werden. Die Menge der möglichen Grauwerte[3] $G = \{g_0 = 0, \dots, g_{255} = 255\}$ der einzelnen Bildpixel $(x, y) \in R$ bilden dabei das Alphabet emittierbarer Zeichen, deren Auftrittswahrscheinlichkeiten $p_R(g_i)$ sich aus den relativen Häufigkeiten der einzelnen Werte in der ausgewählten Region $R$ schätzen lassen:

$$\hat{p}_R(g_i) = \frac{1}{|R|} \sum_{(x,y) \in R} \delta_{I(x,y),g_i}. \tag{7.2}$$

[2] Für das Auftreten der einzelnen Zeichen wird dabei statistische Unabhängigkeit vorausgesetzt.

[3] Die Berechnung der Entropie wurde an dieser Stelle zur Vereinfachung und aus Effizienzgründen auf die Grauwerte von Bildpixeln beschränkt, sie kann jedoch bei Bedarf ohne Schwierigkeiten auch auf Farbinformationen ausgeweitet werden.

Hohe Entropiewerte innerhalb einzelner Bildregionen, die typischerweise eine Größe von etwa $15 \times 15$ Pixeln aufweisen, deuten auf eine große Vielfalt vorkommender Grauwerte hin (Abb. 7.2). Damit lassen sich in einer Szene insbesondere homogene, kontrast- und strukturarme Regionen, die zumeist wenig Interesse beim Betrachter erwecken (etwa einfarbige Wände), von inhomogenen, stärker texturierten Bereichen unterscheiden, die tendenziell mit einem höheren Informationsgehalt assoziiert werden (z.B. Bilder an den Wänden).

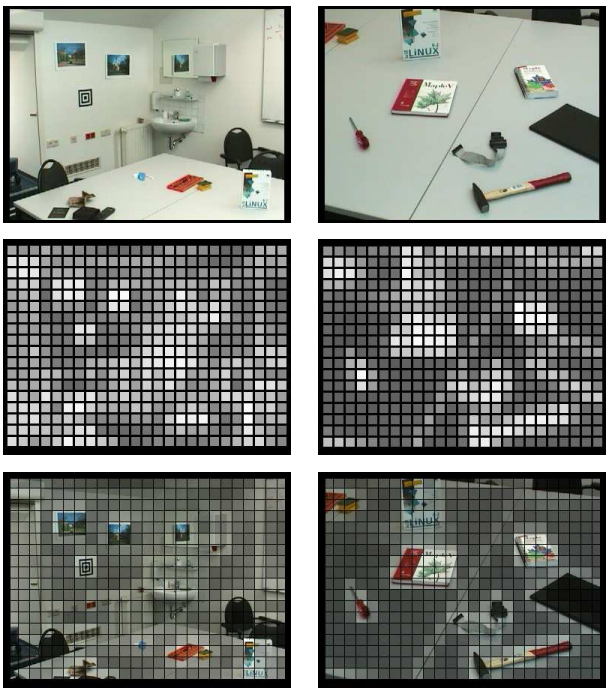


**Abbildung 7.2:** Zwei Beispielbilder (oben) und ihre Entropiekarten (Mitte). Die untere Zeile zeigt zur Verdeutlichung modifizierte Darstellungen der Bilder, die durch eine Gewichtung ihrer Farbwerte mit den lokalen Entropien generiert wurden.

### Bewegungen

Das visuelle System des Menschen zeigt eine besondere Sensitivität im Hinblick auf eine Wahrnehmung von Bewegungen. Während viele Informationen einer Szene nur durch eine vorherige, direkte Fokussierung aufgenommen werden können, genügen zur Bewegungserkennung schon die im Außenbereich der Netzhaut angesiedelten, nur sehr grob auflösenden Nervenzellen aus ([Sch90], S. 315ff.). Diese Eigenschaften des visuellen Systems werden dabei sehr früh in der Entwicklung ausgebildet, da bereits bei wenige Monate alten Säuglingen Reaktionen auf Bewegungen in ihrem Sichtfeld zu beobachten sind ([Ber03], S. 150ff.). Bewegungsdaten sind damit offenbar ein wichtiger Faktor bei der Lenkung der visuellen Aufmerksamkeit eines Menschen und als solche auch bei einer Nachbildung in technischen Systemen von hoher Bedeutung.

Eine der Hauptzielsetzungen bei der Entwicklung des visuellen Speichers in dieser Arbeit ist durch eine adäquate, möglichst vollständige ikonische Modellierung des statischen Hintergrundes einer Szene gegeben (vgl. Kap. 5). In diesem Zusammenhang ist insbesondere eine stetige Überprüfung der Konsistenz und Korrektheit der jeweils aktuellen Repräsentation im zeitlichen Verlauf bedeutsam. Änderungen in der Struktur des Szenenhintergrundes werden durch verschiedenste Ursachen bedingt, die jedoch in aller Regel mit unabhängigen Bewegungen in der Szene korrelieren. Damit erlauben Bewegungsdaten einer zu modellierenden Szene Rückschlüsse auf mögliche Strukturveränderungen innerhalb des statischen Szenenhintergrundes (Details siehe Abschnitt 5.2.2). Sie stellen somit nicht nur aus biologischer Sicht einen wichtigen Faktor zur Lenkung der visuellen Aufmerksamkeit in der hier skizzierten Implementierung dar. Die Informationen über unabhängige Bewegungen in einer Szene resultieren dabei aus der Anwendung der

im fünften Kapitel vorgestellten Detektions- und Trackingalgorithmen, wobei die Informationen des Korrespondenzgraphen direkt bei der Fokuspunkt-Selektion berücksichtigt werden (vgl. auch Unterkapitel 7.3).

### Zeitlicher Kontext

Das Ziel einer aktiven Szenenexploration kann einerseits in einer durch eine spezifische Aufgabe induzierten Suche nach geeigneten Informationen zur Problemlösung bestehen, oder auch einfach das selbständige Erkunden und Kennenlernen einer unbekannten Umgebung umfassen. Insbesondere dieser zweite Anwendungsfall erfordert dabei Mechanismen, die die Aufmerksamkeit des Systems im zeitlichen Verlauf sukzessive auch auf neue oder für längere Zeit nicht fokussierte Teilbereiche in einer Szene lenken und damit gleichsam eine Art „Neugier“ des Systems im Hinblick auf unbekannte Orte bewirken.

Die Berücksichtigung des zeitlichen Kontextes in der hier vorgestellten Implementierung erfolgt durch eine zeitliche Gewichtung der Interessantheit einzelner Punkte bei der Fokuspunkt-Selektion in einer Szene. Dazu werden in der Merkmalsinstanz $M^1_{f_M}$ für alle ausgewählten Szenenpunkte Zeitstempel gespeichert, die auf den Zeitpunkt verweisen, zu dem die entsprechenden Punkte jeweils zuletzt von der Kamera fokussiert wurden. Aus den absoluten Differenzen zwischen diesen Zeitstempeln und der aktuellen, internen Systemzeit können dann direkt die Zeiträume abgelesen werden, in denen die einzelnen Szenenpunkte nicht exploriert wurden. Sie lassen sich unter anderem in Form zusätzlicher Gewichtungsfaktoren in die globale Bewertung der Punkte einbeziehen.

### Räumlicher Kontext

Die Lenkung der Aufmerksamkeit des Menschen bei Problemlösungen oder im Rahmen von Kommunikationssituationen lässt sich im Wesentlichen anhand der Trajektorien der durchgeführten Blicke nachvollziehen (z.B. [Vel96]). Die Charakteristika der Blicktrajektorien erlauben dabei grundsätzlich eine Unterscheidung von zwei verschiedenen Arten von Augenbewegungen: schnelle, ruckartige „Sakkaden“ und langsamere, insgesamt weichere Bewegungen [Kow95]. Während die Sakkaden vorrangig der direkten Fokussierung verschiedener Stimuli dienen, finden die stetigeren Bewegungen etwa bei der Verfolgung von bewegten Objekten Anwendung. Sowohl Sakkaden wie auch die langsameren Augenbewegungen werden zumeist mit Kopfbewegungen kombiniert, da sich auf diese Weise eine höhere Performanz erzielen lässt.

Im Hinblick auf die Steuerung einer aktiven Kamera finden sich sowohl Anwendungsfälle für schnelle, sakkadische Bewegungen wie auch für langsamere, im Allgemeinen weniger weitreichende Kamerarotationen. Für die Kamerabewegungen bei der Aufnahme visueller Daten im Rahmen des visuellen Speichers ist es dabei insbesondere hinsichtlich großer sakkadischer Bewegungen ratsam, maximale Abstände zwischen aufeinander folgenden Fokuspunkten festzulegen. Einerseits soll zwar die Flexibilität bei der Auswahl von Fokuspunkten auf Basis auch außerhalb des aktuellen Blickfeldes liegender Daten so wenig wie möglich eingeschränkt werden. Andererseits lassen sich sehr große Sprünge

zumeist nur schrittweise realisieren, um die Mosaikbildberechnung nicht zu erschweren und die räumlichen und zeitlichen Zusammenhänge in den Bilddaten, etwa aufgrund einer möglicherweise fehlgeschlagenen Registrierung, nicht zu verlieren. Der daraus resultierende, zusätzliche Zeitaufwand bei großen sakkadischen Bewegungen sollte daher immer in Bezug zum aktuell gegebenen Anwendungskontext und der damit verbundenen, tatsächlichen Notwendigkeit der Sprünge gesetzt werden.

Die vorstehenden Überlegungen bilden den Ausgangspunkt für eine explizite Gewichtung jedes Punktes $\tilde{p} = (x, y)$ in einer Szene gemäß seiner räumlichen Distanz $r_{x,y}$ zum aktuell fokussierten Punkt $\tilde{p}_f = (x_f, y_f)$. $r_{x,y}$ kann dabei etwa durch den quadratischen euklidischen Abstand, $r_{x,y} = (x - x_f)^2 + (y - y_f)^2$, oder als Einschlußwinkel zwischen den (3D-)Ortsvektoren von $\tilde{p}$ und $\tilde{p}_f$ definiert werden. Als Gewichtungsfunktion dient eine invertierte „Mexikanerhut"-Funktion $h(x, y)$, die auch in Abbildung 7.3 dargestellt ist (vgl. [Zei96]):

$$h_{\tilde{p}_f}(x, y) = -(1 - d \cdot r_{x,y}) \cdot e^{-\frac{d}{2} \cdot r_{x,y}}. \tag{7.3}$$

In der Gleichung definiert $d$ einen Gewichtungsfaktor, über den sich die räumliche Ausdehnung der Funktion variieren lässt, so dass die maximale Größe zulässiger Kamerabewegungen in Abhängigkeit vom aktuellen Anwendungsfeld flexibel reguliert werden kann. Durch die Funktion werden drei konzentrische Regionen um den aktuellen Fokuspunkt herum definiert, die sich jeweils in den Vorzeichen ihrer Gewichtung unterscheiden. Pixel in den inneren und äußeren beiden Bereichen erhalten eine negative Gewichtung, so dass sowohl große Sprünge wie auch die kontinuierliche Fokussierung eines einzelnen Szenenpunktes unterbunden werden. Insbesondere der zweite Aspekt ist dabei eng mit der Berücksichtigung des zeitlichen Kontextes verknüpft (vgl. vorhergehenden Unterabschnitt). Bildpunkte innerhalb der mittleren Zone werden positiv bewertet und damit bei der Auswahl des nächsten Fokuspunktes bevorzugt.

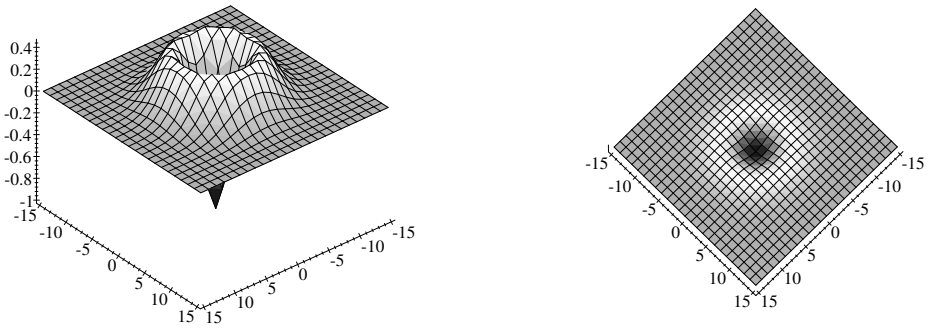



**Abbildung 7.3:** Die inverse „Mexikanerhut"-Funktion zur Berücksichtigung räumlicher Distanzen bei der Fokuspunkt-Selektion: Sie definiert im Wesentlichen drei „Zonen" (grob schwarz, weiss und grau markiert), die sich durch die Vorzeichen ihrer Gewichtung voneinander unterscheiden.

## 7.3 Explorationsstrategien

Mit den zusätzlichen Merkmalsinstanzen innerhalb der Multi-Mosaikbilder steht interaktiven Systemen eine geeignete Datenbasis zur Realisierung einer durch visuelle Auf-

merksamkeit geleiteten Szenenexploration zur Verfügung. Neben den reinen Daten sind dafür jedoch zusätzlich Mechanismen und Heuristiken zur Auswertung dieser Daten notwendig. Dabei stehen insbesondere Fragen der finalen Auswahl einzelner Fokuspunkte und die Aktualisierung der Daten im zeitlichen Verlauf einer Explorationssitzung im Vordergrund.

Die Hauptzielsetzung einer aktiven Szenenexploration im Rahmen dieser Arbeit besteht in einer geeigneten Kamerasteuerung zur Akquisition visueller Daten einer Szene, die die Basis einer ikonischen Szenenrepräsentation durch Multi-Mosaikbilder bilden. Ein interaktives System soll dabei zunächst primär interessante Bereiche innerhalb einer Szene explorieren, die mit einem hohen Informationsgehalt assoziiert werden, im weiteren zeitlichen Verlauf aber auch Regionen fokussieren, die zwar weniger Interesse wecken, aber dennoch potenziell wichtige Daten enthalten können. Im Hinblick auf eine korrekte Modellierung des statischen Szenenhintergrundes sind darüber hinaus Konsistenzprüfungen der aktuellen Hintergrundrepräsentation ein wichtiger Aspekt. Die Daten der ikonischen Repräsentation sollen jeweils, mit möglichst geringem Zeitversatz, auch an veränderte Strukturen innerhalb des Hintergrundes angepasst werden. Hinweise für derartige Änderungen liefern dabei insbesondere Informationen über unabhängige Bewegungen in einer Szene, denen daher hier eine besondere Bedeutung zukommt.

Die vorstehenden Überlegungen bilden die Grundlage für eine Unterteilung der Exploration in drei mögliche Zustände, in denen sich das System befinden kann. Diese Aufteilung wird zunächst im nachfolgenden Abschnitt skizziert, bevor in Abschnitt 7.3.2 eine Beschreibung der realisierten Ansätze zur Fokuspunkt-Selektion und zur kontinuierlichen Aktualisierung der gespeicherten Daten im zeitlichen Verlauf folgt. Abschließend zeigt Abschnitt 7.3.3 zusätzliche Möglichkeiten auf, wie neben reinen Pan- und Tilt-Bewegungen der Kamera bei einer festgewählten Bildauflösung auch Zoom effektiv im Rahmen der aktiven Szenenexploration eingesetzt werden kann.

### 7.3.1 Definition der Systemzustände

Die Selektion eines neuen Fokuspunktes in einer betrachteten Szene orientiert sich an den im vorhergehenden Abschnitt skizzierten Zielsetzungen. Zur algorithmischen Umsetzung dieser Ziele wurden im Kern drei verschiedene Zustände implementiert, in denen sich das System befinden kann und die in Abbildung 7.4 veranschaulicht sind. Die Zustände „ExpN“ („normale“ Exploration) und „ExpA“ (Exploration mit erhöhter Aufmerksamkeit) dienen dabei vorrangig einer adäquaten Behandlung dynamischer Daten, die im gegebenen Anwendungskontext einen besonderen Stellenwert einnehmen. Der dritte Zustand „ExpD“ umfasst dagegen eine Detailexploration der Szene, die dem lokal variierenden Informationsgehalt in den visuellen Daten Rechnung trägt und in Abschnitt 7.3.3 genauer beschrieben wird.

Zustandsänderungen des Systems werden primär durch hohe lokale Entropiewerte oder Bewegungsereignisse innerhalb einer Szene induziert. Während dem ersten Fall, der erst in Abschnitt 7.3.3 genauer betrachtet wird, eine Schwellwert-basierte Auswertung der

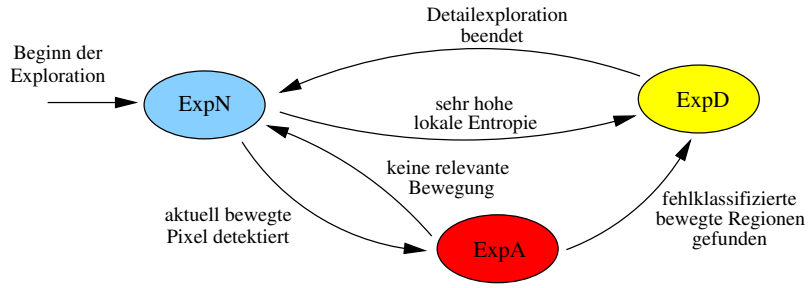



**Abbildung 7.4:** Skizze des implementierten Zustandsdiagramms für die aktive Szenenexploration: Die beiden Zustände „ExpN“ und “ExpA“ dienen zur Trennung von Situationen, in denen die Szene nur statisch ist, und Zeitpunkten, zu denen unabhängige Bewegungen präsent sind. Der Zustand „ExpD“ schließlich umfasst lokale Detailexplorationen unter Verwendung des Kamerazooms.

Entropie zu Grunde liegt, bilden im zweiten Fall die Bewegungsdaten aus dem Korrespondenzgraphen den Ausgangspunkt (vgl. Kap. 5.2.1). Der Normalzustand des Systems ist dabei der Zustand „ExpN“, in dem eine durch die lokale Entropie geleitete Exploration der Szene in grober Auflösung stattfindet. Sobald Bewegungen in der Szene registriert werden, d.h. wenn eine empirisch ermittelte Mindestanzahl von Pixeln ($\approx 350$) als „aktuell bewegt“ klassifiziert wird, findet ein Wechsel in den Zustand „ExpA“ erhöhter Aufmerksamkeit statt. Dabei wird die zuvor ausgeführte Exploration zwar grundsätzlich fortgeführt, jedoch mit in ihrer Reichweite signifikant reduzierten Kamerabewegungen. Diese Maßnahme bindet die Aufmerksamkeit des Systems an den Szenenbereich, in dem die Bewegungen detektiert wurden, ohne diese jedoch unmittelbar zu fokussieren. Damit lässt sich eine direkte Systemreaktion auf nicht relevante Bewegungen unterbinden, wie sie etwa durch Rauschen oder veränderte Beleuchtungsverhältnisse verursacht werden können. Der betroffene Bereich wird zunächst nur im Sichtfeld der Kamera gehalten, so dass eine weitere Analyse Aufschluss über die tatsächliche Relevanz geben kann.

Resultieren aus einer solchen Analyse der Bewegungsdaten keine Hinweise auf tatsächliche Modifikationen der Szenenstruktur, d.h. werden die Bewegungsinformationen als nicht bedeutsam eingestuft, so wechselt das System nach wenigen Zeitschritten zurück in den Zustand „ExpN“ normaler Exploration. Fehlklassifizierte bewegte Regionen, die mit Hilfe der Trajektorienanalyse (Abschnitt 5.2.2) detektiert werden können, bewirken dagegen einen Übergang in den Zustand „ExpD“, in dem eine Detailexploration der korrespondierenden Bereiche stattfindet.

### 7.3.2 Punktselektion und Aktualisierungsheuristiken

Im Rahmen der Zustände „ExpN“ und „ExpA“ findet eine Szenenexploration in grober Auflösung statt. Zur Lokalisation neuer Fokuspunkte $\tilde{p}_n$ werden dabei in jedem Zeitschritt die zuvor skizzierten Aufmerksamkeitsmaße gemäß der nachfolgenden Berechnungsvorschrift ausgewertet:

$$\tilde{p}_n := (x_n, y_n) = \underset{(x,y)}{\operatorname{argmax}} \left( M^2_{f_M}(x,y) \cdot (t - M^1_{f_M}(x,y) + \epsilon) \cdot h_{\tilde{p}_f}(x,y) \right). \qquad (7.4)$$

$M^1_{f_M}$ und $M^2_{f_M}$ bezeichnen die Merkmalsinstanzen innerhalb des Multi-Mosaikbildes, in denen jeweils die Zeitstempel der letzten Fokussierung eines Punktes sowie dessen aktuelle, lokale Entropie gespeichert sind. Die Entropie wird dabei auf einen Wertebereich von 0 bis 1 normiert. $t$ entspricht der aktuellen, internen Systemzeit, wobei der Parameter $\epsilon$ einen Offset darstellt, der den Einfluss einer verschwindenden Zeitdifferenz abschwächt. $h_{\tilde{p}_f}(x, y)$ schließlich ist die weiter oben eingeführte Mexikanerhut-Funktion, die den räumlichen Abstand zwischen dem aktuellen Fokuspunkt $\tilde{p}_f$ und den potenziellen Zielpunkten der nachfolgenden Kamerabewegung bewertet. Der nächste Fokuspunkt ergibt sich nach dieser Vorschrift jeweils als Szenenpunkt mit maximaler Bewertung.

Die Verwaltung und Aktualisierung der Aufmerksamkeitsdaten über die Zeit folgt denselben Prinzipien, die auch bei der Verwaltung der Bilddaten des Multi-Mosaikbildes Anwendung finden (s. Abschnitt 6.4). Um die Handhabung der Daten zu erleichtern, wird analog zur Fokus-Bildebene eine Aufmerksamkeitsebene verwaltet, die dieselbe Position und Orientierung wie die Fokus-Bildebene aufweist, jedoch die reduzierte Auflösung der Merkmalsinstanzen widerspiegelt. Sie wird parallel zu der eigentlichen Fokus-Bildebene aktualisiert, wobei auch der Datenaustausch mit den Merkmalsinstanzen analog verläuft.

In jedem Zeitschritt wird jeweils der Bereich der Aufmerksamkeitsebene aktualisiert, der sich gerade im Kamerabild befindet. Die in der Ebene gespeicherten Interessantheitsdaten werden dabei mit den neuen Werten überschrieben, wodurch eine Anpassung an die aktuellen Gegebenheiten in der Szene stattfindet. Die explizite Speicherung der Zeitmarken, die den jeweils letzten Zeitpunkt vermerken, zu dem ein Punkt von der Kamera fokussiert wurde, erlaubt dabei eine effiziente, implizite Aktualisierung *aller* gespeicherten Daten. Während bei der Integration neuer Daten die Zeitmarken innerhalb des direkt fokussierten Bereiches inkrementiert werden, bleiben die Zeitstempel der nicht betrachteten Bereiche unverändert. Die Differenz zur aktuellen Systemzeit wächst dort somit stetig an, so dass das Gewicht der Punkte im zeitlichen Verlauf implizit zunimmt.

### 7.3.3 Lokale Detailanalyse

Menschliche Aufmerksamkeit zur Selektion visueller Informationen basiert auf einem Wechselspiel zwischen einer groben Szenenexploration einerseits, und einer lokalen Detailanalyse andererseits. Die grobe Exploration dient dabei der Erfassung von grundlegenden Strukturen in einer Szene und verhilft dem Menschen damit beispielsweise zu einer schnellen Orientierung in unbekannten Umgebungen. Um jedoch detailliertere Informationen, etwa zur Lösung spezifischer Fragestellungen zu finden, bietet eine solche Vorgehensweise keine ausreichende Flexibilität. Hierzu sind vielmehr lokal sehr begrenzte, hoch auflösende Fokussierungsmechanismen notwendig. Erst durch sie wird es etwa möglich, nach der Groblokalisation eines Bücherregals in einem Raum auch die Suche nach einem bestimmten Buch anhand seines Titels zu beginnen.

Die im Rahmen dieser Arbeit auf Basis der Multi-Mosaikbilder implementierte Szenenexploration hat primär die aktive Akquisition von Bilddaten zum Aufbau einer adäquaten, ikonischen Szenenrepräsentation zum Ziel. Die in Abschnitt 7.3.2 skizzierten Krite-

rien zur Fokuspunkt-Selektion realisieren dabei eine grobe Explorationsstrategie, mit der ein globaler Überblick über eine Szene gewonnen werden kann. Um zusätzlich auch der lokalen Varianz im Informationsgehalt der visuellen Informationen einer Szene gerecht werden zu können, werden in diesem Abschnitt weiterführende Explorationsstrategien vorgestellt. Basierend auf einer aktiven Steuerung des Kamerazooms ermöglichen sie die Aufnahme ausgewählter Detailinformationen und bilden damit die Grundlage für eine automatische, aufmerksamkeitsgesteuerte Aufnahme von Multi-Mosaikbildern mit verschiedenen Auflösungsebenen (vgl. auch Unterkap. 6.2).

Die Auswahl von interessanten Punkten in einer Szene, die sich für eine Detailexploration empfehlen, wird oftmals stark von zu Grunde liegenden Aufgaben und Problemstellungen geprägt. Durch die signalnahe, ikonische Datenrepräsentation des visuellen Speichers und die Möglichkeit einer direkten Anwendung vielfältiger Bildverarbeitungsmodule eröffnen sich zahlreiche Perspektiven für verschiedenste Anwendungsmöglichkeiten. Sobald eine konkrete Zielapplikation ausgewählt wurde (wie etwa eine Szenenmodellierung zur Unterstützung ansichts-basierter Objekterkennungsverfahren, s. Kap. 8), können applikationsspezifische Explorationsstrategien realisiert werden, die eine zielgerichtete Akquisition der jeweils benötigten, hoch aufgelösten Daten erlauben. Ohne eine solche Applikation sind dagegen kaum externe Anhaltspunkte für eine Auswahl von derartigen Bereichen in einer Szene gegeben. Die Selektion von einzelnen Fokuspunkten, an denen eine Detailexploration lohnenswert erscheint, orientiert sich daher im Rahmen des hier vorgestellten Ansatzes primär an der grundlegenden Zielsetzung des visuellen Speichers und den Eigenschaften der derzeit berücksichtigten Interessantheitsmaße.

Ein Hauptanliegen bei der Entwicklung des visuellen Speichers ist die adäquate Repräsentation des statischen Hintergrundes einer Szene. Dies schließt insbesondere eine stetige Überprüfung der Konsistenz der aktuellen Repräsentation ein. Wie bereits in Abschnitt 7.2.2 skizziert wurde, sind in diesem Zusammenhang insbesondere unabhängige Bewegungen in einer Szene von hohem Interesse. Die Aktualisierung der Hintergrundrepräsentation lässt sich dabei zwar auch im Rahmen einer groben Exploration realisieren, um jedoch wichtige Detailinformationen über die von den Veränderungen betroffenen Szenenbereiche zu bekommen, bedarf es einer genaueren Analyse.

Darüber hinaus lassen sich auch aus den errechneten Entropiewerten in einer Szene Hinweise auf mögliche Ziele einer Detailexploration gewinnen. Verschiedene Teilbereiche einer Szene sind durch eine, gegenüber den übrigen Bereichen außergewöhnlich hohe Entropie gekennzeichnet, die einen großen Informationsgehalt in der entsprechenden Region vermuten lässt. Der hier realisierte Ansatz für eine (datengetriebene) Detailexploration fokussiert somit auf Strukturveränderungen und maximale Entropiewerte.

Zur Detailexploration einzelner Szenenbereiche wechselt das System in den Zustand „ExpD“ (Abb. 7.4). Die Notwendigkeit zu einem solchen Wechsel wird dabei fortwährend im Verlauf der groben Szenenexploration überprüft. Dazu findet nach der Berechnung der Interessantheitsmaße aller aktuell betrachteten Szenenpunkte eine Klassifikation der Punkte statt. Im Zustand „ExpN“ wird lediglich die Entropie der aktualisierten Punkte mit einem empirisch festgelegten Schwellwert verglichen (s. unten). Punkte, deren Entro-

pie diesen Wert überschreitet, werden zur Detailexploration ausgewählt. Befindet sich das System dagegen im Zustand „ExpA“, so werden fehlklassifizierte, nicht länger bewegte Regionen gesucht, die sofortige Aufmerksamkeit erlangen und deren Schwerpunkt direkt als Zielposition einer Detailexploration ausgewählt wird.

Die Exploration der ausgewählten Punkte selbst folgt derzeit einem festkodierten Schema. Jeder selektierte Punkt in der Szene wird zunächst durch Pan- und Tiltbewegungen der Kamera ins Zentrum des aktuellen Bildes gebracht und dann unter einer schrittweisen Erhöhung des Kamerazooms genauer untersucht. Die Anzahl der bei der Exploration eines zuvor bewegten Objektes durchgeführten Zoomschritte leitet sich dabei aus der Größe des bewegten Bereichs ab. Die Bildweite wird in diesem Fall schrittweise erhöht, bis der bewegte Bereich einen festgelegten Anteil der Bildfläche einnimmt. Demgegenüber sind zur Festlegung der Anzahl an Zoomschritten bei der Exploration einer Region mit hoher Entropie a priori keine Anhaltspunkte gegeben. Zwar lassen sich prinzipiell aus einer detaillierteren Analyse der Bildinhalte, etwa mit Konturextraktionsalgorithmen oder Objekterkennungsverfahren, derartige Hinweise ableiten, eine Einbindung solcher Ansätze übersteigt jedoch den Rahmen der vorliegenden Arbeit. Bei der entropiegeleiteten Detailexploration wird die Bildweite daher derzeit unabhängig von den Bildinhalten schrittweise um 300 Pixel erhöht und anschließend wieder bis zur Ausgangsbildweite reduziert.

## 7.4 Ergebnisse & Diskussion

Der folgende Abschnitt stellt nun Ergebnisse vor, die mit den skizzierten Ansätzen und Heuristiken erzielt werden konnten. Den beiden dazu exemplarisch betrachteten Beispielen[4] liegen dabei Grauwertbilder zu Grunde, da sich die Berechnung des maßgebenden Entropiekriteriums derzeit ausschließlich auf Intensitätswerte stützt. Des Weiteren wurde die Fokuspunkt-Selektion in den Beispielen auf Basis der Aufmerksamkeitsebene realisiert und eine optionale Auswertung der Daten der Merkmalsinstanzen selbst nicht durchgeführt. Diese Vorgehensweise vereinfacht zum einen die Visualisierung der Bilder und Aufmerksamkeitskarten. Zum anderen resultiert daraus direkt eine Obergrenze für die Kamerabewegungen, durch die das Risiko von Fehlern bei der Bildregistrierung vermindert und im Allgemeinen eine robuste Mosaikbildberechnung garantiert werden kann (vgl. S. 125). Die Vorteile einer aktiven Szenenexploration auf Basis von Multi-Mosaikbildern zeigen sich aber auch bereits bei dieser Herangehensweise deutlich.

Die Festsetzung der verschiedenen Parameter des Explorationsprozesses, die gemäß der Gleichungen 7.3 und 7.4 zu wählen sind, hängt stark von dem gewünschten Explorationsverhalten des Systems und damit auch dem spezifischen Anwendungskontext ab. Die in den Beispielen verwendeten Werte können daher nur als grobe Anhaltspunkte dienen. In allen Beispielen wurde der räumliche Kontext höher gewichtet als der zeitliche Rahmen, wobei der räumlichen Gewichtung die Einschlusswinkel zwischen den 3D-Ortsvektoren

[4]Weitere Bildfolgen finden sich unter `http://www.informatik.uni-halle.de/~moeller/phd/`.

der jeweils betrachteten Punkte zu Grunde lagen und der Gewichtungsfaktor $d$ (Gl. 7.3) zumeist Werte zwischen 0,02 und 0,05 aufwies. Darüber hinaus wurden bei der Fokussierung eines Punktes nur die Zeitstempel der Szenenpunkte in seiner unmittelbaren Nähe inkrementiert (d.h. im inneren Drittel des aktuellen Bildes), nicht jedoch im gesamten Sichtbereich. Auf diese Weise lässt sich das System beständig in Richtung der Ränder des betrachteten Bereichs lenken, ohne dafür eine explizite Initialisierung noch nicht gesehener Szenenpunkte mit hypothetisierten Entropiewerten zu erfordern. Ferner erlaubt diese Vorgehensweise auch die Vermeidung eines zyklischen Explorationsverhaltens. Der Wert für $\epsilon$ aus Gleichung 7.4 wurde auf 0,1 gesetzt.

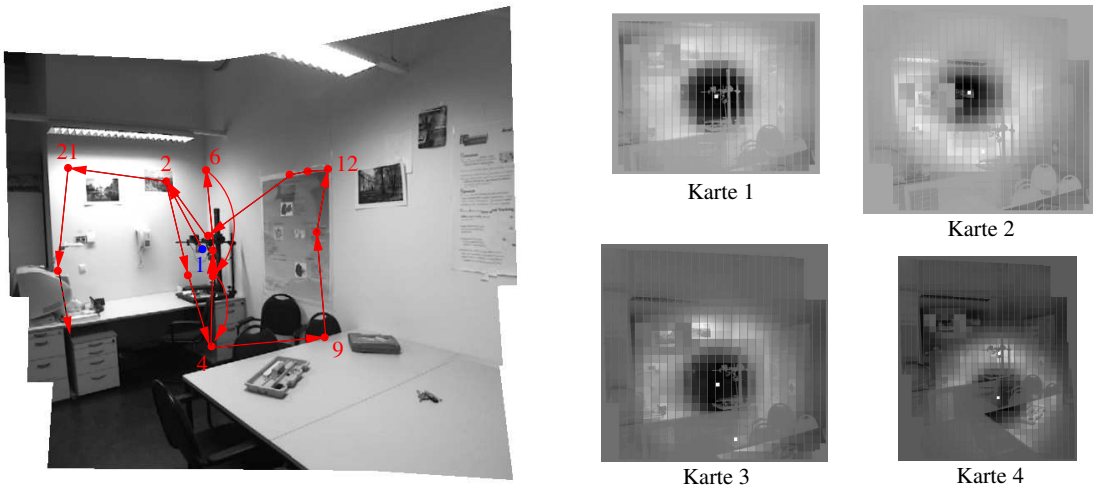



**Abbildung 7.5:** Mosaikbild, berechnet aus einer gemäß den Heuristiken der aktiven Szenenexploration aufgenommenen Bildfolge. Die Trajektorie kennzeichnet grob die Blickbewegungen der Kamera, während rechts einige exemplarische (skalierte und kontrastverstärkte) Aufmerksamkeitskarten gezeigt sind. Zur besseren Visualisierung wurden die Karten und die jeweiligen Bilddaten dabei überlagert.

In Abbildung 7.5 sind zunächst das Mosaikbild eines Raumes, in das die „Blicktrajektorie“ der Kamera im Verlauf der Explorationssitzung eingezeichnet wurde, sowie einige beispielhafte Aufmerksamkeitskarten dieser Sitzung gezeigt. Wie die Trajektorie verdeutlicht, hat das System anfangs den zentralen Bereich des Mosaikbildes fixiert, sich im weiteren Verlauf jedoch auch räumlich entfernteren Szenenpunkten zugewandt (s. auch die Bilder in Abb. A.7). Dies entspricht dem eingangs als Ziel für eine aktive Szenenexploration im Rahmen des visuellen Speichers formulierten Explorationsverhalten. Die Fokuspunkte konzentrieren sich dabei nicht nur auf Objekte, da auch an Kanten und innerhalb von Grauwertverläufen eine große Varianz in den Pixelwerten auftreten kann.

Im zweiten Beispiel, das in Abbilung 7.7 veranschaulicht ist, wird die zu explorierende Szene von einem Tisch dominiert, auf dem verschiedene Gegenstände platziert sind. Das System fokussiert dabei nach und nach verschiedene Objekte auf dem Tisch, wobei auch hier zunächst die initial betrachtete, rechte Tischhälfte im Mittelpunkt steht und die linke Hälfte erst später die Aufmerksamkeit des Systems erlangt (vgl. auch Abb. 7.6).

Bei dem zweiten Beispiel war, im Gegensatz zum ersten, auch die Detailexploration aktiviert. Basierend auf hohen Entropiewerten ($> 0{,}75$ bei normierter Entropie) und

**Abbildung 7.6:** Exemplarische Beispielbilder, die im Rahmen einer aktiven Szenenexploration akquiriert wurden. Die roten Markierungen im dritten und vierten Bild verweisen auf ein Objekt, das der Szene nachträglich hinzugefügt wurde. Das zugehörige Mosaikbild ist in Abb. 7.7 zu sehen.

Bewegungen erfolgte eine Selektion und Exploration einzelner Szenenpunkte im Detail. Einige der dabei ausgewählten Ziele sind in der Abbildung jeweils vergrößert dargestellt. Das Heft im linken oberen Ausschnitt wurde durch die Bewegungsanalyse selektiert. Wie aus der Bildfolge in Abbildung 7.6 zu entnehmen ist, wurde es der Szene erst im Verlauf der Exploration zugefügt, so dass eine strukturelle Veränderung des statischen Hintergrundes die Folge war. Im Rahmen der daraufhin initiierten Detailexploration erfolgten Zoomveränderungen, bis das Heft rund 40% der Bildfläche einnahm.

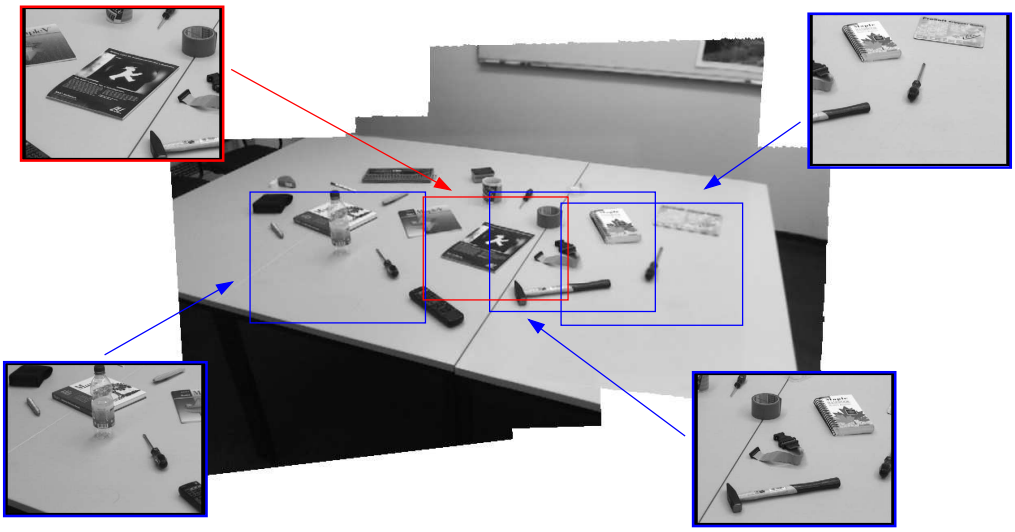


**Abbildung 7.7:** Dieses Mosaik ist das Ergebnis einer aktiven Szenenexploration einschließlich einer Detailexploration verschiedener Teilbereiche. Die mit erhöhter Auflösung explorierten Bereiche sind vergrößert dargestellt. Der Ausschnitt links oben wurde dabei aufgrund von Bewegungsdaten selektiert und beinhaltet ein Objekt, das der Szene erst nach Beginn der Exploration zugefügt wurde (Abb. 7.6).

Die diskutierten Beispiele zeigen einerseits, dass bereits auf Basis rudimentärer Interessantheitsmaße, wie der Entropie und dem räumlichen und zeitlichen Kontext, das gewünschte Explorationsverhalten bei der aktiven Szenenexploration erzielt werden kann. Darüber hinaus wird auch deutlich, dass der visuelle Speicher wichtige Vorteile für eine aktive Szenenexploration mit sich bringt. Dem System stehen jeweils alle bislang aufgenommenen Bilddaten für die Selektion eines Fokuspunktes zur Verfügung, so dass eine sehr viel breitere Datenbasis ausgewertet werden kann. Dies erlaubt schlussendlich die Realisierung eines stärker zielgerichteten Explorationsverhaltens, wie es im Hinblick auf eine effiziente Auswertung visueller Daten für interaktive Systeme unerlässlich ist.

Grundsätzlich wird das angestrebte Verhalten eines Systems bei einer aktiven Exploration vom gegebenen Anwendungskontext beeinflusst. So folgen geeignete Interessantheitsmaße und auch die Bedeutungen des räumlichen und zeitlichen Kontextes zumeist direkt aus dem Szenario. Das vorgestellte Konzept bietet in dieser Hinsicht eine gute Grundlage für eine einfache Adaption des Systems an neue Einsatzfelder und Anforderungen. Durch die modulare Struktur können ohne Aufwand neue Interessantheitsmaße durch zusätzliche Merkmalsinstanzen integriert werden. Darüber hinaus lassen sich aber auch, bei Bedarf sogar dynamisch im Verlauf einer Explorationssitzung, neue Merkmale auf Basis der zusätzlich gespeicherten Bilddaten gewinnen. Werden etwa in interaktiven Systemen vom Benutzer Objekte mit einer spezifischen Farbe referenziert, so lässt sich direkt, ohne einen erneuten Hardwarezugriff, eine Farbkarte aller bereits explorierten Szenenbereiche erstellen. Das System wird somit in die Lage versetzt, flexibel auf variierende Anforderungen zu reagieren, ohne eine Beschränkung der zu Grunde gelegten Merkmale im Vorfeld zu bedingen.

Auch die Detailexploration hat sich als geeignet erwiesen, um vermeintlich besonders interessante Szenenausschnitte zu lokalisieren und in erhöhter Auflösung zu explorieren. Allerdings sind perspektivisch im Hinblick auf eine robuste Detailexploration weitere Verbesserungen der dazu eingesetzten Algorithmen empfehlenswert. Einerseits wird bei einer entropiegeleiteten Detailexploration derzeit nur eine festgewählte Anzahl von Zoomschritten durchgeführt, die die Eigenschaften von Objekten und insbesondere deren Größe nicht berücksichtigt. Bei einer Exploration struktureller Veränderungen im Szenenhintergrund auf Basis von Bewegungsdaten wird die aus den Daten ableitbare Objektgröße zwar in die Festlegung der Zoomstufen einbezogen, auch hier sind jedoch weitere Verifikationen des ausgewählten Bereichs notwendig. Eine Erhöhung des Zooms bedingt oftmals ein ansteigendes Risiko, für die Registrierung wichtige Strukturen im Bild zu verlieren und somit eine erfolgreiche Bildregistrierung zu gefährden. Dieser Aspekt ist von großer Bedeutung, da derzeit nicht zwischen strukturellen Veränderungen im Hintergrund aufgrund des Hinzufügens bzw. Entfernens von Objekten unterschieden wird. Insbesondere im letzten Fall ist in der explorierten Region zumeist nur noch der Szenenhintergrund sichtbar, der im Allgemeinen keine hinreichende Struktur aufweist.

Der hohen Bedeutung dynamischer Komponenten in einer Szene wird derzeit vorrangig durch eine Detailexploration struktureller Veränderungen im Hintergrund Rechnung getragen, um die Konsistenz der Mosaikrepräsentation zu gewährleisten. Die dynamischen Daten bewegter Objekte werden dabei zwar auch in Form von Trajektorien extrahiert und gespeichert (Abschnitt 5.2.1), bleiben aber bei der aktiven Exploration noch unberücksichtigt. Sie können jedoch eine gute Grundlage für eine weitere Erhöhung der Leistungsfähigkeit des Systems bilden, etwa im Rahmen eines gezielten Objekt-Trackings.

Zusammenfassend kann gefolgert werden, dass der realisierte Ansatz einer aktiven Szenenexploration auf Basis der Multi-Mosaikbilder eine gute Grundlage zur Entwicklung derartiger Algorithmen bildet. Durch seine modulare Struktur stellt er eine hohe Flexibilität zur Erschließung verschiedenster Anwendungsfelder bereit, von der die Weiterentwicklung interaktiver Systeme profitieren kann.

# 8 Multi-Mosaikbilder in interaktiven Systemen

Ein wichtiges Ziel bei der Entwicklung interaktiver Systeme ist die fortwährende Erschließung neuer Einsatzgebiete. Dazu ist es unerlässlich, die Systeme so flexibel wie möglich zu gestalten, um ohne großen Aufwand eine schnelle Anpassung an veränderte Aufgabenfelder gewährleisten zu können. Die aktuelle Forschung in diesem Bereich wurde nicht zuletzt auch als Antwort auf die daraus resultierenden Anforderungen unter anderem auf *mobile*, *interaktive* Roboter fokussiert [Kes00, Bre03, Haa04]. Ortsungebundene Systeme, die mit verschiedenen Sensoren und Aktoren, sowie effizienten, internen Verarbeitungspfaden für akquirierte Informationen ausgestattet sind, versprechen die größtmöglichen Spielräume bei der Erschließung neuer Anwendungsszenarien.

In vielen Fällen zielt eine Ausweitung des Einsatzbereichs mobiler, interaktiver Systeme auf eine stärkere Einbindung der Technik in die Alltagswelt des Menschen, wobei jedoch vielfältige Probleme zu lösen sind. Die mobilen Systeme benötigen einerseits Fähigkeiten zur selbständigen Navigation und Lokalisation in unbekannten Umgebungen. Darüber hinaus ist es im Allgemeinen erwünscht, dass Mensch und Maschine in direkten Kontakt miteinander treten und dabei auch natürlich kommunizieren. Eine breite Akzeptanz von künstlichen Systemen bedingt damit zwangsläufig, dass eine intuitive Mensch-Maschine-Interaktion ermöglicht wird, die sich nicht wesentlich von der zwischenmenschlichen Kommunikation unterscheidet. Aus den Vorgehensweisen verschiedener Forschungsprojekte [Dar96, Ric96, Böh03] zur Realisierung einer solchen, oftmals situationsgebundenen Interaktion lässt sich dabei ersehen, dass dieses Ziel ohne eine multimodale Sensorik und damit unmittelbar verknüpfte, ausgereifte Verarbeitungsstrategien für aufgenommene Informationen kaum realistisch erscheint.

Dem Menschen steht zur Kommunikation und Interaktion mit seiner Umwelt ein breites Spektrum verschiedenster Modalitäten zur Verfügung, die sich von der Sprache, über Mimik und Gestik bis hin zu visuellen Informationen über seine Umgebung erstrecken. Insbesondere die visuellen Daten bilden dabei eine unverzichtbare Grundlage für eine effektive und zielgerichtete Interaktion. Sie erlauben beispielsweise eine Erfassung der räumlichen Strukturen der Umgebung sowie die Erkennung von Personen und Objekten, zwei notwendige Grundvoraussetzungen für selbständiges Interagieren in unbekannten Situationen. Damit kommt den visuellen Daten auch bei der Nachbildung menschlicher Kommunikationsfähigkeiten in interaktiven, technischen Systemen eine Schlüsselrolle zu.

Eine detaillierte Analyse derartiger Daten basiert oftmals auf *zeitlich integrierten* Informationen, so dass Mosaikbilder im Allgemeinen und die online berechenbaren Multi-

Mosaikbilder der vorliegenden Arbeit im Besonderen an dieser Stelle einen wichtigen Beitrag zu einer verbesserten, intuitiven Mensch-Maschine-Interaktion leisten können.

Während Mosaikbilder in der Computergrafik und Nachrichtentechnik bereits seit langem bei der Lösung verschiedenster Fragestellungen Anwendung finden, bilden sie im Bereich der interaktiven Systeme und speziell im Forschungsfeld der mobilen Robotik bislang vorrangig die Grundlage zur Lösung von Navigations- und Lokalisationsaufgaben [Ish94, Gau97]. Ihre Leistungsfähigkeit ist damit jedoch nicht ausgeschöpft, da sie auch eine Realisierung erweiterter kognitiver Fähigkeiten, wie z.B. das Lernen und Erkennen von Objekten, effizient unterstützen können. Das vorliegende Kapitel zeigt Perspektiven für eine Verwendung der Multi-Mosaikbilder in einem solchen Kontext auf. Den Ausgangspunkt bildet dabei BIRON (Bielefeld Robot Companion), ein mobiler Roboter, der in der Arbeitsgruppe „Angewandte Informatik“ der Universität Bielefeld als Forschungsplattform für die Entwicklung von Ansätzen im Bereich der multimodalen Mensch-Roboter-Interaktion dient.

Im nachfolgenden Abschnitt erfolgt zunächst eine kurze Einführung in die Thematik der intuitiven Mensch-Roboter-Interaktion am Beispiel des derzeitigen Entwicklungsstandes von BIRON. Im Rahmen einer Erweiterung seiner kognitiven Fähigkeiten hinsichtlich eines robusten Erlernens und Erkennens von Objekten wurde in Zusammenarbeit mit den Kollegen aus Bielefeld ein Konzept für die Einbindung der Multi-Mosaikbilder in die Architektur von BIRON entworfen. Das Unterkapitel 8.2 gibt dazu einen kurzen Überblick über allgemeine Fragestellungen bezüglich des Lernens und Erkennens von Objekten in technischen Systemen, bevor Abschnitt 8.3 die geplante Integration der Mosaikbilder in die Architektur von BIRON beschreibt. Das Kapitel schließt mit der Diskussion vorläufiger Ergebnisse, die im Rahmen von Vorversuchen erzielt wurden und Rückschlüsse auf die zu erwartende Leistungsfähigkeit des Ansatzes bei einer vollständigen Integration der Multi-Mosaikbilder zulassen (Abschnitt 8.4).

## 8.1 Intuitive Mensch-Roboter-Interaktion: BIRON

Den Ausgangspunkt zur Integration der entwickelten Multi-Mosaikbilder in ein interaktives System bildet BIRON (Abb. 8.1), der *Personal Robot* der Arbeitsgruppe „Angewandte Informatik“ der Universität Bielefeld [Haa04]. Die Zielsetzung bei der Entwicklung von Personal Robots besteht darin, mobile Roboter mit Fähigkeiten auszustatten, die es ihnen ermöglichen, gemeinsam mit Menschen in einem Haushalt zu „leben“ und dabei alltägliche Aufgaben selbständig auszuführen. Die Roboter müssen sich dazu einerseits sicher in unbekannten Umgebungen orientieren und mit Menschen natürlich kommunizieren können, andererseits aber auch durch aktives, instruiertes Lernen eine Erweiterung ihrer Aufgabenbereiche und Handlungsspielräume ermöglichen. Eine dynamische Ergänzung ihres Wissens im Hinblick auf neue Objekte und Orte erlaubt es den Robotern, auch in unbekannten Situationen autark und flexibel zu agieren. Dieser zweite Aspekt ist insbesondere im Hinblick auf eine problemlose Integration der Roboter in die menschliche Alltagswelt von hoher Bedeutung, da jeder Nutzer individuelle, im Vorfeld

nicht bekannte Präferenzen hat, und jede neue Umgebung durch spezifische Merkmale charakterisiert wird (z.B. Raumanordnung in einer Wohnung, Positionen von Möbeln), die sich ein Roboter nur durch eine aktive Adaption erschließen kann.

Um aus diesem so genannten *Home Tour Scenario* die für die Praxis resultierenden, notwendigen Basisfähigkeiten eines Personal Robots ableiten zu können, empfiehlt sich ein detaillierterer Blick auf den skizzierten Anwendungskontext. Der Beginn einer zielgerichteten Interaktion mit einem Kommunikationspartner setzt zunächst dessen Detektion und Lokalisation voraus, so dass nachfolgend eine Ebene gemeinsamer Aufmerksamkeit *(Joint Attention)* [Kap04] als Grundlage für einen weiteren Informationsaustausch etabliert werden kann. BIRON verfügt zu diesem Zweck über ein multimodales Aufmerksamkeitssystem [Lan03], mit dessen Hilfe er potenzielle Kommunikationspartner erkennen und lokalisieren kann.


**Abbildung 8.1:** BIRON.

Den Kern des Systems bildet eine Auswertung von akustischen und visuellen Sensorinformationen sowie den Entfernungsdaten eines Laser-Distanzmessgerätes. Komponenten zur Spracherkennung und zum Sprachverstehen stellen den Grundstein für eine natürlich sprachliche Interaktion mit dem Roboter bereit. Dabei wird durch die Verwendung von zwei Mikrophonen gleichzeitig eine Lokalisation der Quellen akustischer Signale möglich, die Rückschlüsse auf die Standorte potenzieller Kommunikationspartner erlaubt.

Die Analyse visueller Daten umfasst vorrangig eine merkmalsbasierte Gesichtserkennung und -identifikation [Jon01], während sich mit Hilfe der Entfernungsdaten des oberhalb der Räder angebrachten Lasers menschliche Beinpaare in der Nähe des Roboters lokalisieren lassen. Jede dieser drei Modalitäten liefert Hinweise auf potenzielle Kommunikationspartner, die Daten einzelner Quellen weisen jedoch nicht immer die zur Initiierung einer direkten Kommunikation notwendige Genauigkeit auf. Die finale Fokussierung auf einzelne Personen beruht daher auf einer Fusion der Analyseresultate aller drei Datenquellen *(Multi-Modal Anchoring)* [Fri03].

Die Detektion eines Kommunikationspartners und die nachfolgende Etablierung einer Ebene gemeinsamer Aufmerksamkeit stellt in der Regel den Auftakt zur Durchführung komplexerer Interaktionshandlungen dar. Im Hinblick auf Personal Robots folgt dieser initialen Kontaktaufnahme daher oftmals ein wechselseitiger Informationsaustausch, der den Roboter befähigen soll, konkreten Instruktionen zu folgen. In dem skizzierten Home Tour Scenario beziehen sich die Instruktionen dabei zumeist auf Interaktionen mit Gegenständen der Umgebung. Daraus folgt, dass eine weitere, essenzielle Grundvoraussetzung für sinntragendes Handeln von Personal Robots durch die Fähigkeit gegeben ist, Objekte lernen und später auch wiedererkennen zu können. Diesen Anforderungen trägt ein derzeit in der Entwicklung befindliches, so genanntes Objektaufmerksamkeits-System

Rechnung, das BIRONs Architektur um die dafür notwendigen, kognitiv motivierten Strukturen erweitern soll. Neben Verfahren zur Erkennung von Objekten umfasst das System insbesondere Algorithmen zur gezielten Lokalisierung und Fokussierung einzelner Objekte in einer Szene, auf deren Basis eine effiziente, objekt-bezogene Mensch-Roboter-Interaktion realisiert werden kann.

Innerhalb der Systemarchitektur von BIRON bildet das Objektaufmerksamkeits-System die Schnittstelle zwischen den Instruktionen des Nutzers auf der einen und eines in Form einer Wissensbasis verwalteten Szenenmodells auf der anderen Seite (Abb. 8.2). Es wird aktiviert, wenn sich – nach Fokussierung auf einen Kommunikationspartner – in dessen sprachlichen Äußerungen Hinweise für die Referenzierung eines spezifischen Objektes finden. Durch eine kombinierte Analyse der aktuellen Bilddaten, der sprachlichen Verweise des Menschen [Top04] und der Resultate einer zusätzlich initiierten Gestenerkennung [Hof04] wird das referenzierte Objekt dann (interaktiv) lokalisiert und identifiziert bzw. bei Bedarf neu gelernt und dem Szenenmodell zugefügt.

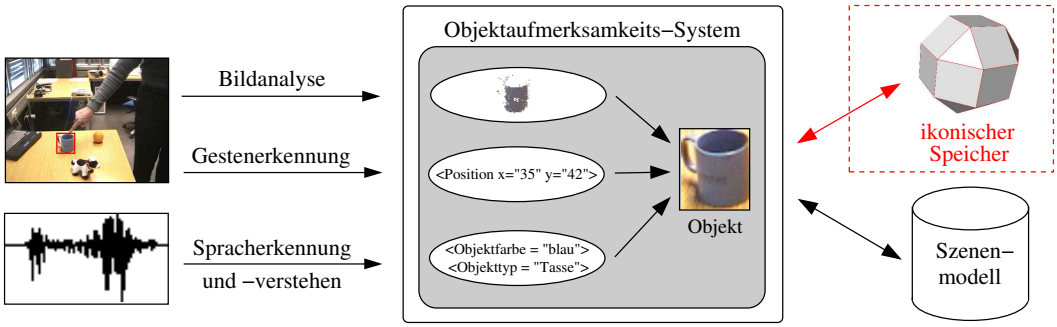



**Abbildung 8.2:** Realisierung von Objektaufmerksamkeit für BIRON: Ein Objektaufmerksamkeits-System dient als Schnittstelle zwischen den multimodalen Eingaben (Sprache, Gestik) eines Kommunikationspartners und dem Szenenmodell des Roboters. Die Multi-Mosaikbilder (oben rechts) stellen zusätzliche Daten für eine flexiblere, ansichts-basierte Repräsentation der Objekte bereit.

## 8.2 Erkennen und Lernen von Objekten

Die Nachbildung der menschlichen Leistungsfähigkeit bei der Erkennung von Gegenständen und Personen stellt ein im Rahmen der automatischen Analyse und Klassifikation von Bilddaten bereits seit vielen Jahren bearbeitetes Problem dar. Eine allgemeingültige Lösung konnte dennoch bislang nur in Ansätzen gefunden werden. Eine der Hauptschwierigkeiten, die es zu bewältigen gilt, ist die Entwicklung geeigneter Repräsentationsstrukturen für erworbenes Objektwissen, auf deren Basis sich eine robuste Erkennung realisieren lässt. Unter anderem müssen beispielsweise geeignete Merkmale gefunden werden, die ein Objekt einerseits exakt genug charakterisieren, um es leicht von anderen Objekten unterscheiden zu können, gleichzeitig aber keinen zu großen Berechnungsaufwand bei der Objekterkennung selbst verursachen.

Im Laufe der Jahre wurden verschiedenste Ansätze zur Beschreibung von Objekteigenschaften untersucht (eine einführende Übersicht findet sich z.B. in [Sue92]). Neben weit verbreiteten Merkmalen auf Basis von Textur, Farbe oder Struktur, die ein Objekt

zumeist als Einheit auffassen, können zur Charakterisierung auch formbeschreibende, geometrische 2D- oder 3D-Primitive herangezogen werden, etwa parallele Kanten oder geschlossene Konturzüge [Mil94]. Derartige Herangehensweisen sind eng mit Ansätzen verknüpft, die auf einer Zerlegung von Objekten in konstituierende Teile gründen. Die Teile können dabei unabhängig voneinander im Bild detektiert und anschließend wieder zu vollständigen Objekten zusammengesetzt werden (vgl. auch die Arbeit von Biederman zur Objekterkennung aus Komponenten [Bie85]).

Bei der Wahl einer konkreten Repräsentation für erworbenes Objektwissen spielt neben den Merkmalen selbst die Robustheit des Ansatzes im Hinblick auf Varianzen innerhalb der Objekteigenschaften eine entscheidende Rolle. Bedingt durch Umwelteinflüsse oder auch spezifische Eigenarten der Objekte kann ihr Erscheinungsbild variieren und eine robuste Erkennung damit nur unter expliziter Berücksichtigung solcher Effekte gelingen. Aufgrunddessen haben zur Festlegung der charakterisierenden Eigenschaften von Objekten insbesondere Ansätze eine hohe Bedeutung erlangt, in denen die Merkmale und Eigenschaften von Einzelobjekten oder auch vollständigen Objektklassen aus Trainingsmengen mit Beispielobjekten „gelernt“ werden (z.B. [Tur91, Web00]). Da die Objekte jedoch häufig dreidimensional sind, hängen die gelernten Merkmale dabei von der konkreten Blickrichtung auf ein Objekt ab und unterlaufen damit die zumeist gewünschte und im Hinblick auf eine robuste Erkennung auch notwendige Generalisierbarkeit der gelernten Repräsentation.

Dieses Problem hat unter anderem die Entwicklung *ansichts-basierter* Objekterkennungsansätze vorangetrieben, in denen verschiedene, charakteristische Beispielansichten eines Objektes explizit in die Darstellung einbezogen werden. Die Robustheit und Flexibilität dieser Verfahren, die inzwischen weit verbreitet sind, nimmt dabei im Allgemeinen mit einer steigenden Anzahl verfügbarer Ansichten zu. Allerdings erfordert die Akquisition verschiedener Ansichten auch ausgereifte Planungsstrategien zur Festlegung geeigneter Positionen für deren Aufnahme [Bor98, Pal00, Den02]. Im Kontext von Personal Robots bedingt eine solche Herangehensweise, dass der Roboter ein unbekanntes Objekt beim Lernen von verschiedenen Seiten betrachten und damit gegebenenfalls auch um das Objekt herum manövrieren muss.

Ein möglicher Ansatz, um den damit verbundenen, oftmals unverhältnismäßig hohen Zeitaufwand zu reduzieren und den Einsatz einer ansichts-basierten Objekterkennung in mobilen Systemen zu vereinfachen, besteht in der Einbindung eines ikonischen Speichers, in dem Bilddaten einer Szene abgelegt werden können. Sind anschließend zum Lernen eines neuen Objektes zusätzliche Ansichten erforderlich, kann der Roboter diese direkt aus dem Speicher abfragen, ohne weitere Positionen innerhalb der aktuellen Szene wiederholt anfahren zu müssen. In den nachfolgenden Abschnitten wird ein solches Konzept exemplarisch anhand des Zusammenspiels des im Rahmen dieser Arbeit entwickelten ikonischen Speichers mit den Objektaufmerksamkeits-Komponenten von BIRON veranschaulicht. Die Multi-Mosaikbilder stellen dabei eine zusätzliche Informationsquelle für das bereits skizzierte Szenenmodell bereit, aus der ikonische Daten zur ansichts-basierten Repräsentation neuer Objekte entnommen werden können (vgl. auch Abb. 8.2).

## 8.3 Szenenrepräsentation durch Multi-Mosaikbilder

Ein Personal Robot dient unter anderem der selbständigen Ausführung von Aufgaben im Haushalt, z.B. dem Transport oder der Manipulation von Gegenständen. Dabei agiert er gemäß den Instruktionen eines menschlichen Kommunikationspartners, der die Aufmerksamkeit des Roboters beispielsweise durch Gestik und Sprache auf die relevanten Bereiche oder Objekte in einer Szene lenkt. Die konkreten Zielobjekte für auszuführende Handlungen sind dabei zumeist im Vorfeld unbekannt, so dass der Roboter sich die zur Lösung einer gestellten Aufgabe notwendigen Informationen erst während einer Interaktion erarbeiten kann. Im Vordergrund steht dabei die Identifikation der zu manipulierenden Objekte in der Szene.

BIRON nutzt zu diesem Zweck ein ansichts-basiertes Objekterkennungsverfahren, dessen Robustheit mit der Anzahl gegebener Objektansichten skaliert (vgl. Abschnitt 8.2). Dabei ist eine ansichts-basierte Darstellung *aller* später potenziell referenzierbaren Objekte in einer Szene mit einem hohen Rechen- und Speicheraufwand verbunden und damit nicht realistisch. Darüber hinaus bedingt eine solche Vorgehensweise auch, dass eine Vielzahl niemals benötigter Repräsentationen erzeugt wird. Ein vielversprechenderer Ansatz besteht daher darin, nach und nach eine vollständige ikonische Repräsentation der Szene aufzubauen, aus der später gezielt die jeweils zum Erkennen bzw. Lernen eines spezifischen Objektes relevanten Daten extrahiert werden können.

Eine mögliche Grundlage für eine solche Szenenrepräsentation bildet eine Menge von Multi-Mosaikbildern, die an verschiedenen Positionen in einer Szene aufgenommen werden (Abb. 8.3). Sie unterstützen die Erkennung von Objekten einerseits durch ihre stückweise Planarität, die eine direkte Anwendung gängiger Bildverarbeitungsalgorithmen und insbesondere auch Objekterkennungsverfahren erlaubt. Darüber hinaus ist die in die Multi-Mosaikbilder integrierte Auflösungshierarchie in diesem Anwendungskontext von hoher Bedeutung, da sie der räumlichen Variabilität der visuellen Informationen in einer Szene Rechnung trägt (vgl. Unterkap. 6.2). Während zur Repräsentation der groben räumlichen Strukturen Bilddaten in niedriger Auflösung ausreichend sind, erfordert eine Identifikation von Objekten im Allgemeinen höher aufgelöste Daten. Auch wenn der Roboter die später referenzierten Objekte zum Zeitpunkt der Szenenexploration noch nicht kennt und folglich auch nicht näher auf charakterisierende Eigenschaften untersuchen kann, lässt sich dennoch mit Hilfe geeigneter Explorationsstrategien (s. auch Kap. 7) eine lokal im Detail variierende, ikonische Repräsentation aufbauen, die später zu einer signifikanten Vereinfachung der Lern- und Erkennungsprozesse beitragen kann.

Ein mit der skizzierten Szenenmodellierung vergleichbarer Ansatz findet sich in [Tel98], wo eine Menge sphärischer Mosaikbilder die Grundlage für das Rendering realistischer 3D-Modelle eines Stadtteils bildet. Im Gegensatz zu dieser Arbeit werden dort allerdings weder eine Online-Berechnung der Bilder noch ein direkter Datenzugriff angestrebt. Vielmehr besteht das Ziel in einer möglichst exakten Rekonstruktion der Umgebung, so dass auch Rechenzeiten in der Größenordnung mehrerer Stunden akzeptabel sind.

Der nachfolgende Abschnitt skizziert zunächst die Aufnahme der Mosaikbilder zur Szenenrepräsentation, bevor eine Beschreibung der Datenabfrage folgt (Abschnitt 8.3.2).

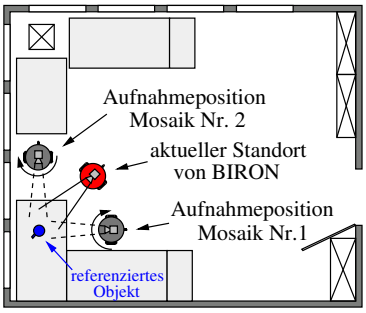



**Abbildung 8.3:** Szenenrepräsentation auf Basis von Multi-Mosaikbildern.

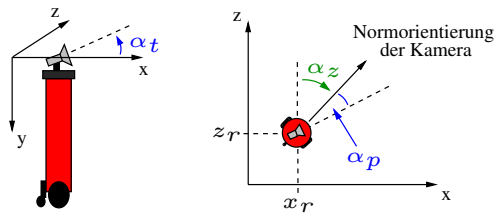



**Abbildung 8.4:** Definition des 3D-Weltkoordinatensystems in einer Szene.

### 8.3.1 Aufnahme der Mosaikbilder

Während BIRON potenzielle Kommunikationspartner erwartet, d.h. bevor eine spezifische Kommunikationssituation eintritt, exploriert er aktiv seine aktuelle Umgebung. Er nimmt dabei unter Umständen bereits visuelle Daten auf, die für die Lösung späterer Aufgaben von Interesse sein können. Die akquirierten Informationen werden aus diesem Grund in Multi-Mosaikbilder integriert und dort für Datenabfragen bereitgehalten. Dabei kann die Datenintegration durch die realisierte Online-Berechnung der Multi-Mosaikbilder (Unterkap. 6.4) jederzeit, beispielsweise aufgrund von eintretenden Nutzerinteraktionen, ohne Datenverlust unterbrochen werden.

Das Konzept der Multi-Mosaikbilder bedingt, dass der Roboter zum Zeitpunkt der Bildakquisition eine feste Position in der Szene einnehmen muss und diese während der Aufnahme der Mosaikbilder nicht verlassen darf. Da er jedoch unter anderem mit einer aktiv steuerbaren Pan-Tilt-Kamera ausgestattet ist (vgl. hierzu Abb. 8.1 und Abb. 2.5), kann eine großwinklige Szenenexploration auch von einem einzelnen Standort aus erfolgen. Eine Ausdehnung der Berechnung von Mosaikbildern auf Phasen, in denen der Roboter fährt, ist zwar prinzipiell ebenfalls möglich, bringt jedoch eine Reihe von Schwierigkeiten mit sich. Einerseits lassen sich beliebige Kamerabewegungen in 3D-Szenen nicht mehr in geschlossener Form modellieren, und die für derartige Konstellationen entwickelten, so genannten „Manifold Mosaics“ [Pel97, Pel00] weisen (ohne eine Einschränkung der zulässigen Kamerabewegungen) oftmals Verzerrungen auf [Fel03], die einer direkten Analyse mit konventionellen Techniken entgegenstehen.

Unabhängig von den vorstehenden Aspekten ist es darüber hinaus aber auch zum Aufbau einer ikonischen Szenenrepräsentation zumeist nicht notwendig, *alle* im Verlauf einer Szenenexploration durch die rotatorischen und translatorischen Bewegungen der Kamera und des Roboters akquirierten, visuellen Daten zu speichern. Analog zur schritthaltenden Modellierung des statischen Hintergrundes einer Szene anhand ausgewählter Momentaufnahmen, die Zeitpunkte ohne unabhängige Bewegungen in der Szene widerspiegeln (vgl. Abschnitt 4.1.2 bzw. auch Kap. 5), lässt sich eine Szene vielmehr in der Regel bereits durch Aufnahmen von einigen wenigen Multi-Mosaikbildern an ausgewählten „Schlüsselpositionen“ hinreichend exakt modellieren.

Eine Extraktion spezifischer Ansichten einzelner Szenenbereiche bzw. Objekte aus den verschiedenen Mosaikbildern erfordert die Kenntnis der räumlichen Zusammenhänge zwischen den relevanten Zielregionen, dem aktuellen Standort des Roboters und der Aufnahmepositionen der einzelnen Multi-Mosaikbilder in der Szene (vgl. Abschnitt 8.3.2). Für jede Szene (z.B. einzelne Räume oder auch eine ganze Wohnung) wird daher ein globales 3D-Weltkoordinatensystem definiert, das die Grundlage zur Spezifikation der benötigten räumlichen Strukturen bildet (Abb. 8.4). Das Koordinatensystem ist immer rechtshändig orientiert, wobei die y-Achse jeweils nach unten zeigt. Die Lage des Koordinatenursprungs in der Szene kann grundsätzlich beliebig gewählt werden. Da die Höhe der Kamera auf dem Roboter aktuell jedoch nicht variabel ist, bietet es sich zumindest an, die xz-Ebene des Koordinatensystems auf die Höhe der Kamera zu legen, so dass die y-Koordinate von Positionsangaben im Allgemeinen den Wert 0 aufweist.

Die Orientierung des Roboters an einer spezifischen Position $(x_r, y_r, z_r)$ innerhalb dieses Koordinatensystems wird durch den Winkel $\alpha_z$ angegeben, den die Robotervorderseite relativ zur z-Achse aufweist. Die Kamera zeigt dabei in ihrer als Referenz dienenden Grundausrichtung geradeaus, d.h. die Pan- und Tiltwinkel werden jeweils zu Null angenommen (Normorientierung). Mit jedem Multi-Mosaikbild der Repräsentation lassen sich auf dieser Basis 3D-Koordinaten innerhalb der Szene assoziieren, wobei im Wesentlichen die Odometriedaten des Roboters als Anhaltspunkte zur Bestimmung der Position dienen. Die spezifische Orientierung eines Multi-Mosaikbildes, die durch die Ausrichtung der Basisebene definiert wird (vgl. Abschnitt 6.1.3), ergibt sich aus der Grundorientierung $\alpha_z$ des Roboters und der Orientierung der Kamera in der Horizontalen $\alpha_p$ und Vertikalen $\alpha_t$ bei der Bildaufnahme relativ zu der zuvor definierten Normorientierung.

### 8.3.2 Extraktion von Objektansichten

Das Ausführen bzw. die korrekte Interpretation einer Instruktion (z.B. „Das ist meine blaue Tasse.“) ist für den Roboter vorrangig mit der Identifikation des referenzierten Objektes verbunden. Dazu werden zunächst aus den sprachlichen Äußerungen des Instrukteurs sowie den Resultaten der automatischen Handgestendetektion Zielpositionen für das Objekt innerhalb der Szene bestimmt (vgl. Abb. 8.2). Gleichzeitig erfolgt ein Abgleich der Objektdaten (z.B. von Bezeichnungen wie „meine Tasse“ oder Eigenschaften wie „blau“) mit den Objektrepräsentationen, die bereits im Szenenmodell gespeichert sind, um das Objekt identifizieren bzw. gegebenenfalls als noch unbekannt einstufen zu können. Eine Objektrepräsentation umfasst dabei neben semantischen Beschreibungen, die sich etwa aus sprachlichen Äußerungen ableiten lassen (z.B. „meine Tasse“), insbesondere eine Menge verschiedener Ansichten des Objektes zur adäquaten Charakterisierung seiner visuellen Eigenschaften. Ist das Objekt bereits bekannt, wird die aktuelle Ansicht der vorhandenen Repräsentation zur Vervollständigung zugefügt, während ein unbekanntes Objekt die Instanziierung einer neuen Objektrepräsentation erfordert. Die dazu notwendigen Ansichten können in einem solchen Fall ohne nennenswerten Aufwand aus den zuvor aufgenommenen Multi-Mosaikbildern gewonnen werden.

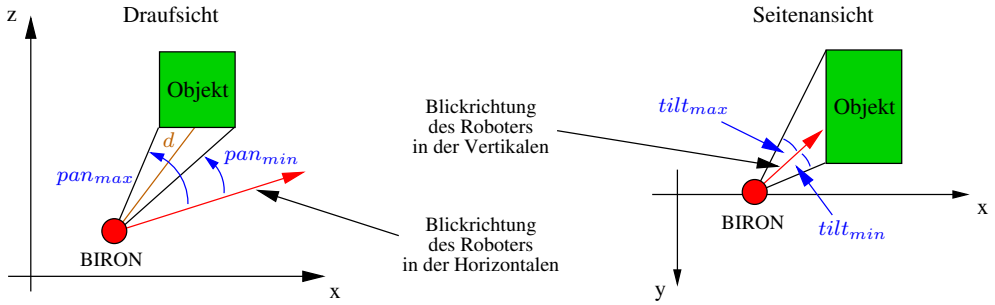



**Abbildung 8.5:** Spezifikation der Position des referenzierten Objektes: Das Objektvolumen ergibt sich durch 3D-Triangulationen aus dem Abstand $d$ der Frontfläche des Objektes zum Roboter und den Blickwinkeln $pan_{min}$ und $pan_{max}$ bzw. $tilt_{min}$ und $tilt_{max}$ des Roboters auf diese Fläche.

Den Ausgangspunkt zur Extraktion dieser Ansichten bilden die 3D-Weltkoordinaten des Roboters sowie seine aktuelle Blickrichtung, die Position des referenzierten Objektes relativ zum Roboter und die Positionen der Multi-Mosaikbilder in der Szene. Das Objekt selbst wird derzeit durch einen Quader repräsentiert, wobei die Annahme zu Grunde liegt, dass eine derartige, leicht zu handhabende, jedoch nur approximative Beschreibung zur Definition des relevanten Bildbereichs in den Mosaikbildern ausreichend ist. Die Objektposition und -größe wird durch den geschätzten Abstand zum Roboter und die begrenzenden Winkel des Sichtkegels vom Roboter aus auf das Objekt spezifiziert (Abb. 8.5). Die Entfernung lässt sich dabei etwa mit Hilfe der Stereokamera abschätzen, die unterhalb des oberen Sonarrings des Roboters montiert ist (Abb. 8.1), oder aber über ein zusätzliches Laser-Distanzmessgerät, das direkt auf der Kamera angebracht werden kann [Haa03].

Aus der Lage des Objektes relativ zum Roboter lassen sich zunächst die vier Eckpunkte der Frontfläche des Objektquaders über 3D-Triangulationen berechnen. Die Fläche bildet den Ausgangspunkt zur Extrapolation der Koordinaten der übrigen vier Eckpunkte des Quaders, wobei dessen Ober- und Unterseiten zur Vereinfachung als quadratisch und parallel zur xz-Ebene ausgerichtet angenommen werden. Unter diesen Voraussetzungen lassen sich die restlichen Punkte leicht geometrisch bestimmen. Ist die exakte Position des Objektvolumens schließlich vollständig bekannt, kann in einem nächsten Schritt die Extraktion der visuellen Daten aus den einzelnen Multi-Mosaikbildern erfolgen.

Dazu wird für jedes Mosaikbild bzw. für alle zuvor berechneten Projektionsinstanzen (Unterkap. 6.2) zunächst jeweils die Teilebene bestimmt, deren Normalenvektor die geringste Winkeldifferenz zum Richtungsvektor zwischen dem Zentrum des Mosaikbildes und dem Objektmittelpunkt aufweist. Sie beinhaltet die vermeintlich beste Ansicht des referenzierten Objektes. Durch eine Schnittpunktbildung der Richtungsvektoren der acht Eckpunkte des Objektvolumens mit der ausgewählten Ebene lässt sich das Objekt auf diese Ebene projizieren. Nach einer Umrechnung der resultierenden Projektionspunkte in lokale 2D-Bildkoordinaten kann dann das umschließende Rechteck des Bereichs im Mosaikbild bestimmt werden, dessen visuelle Daten dem referenzierten Szenenausschnitt entsprechen (Abb. 8.6).

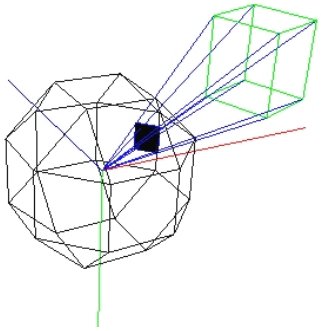


**Abbildung 8.6:** Skizze der Rückprojektion eines referenzierten Objektes (grüner Quader) auf die Teilflächen eines Multi-Mosaikbildes: Die Ortsvektoren der acht Eckpunkte des Objektvolumens sind blau eingezeichnet, während die Projektion durch das schwarze Rechteck angedeutet wird.

Bei der Extraktion der visuellen Daten ist zu berücksichtigen, dass das berechnete Rechteck für sehr große Objekte in Abhängigkeit von der gewählten Skalierung der Teilflächen (vgl. Abschnitt 6.1.3) nicht zwingend nur eine einzelne Teilfläche des Multi-Mosaikbildes überdeckt, sondern sich auch über mehrere benachbarte Flächen erstrecken kann. Zur Generierung eines Bildes des gewünschten Szenenbereichs müssen in diesen Fällen Informationen von verschiedenen Teilflächen verschmolzen werden. Aktuell geschieht dies durch eine Projektion der Daten von Nachbarebenen auf die ausgewählte Teilfläche, wobei zur Projektion der Bilddaten dasselbe Rekursionsschema Anwendung findet, das bereits im Zusammenhang mit einer Aktualisierung der Fokus-Bildebene skizziert wurde (Abschnitt 6.4.2).

## 8.4 Ergebnisse & Schlussfolgerungen

Das Konzept der Objektaufmerksamkeit für BIRON befindet sich derzeit noch in der Entwicklung, so dass die in den vorhergehenden Abschnitten skizzierte Einbettung der vorgestellten, ikonischen Szenenrepräsentation in die Architektur von BIRON noch nicht vollständig realisiert werden konnte. Zur Abschätzung der Praxistauglichkeit und der zu erwartenden Leistungsfähigkeit des Ansatzes wurde daher alternativ ein prototypischer Anwendungsfall für das Konzept entworfen und evaluiert. Die daraus resultierenden Ergebnisse und Schlussfolgerungen werden in diesem Abschnitt vorgestellt. Das angenommene Szenario beinhaltet dabei das Lernen eines Objektes, auf das der Roboter durch sprachliche Verweise und Zeigegesten des Benutzers aufmerksam gemacht wurde. Darüber hinaus liegt dem Fallbeispiel die Annahme zu Grunde, dass der Roboter vor Beginn der Interaktion bereits zwei Multi-Mosaikbilder an verschiedenen Positionen innerhalb der Szene berechnet hat. Die räumlichen Zusammenhänge zwischen den Aufnahmepositionen der Mosaikbilder, dem aktuellen Standort des Roboters und der Objektposition sind grob in Abbildung 8.3 dargestellt. Das referenzierte Objekt, in diesem Fall eine blaue Tasse, steht unmittelbar vor BIRON auf dem Tisch, wobei Anfragen an das Szenenmodell ergeben haben, dass die Tasse noch unbekannt ist und somit neue Objektansichten gelernt werden müssen.

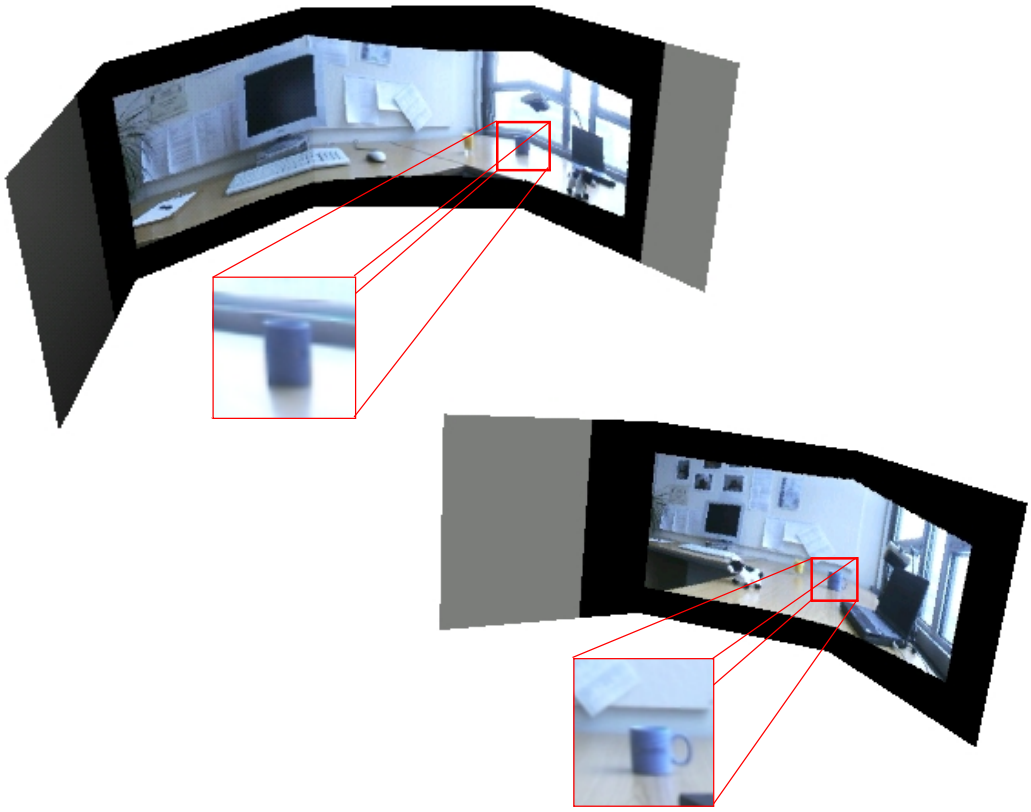


**Abbildung 8.7:** Exemplarische Darstellung einer ikonischen Szenenrepräsentation, die aus zwei Multi-Mosaikbildern besteht (die ungefähre räumliche Lage der Aufnahmepositionen und des referenzierten Objektes kann der Skizze in Abb. 8.3 entnommen werden). Die vergrößerten Bildausschnitte zeigen jeweils die Ansichten des Objektes in den beiden Mosaikbildern, die zur ansichts-basierten Repräsentation aus dem visuellen Speicher extrahiert werden.

In dieser Situation lassen sich nun aus den beiden berechneten Multi-Mosaikbildern, die in Abbildung 8.7 gezeigt sind, Ansichten des referenzierten Objektes extrahieren. Im Gegensatz zu einer expliziten, mit verschiedenen Fahrmanövern verbundenen Neuaufnahme sind diese Bilddaten direkt und ohne Zeitversatz verfügbar. Damit wird insbesondere die laufende Interaktion mit dem Benutzer kaum beeinträchtigt, so dass keine unerwünschten Verzögerungen beim Informationsaustausch auftreten. Die resultierenden Bildausschnitte sind in Abbildung 8.7 vergrößert dargestellt. Es ist deutlich zu erkennen, dass die blaue Tasse, obwohl sie auf den ersten Blick homogen texturiert und gleichmäßig geformt erscheint, zwei deutlich voneinander verschiedene Ansichten aufweist. Insbesondere der Henkel, der ein wesentliches Merkmal zur Identifikation einer Tasse darstellt, ist nur in der zweiten Ansicht zu sehen. Hierdurch bestätigt sich, dass eine Beschränkung der Repräsentation auf einzelne Ansichten oftmals nur unvollständige und schlecht generalisierende Objektbeschreibungen erzeugen kann. Durch die ikonische Darstellung einer Szene in Multi-Mosaikbildern können BIRON dagegen direkt mehrere Ansichten eines Objektes zur Verfügung gestellt werden, so dass flexiblere Objektmodelle resultieren.

Die Anzahl gegebener Ansichten hängt im Allgemeinen von der Anzahl zuvor aufgenommener Mosaikbilder und dem Umfang dabei explorierter Auflösungsstufen ab. Je

länger BIRON vor Beginn einer Interaktion wartet, desto mehr Mosaikbilder können aufgenommen und gespeichert werden. Allerdings darf der zur Verfügung stehende Speicher nicht unberücksichtigt bleiben, der das Datenvolumen der Szenenrepräsentation beschränkt. In der Regel gilt es daher, einen geeigneten Kompromiss zwischen der notwendigen Detailgenauigkeit in der Modellierung einerseits, und den aus den technischen Rahmenbedingungen resultierenden Einschränkungen andererseits zu finden.

Die extrahierten Ansichten bilden die Grundlage zur Instanziierung einer neuen Objektrepräsentation innerhalb des Szenenmodells. Durch eine detailliertere Analyse der Ausschnitte ist es dabei unter anderem möglich, die Objektbeschreibung noch zu verfeinern. Insbesondere sprachliche Hinweise des Benutzers können wertvolle Anhaltspunkte für eine solche Präzisierung liefern. Bezogen auf den hier diskutierten Anwendungsfall erlaubt z.B. eine durch das Wort „blau“ angestoßene Detektion blauer Regionen in den Ausschnitten eine genauere Bestimmung der Objektposition und vorrangig auch der Objektform, die initial durch ein Quader approximiert worden war.

Zusammenfassend zeigen bereits diese Ergebnisse eines prototypischen Anwendungsfalles, dass die Integration eines ikonischen Speichers in die Systemarchitektur von BIRON zu deutlichen Fortschritten im Hinblick auf eine natürliche Interaktion mit Menschen führen kann. Insbesondere die enge Verzahnung der Szenenrepräsentation mit den geplanten Modulen zur Realisierung der Objektaufmerksamkeit verspricht eine hohe Flexibilität beim Lernen und Wiedererkennen von Objekten. Durch die direkte Verfügbarkeit verschiedener Ansichten eines Objektes eröffnen sich Perspektiven für eine flexible Darstellung, die ohne einen visuellen Speicher nur unter erhöhtem Zeitaufwand und mit komplexen Fahrmanövern des Roboters erzielt werden können.

Darüber hinaus kann der visuelle Speicher perspektivisch auch zu einer direkten Verbesserung der Kommunikation zwischen Mensch und Roboter beitragen. Ist es dem Menschen beispielsweise nicht möglich, ein Objekt unmittelbar durch Zeigegesten zu referenzieren, sondern stattdessen nur sprachlich zu beschreiben, eröffnen die in den Multi-Mosaikbildern gespeicherten Daten dem Roboter die Möglichkeit, dort nach geeigneten Kandidaten für das referenzierte Objekt zu suchen und sie dem Menschen vorzuschlagen. Auf diese Weise wird ein effizienterer Informationsaustausch zwischen dem Menschen und der Maschine möglich, zu dem insbesondere die verbesserte Auswertung gespeicherter visueller Daten auf Seiten des Roboters einen entscheidenden Beitrag leistet.

Eine robuste Erkennung von Objekten ist eine unverzichtbare Grundlage einer effizienten und für den Menschen vor allem auch intuitiven Kommunikation mit interaktiven, mobilen Systemen. In dieser Hinsicht untermauern die vorgestellten Ergebnisse auch die eingangs dieser Arbeit aufgezeigte, hohe Bedeutung einer engen Verknüpfung von Interaktionsstrategien mit internen Repräsentationsdatenstrukturen. Sie ist beim Menschen ein unverzichtbarer Bestandteil des kognitiven Systems und nimmt damit auch bei der Entwicklung technischer Systeme eine Schlüsselrolle ein. Der visuelle Speicher aus Multi-Mosaikbildern erweitert in dieser Hinsicht die internen Strukturen zur Verarbeitung aufgenommener Daten von BIRON und bildet damit einen wichtigen Baustein in der Realisierung kognitiver Fähigkeiten in künstlichen, interaktiven Systemen.

# 9 Zusammenfassung & Ausblick

Interaktive Systeme dringen in zunehmendem Maße in die menschliche Alltagswelt vor. Im Gegensatz zu Laborumgebungen, in denen fest definierte Bedingungen vorherrschen, ist dieses neue Anwendungsfeld dabei mit einer sehr viel größeren Komplexität verbunden, die eine hohe Flexibilität zu einer notwendigen Grundvoraussetzung für die Einsetzbarkeit interaktiver Systeme werden lässt. Eine solche Flexibilität bedingt unter anderem Mechanismen für einen effizienten Informationsaustausch des Systems mit seiner Umgebung. Visuelle Daten, insbesondere Bildfolgen, die mit Hilfe aktiver Sensoren zielgerichtet und aufgabenorientiert akquiriert werden können, nehmen dabei eine Schlüsselrolle ein und bilden eine unverzichtbare Grundlage für ein sinntragendes Verhalten interaktiver Systeme. Allerdings bietet eine aktive Sensorik allein noch keine hinreichende Basis für den Einsatz interaktiver Systeme in der Alltagswelt. Auch geeignete Algorithmen für deren Ansteuerung und interne Datenstrukturen zur effizienten Handhabung und weiteren Analyse der aufgenommenen Informationen sind unerlässlich. Erst durch sie wird es möglich, die Daten gezielt zu gliedern und damit als Grundlage für Entscheidungen und angemessene Verhaltensweisen zu nutzen.

Vor diesem Hintergrund wurde in der vorliegenden Dissertation ein neues Konzept für ein visuelles „Gedächtnis“ vorgestellt, das eine effiziente Speicherung und Verarbeitung von Bildfolgen ermöglicht, die mit aktiven Kameras aufgenommen wurden. Die Basis eines solchen, visuellen Speichers bilden Mosaikbilder, deren Grundidee in einer redundanzfreien, speichereffizienten ikonischen Darstellung von Bildfolgen besteht. Die Daten werden dabei weitestgehend unvorverarbeitet und signalnah repräsentiert, so dass eine größtmögliche Flexibilität im Hinblick auf konkrete Anwendungsfelder resultiert. Insbesondere die Eigenschaft von Mosaikbildern, das Sichtfeld einer Kamera in Raum und Zeit zu erweitern, ist dabei von grundlegender Bedeutung.

Mosaikbilder werden im Wesentlichen durch eine Transformation der Eingangsbilder in ein gemeinsames Referenzkoordinatensystem und eine anschließende Fusion ihrer Farbinformationen generiert. Sie finden insbesondere in der Computergrafik als Grundlage realer Texturen, sowie auch bei der Verarbeitung von Videosequenzen zur Hintergrundstabilisierung Anwendung. Ein großer Teil der derzeit existierenden Algorithmen ist dabei auf eine Berechnung hochqualitativer Mosaikbilder ausgerichtet. Dies bedingt einerseits, dass der Aufwand zur Berechnung der Bilder nur eine untergeordnete Rolle spielt und die Bilder einer Folge daher zumeist offline und simultan verarbeitet werden. Andererseits zielen auch die den Mosaikbildern zu Grunde liegenden Koordinatensysteme in derartigen Kontexten vorrangig auf eine optimale, verzerrungsfreie Darstellung der Mosaiks, die nicht notwendigerweise auch eine einfache Weiterverarbeitung der Daten unterstützt.

Interaktive Systeme stellen andere Anforderungen an Verfahren zur Mosaikbildberechnung. In diesem Anwendungsfeld sind vorrangig effiziente Algorithmen zur Berechnung der Bilder notwendig, die nicht im Widerspruch zur Interaktivität der Systeme stehen. Da fortwährend neue Bilddaten zu verarbeiten sind, ist es unerlässlich, Mosaikbilder schrittweise, d.h. online zu generieren und damit eine stetige Aktualisierung der internen Repräsentation und einen zeitlich weitgehend uneingeschränkten Zugriff auf die gespeicherten Daten zu gewährleisten. Darüber hinaus können sich die Vorteile eines visuellen Speichers nur in Kombination mit einem einfachen Datenzugriff und einer effizienten Handhabung voll entfalten. Insbesondere die direkte Anwendbarkeit gängiger Bildanalyseverfahren auf die Mosaikbilder bringt dabei eine große Flexibilität mit sich. Sie lässt sich jedoch nur durch die Bereitstellung euklidischer Koordinaten erreichen, da diese heutzutage die Grundlage nahezu aller existierenden Bildverarbeitungsansätze bilden.

Um den vorstehend skizzierten Anforderungen gerecht zu werden und interaktive Systeme als Einsatzgebiet für Mosaikbilder zu erschließen, wurde im Rahmen dieser Arbeit das neue Konzept der *Multi-Mosaikbilder* entwickelt. Es erweitert gängige Ansätze zur Berechnung von Mosaikbildern im Hinblick auf die spezifischen Rahmenbedingungen interaktiver Systeme, wobei eine adäquate, verzerrungsfreie Repräsentation von Bilfolgen stationärer, rotierender und zoomender Kameras im Vordergrund steht.

Zur Darstellung des vollständigen Sichtbereichs einer rotierenden Kamera finden oftmals zylindrische oder sphärische Koordinatensysteme Anwendung, die durch die Vermeidung geometrischer Verzerrungen die Erstellung hochqualitativer Mosaikbilder ermöglichen. Da ihre praktische Handhabung jedoch schwierig ist und auch keine Kompatibilität zu existierenden Bildanalyseverfahren gegeben ist, sind sie für den visuellen Speicher ungeeignet. Den Multi-Mosaiks liegen daher Referenzkoordinatensysteme auf Basis von Polyedern zu Grunde. Sie approximieren eine Kugel stückweise planar und stellen somit bei weitestgehend reduzierten Verzerrungen euklidische Koordinaten für eine nahtlose Einbindung der Bilder in bestehende Bildanalyse-Architekturen zur Verfügung. Polyeder sind auch in der Computergrafik als Grundlage zur Projektion von Bilddaten verbreitet, sie werden dort jedoch fast ausschließlich in Offline-Verfahren eingesetzt. Durch die Entwicklung geeigneter Algorithmen zur Online-Berechnung polyedrischer Mosaiks gelang es jedoch, diese Konzepte auch auf den vorliegenden Anwendungsfall zu übertragen.

Aktive Kameras sind mit einer hohen Flexibilität bei der Akquisition visueller Daten verknüpft. Dabei bieten neben reinen Rotationen insbesondere auch Veränderungen in der Bildweite große Spielräume für eine zielgerichtete Selektion relevanter Informationen in einer Szene. Eine Repräsentation ikonischer Daten in einem visuellen Speicher muss damit auch verschiedene Bildweiten und eine lokal variierende Granularität in den Eingangsdaten adäquat darstellen können. Die Multi-Mosaikbilder umfassen zu diesem Zweck eine Hierarchie verschieden skalierter Projektionskörper, die in Abhängigkeit von der Bildweite der Eingangsdaten als Repräsentationsziele ausgewählt werden und damit eine weitgehend verlustfreie Darstellung verschiedener Detailgrade ermöglichen.

Das Konzept der Multi-Mosaikbilder gründet auf gängigen Verfahren zur Berechnung von Mosaikbildern, wobei insbesondere robuste Algorithmen zur Registrierung von Bil-

dern sowie geeignete Integrationsheuristiken und Ansätze zur Behandlung unabhängig bewegter Objekte in einer Szene wichtige Bausteine bilden. Ausgehend von bekannten Ansätzen konnten dabei im Rahmen dieser Arbeit auch in diesen Bereichen Detailverbesserungen erzielt werden. Der der Bildregistrierung zu Grunde liegende Algorithmus des *Projective Flow* profitiert dabei unter anderem von einer gezielten Initialisierung der Parameterschätzung auf Basis der intrinsischen und extrinsischen Parameter der verwendeten Kameras, sowie von einer Selektion spezifischer Pixel für die Schätzung. Insbesondere bei einer Verarbeitung von Bildfolgen dynamischer Szenen konnten dadurch robustere Parameter geschätzt werden, die eine verbesserte Mosaikqualität bedingen.

Dynamische Szenen stellen grundsätzlich spezifische Anforderungen an eine Mosaikbildberechnung. Einerseits führen unabhängige Bewegungen, die dem globalen Bewegungsmodell nicht folgen, sowohl zu Schwierigkeiten bei der Registrierung als auch zu Fehlern bei der Bildintegration. Andererseits erfordert eine verlustfreie Repräsentation derartiger Bildfolgen auch einen geeigneten Umgang mit den statischen und dynamischen Daten, die sich nur schwer innerhalb eines einzelnen Mosaikbildes vereinen lassen. In der vorliegenden Arbeit wurden daher zusätzlich Mechanismen integriert, die neben der Erstellung eines ausschließlich die statischen Szenenanteile umfassenden Mosaikbildes auch eine Extraktion und gesonderte Darstellung dynamischer Daten ermöglichen.

Die Qualität der Multi-Mosaikbilder ist im Allgemeinen gut. Obgleich sich lokale Unstimmigkeiten innerhalb der Repräsentation aufgrund des unerlässlichen Online-Ansatzes nicht gänzlich ausschließen lassen, bilden die Bilder eine vielversprechende Grundlage zur besseren Handhabbarkeit von Bildsequenzen in interaktiven Systemen. Zur weiteren Verbesserung der Qualität der Bilder könnte jedoch perspektivisch eine Integration von Mechanismen zur Online-Korrektur von Registrierungsfehlern hilfreich sein. Neben den mit einer robusten, automatischen Detektion solcher Fehler verbundenen Schwierigkeiten gilt es dabei jedoch insbesondere Ansätze zu entwickeln, deren Aufwand nicht im Widerspruch zu einer Online-Verarbeitung steht. Eine mögliche Entwicklungsrichtung ist dabei etwa durch eine Parallelisierung geeigneter Algorithmen gegeben, die allerdings ebenso die Ausstattung der interaktiven Systeme berücksichtigen müssen.

Die mit einer Online-Korrektur von Registrierungsfehlern verbundene, robuste Detektion der Fehler ist eng mit der grundsätzlichen Entwicklung geeigneter Fehlermaße verknüpft, die eine über die heutzutage gängige, nahezu ausschließlich qualitative, visuelle Beurteilung von Mosaikbildern hinausgehende Bewertung der Qualität erlauben. Die Charakteristika der bei einer Registrierung von Bildern auftretenden Fehler unterscheiden sich signifikant von Artefakten, die etwa bei der Bildkompression eine verminderte Bildqualität bedingen, und tangieren damit ein im Rahmen der Entwicklung von Qualitätsmaßen für Bilder bislang noch kaum berücksichtigtes Forschungsfeld.

Weitere Verbesserungen der Parameterschätzung und damit der Mosaikqualität sind auch bei einer expliziten Berücksichtigung von Linsenverzerrungseffekten zu erwarten, die in der derzeitigen Implementierung noch nicht behandelt wurden, sich jedoch bei der Kalibrierung und in Bildfolgen mit einem insgesamt nur geringen Überlapp deutlich bemerkbar gemacht haben. Während diese Effekte einerseits im Rahmen einer geziel-

ten Berechnung geeigneter Korrekturfaktoren behandelt werden können, ist alternativ auch eine direkte Einbettung einer Linsenentzerrung in die Parameterschätzung selbst denkbar. Eine derartige Vorgehensweise, bei der sich unter Umständen auch gleichzeitig Parameter für eine radiometrische Bildkorrektur schätzen lassen, ist eng mit dem Ansatz des *Bundle Adjustment* verknüpft [Tri00]. Dieser setzt jedoch ebenfalls eine Offline-Verarbeitung der Bilddaten voraus, so dass seine Übertragbarkeit auf den vorliegenden Kontext zunächst detailliert zu prüfen ist.

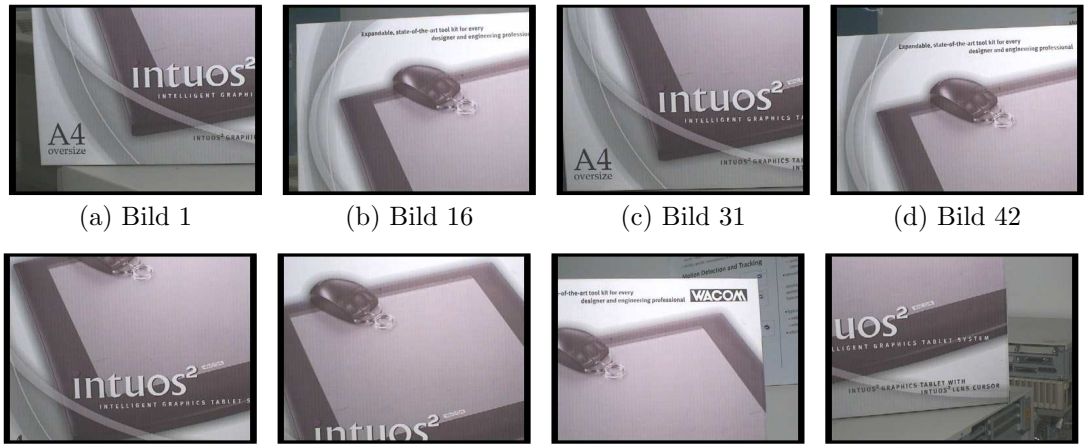
Die signalnahe Repräsentation visueller Daten in Multi-Mosaikbildern bietet grundsätzlich einen großen Spielraum bei der Einbindung des Konzeptes in verschiedene Anwendungsbereiche. Insbesondere die Bereitstellung euklidischer Koordinaten garantiert eine schnelle Adaption, die kaum Einschränkungen im Hinblick auf mögliche Einsatzfelder bedingt. Zur Illustrierung dieser großen Flexibilität wurden im Rahmen dieser Arbeit zwei Anwendungsfälle skizziert und evaluiert, die von den Multi-Mosaikbildern profitieren können. Einerseits konnten im Rahmen einer prototypischen Implementierung die Vorteile aufgezeigt werden, die der visuelle Speicher bei einer aktiven Szenenexploration bietet. Insbesondere das in Raum und Zeit erweiterte Sichtfeld der Kamera führt zu einer hohen Flexibilität bei der Auswahl geeigneter Fokuspunkte in einer Szene. Damit kann eine stärker zielgerichtete und damit insgesamt effizientere Auswertung der visuellen Daten erreicht werden. Letzteres ist insbesondere bei mobilen, interaktiven Systemen von hoher Bedeutung, die eine effiziente Mensch-Maschine-Interaktion anstreben.

In diesem Forschungsfeld liegt auch das zweite, exemplarisch betrachtete Anwendungsszenario, das in enger Zusammenarbeit mit Kollegen der Universität Bielefeld entwickelt wurde. Im Mittelpunkt der Arbeiten steht dabei BIRON, ein multimodal interagierender, mobiler Roboter. Zur Erweiterung seiner kognitiven Fähigkeiten wird derzeit ein Objektaufmerksamkeits-System entwickelt. Das diesem System zu Grunde liegende Objekterkennungsverfahren gründet auf der Auswertung verschiedener Ansichten eines Objektes. Da eine Akquirierung dieser Ansichten für einen mobilen Roboter mit großem Aufwand verbunden sein kann, soll der visuelle Speicher als zusätzliche Informationsquelle zur verbesserten Ausnutzung aufgenommener visueller Daten in BIRONs Architektur integriert werden. Erste Versuche haben dabei gezeigt, dass dieser Schritt zu einer insgesamt verbesserten und vor allem flexibleren Mensch-Roboter-Interaktion beitragen kann.

Zusammenfassend lässt sich folgern, dass die eingangs formulierten Ziele für die Entwicklung eines ikonischen Speichers als visuelles „Gedächtnis“ interaktiver Systeme durch das Konzept der Multi-Mosaikbilder erfolgreich erreicht werden konnten. Die ausgearbeiteten Datenstrukturen und Mechanismen für deren Handhabung und den Datenzugriff erfüllen die mit einem Einsatz des Speichers in interaktiven Systemen verbundenen Anforderungen. Im Rahmen zweier exemplarischer Anwendungsszenarien konnte darüber hinaus auch die praktische Eignung des Konzepts untermauert werden. Es empfiehlt sich damit als zusätzliche Komponente in der Architektur interaktiver Systeme, die auf diese Weise eine größere Flexibilität bei der Verarbeitung visueller Daten erlangen – eine unabdingbare Grundlage zur Erschließung neuer, jenseits der Labore gelegener Anwendungsfelder.

# A Bildsequenzen zur Evaluation

## A.1 Bildsequenz „2D-Verpackungskarton“

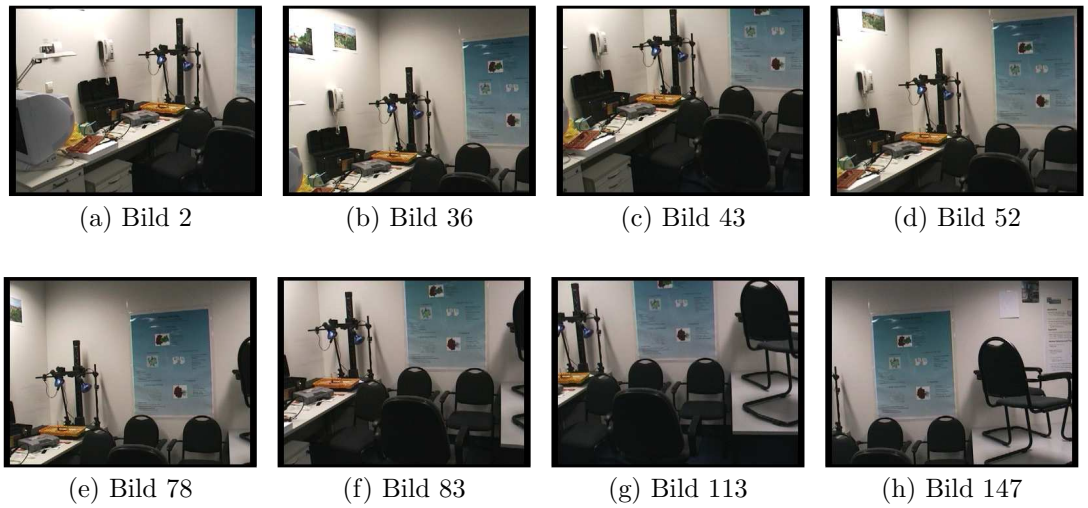



(a) Bild 1 (b) Bild 16 (c) Bild 31 (d) Bild 42

(e) Bild 79 (f) Bild 122 (g) Bild 182 (h) Bild 224

**Abbildung A.1:** Acht Beispielbilder der Sequenz, die dem Mosaik in Abbildung 3.9 zu Grunde lag.

## A.2 Bildsequenz „Labor-Scan“



(a) Bild 2 (b) Bild 36 (c) Bild 43 (d) Bild 52

(e) Bild 78 (f) Bild 83 (g) Bild 113 (h) Bild 147

**Abbildung A.2:** Acht Bilder der Folge, aus der das Mosaik in Abbildung 3.11 berechnet wurde.

## A.3 Bildsequenz „Frame-to-Mosaic“

(a) Bild 1 (b) Bild 21 (c) Bild 46 (d) Bild 55

(e) Bild 66 (f) Bild 81 (g) Bild 89 (h) Bild 97

**Abbildung A.3:** Acht Beispielbilder der Sequenz, die zur Evaluation des Frame-to-Mosaic-Modus verwendet wurde (vgl. Abb. 3.18 und 3.19).

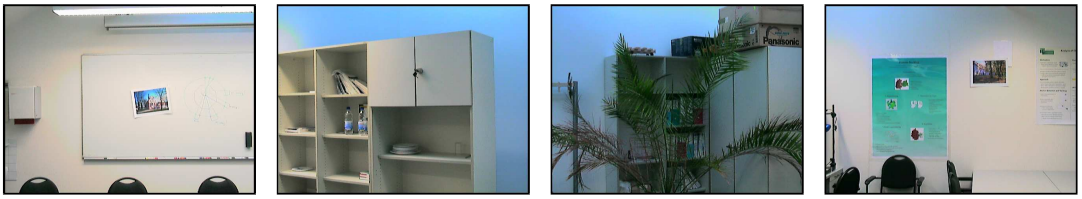
## A.4 Bildsequenz „Multi-Mosaik-Scan“

(a) Bild 1 (b) Bild 5 (c) Bild 9 (d) Bild 13

(e) Bild 17 (f) Bild 21 (g) Bild 25 (h) Bild 31

**Abbildung A.4:** Acht Beispielbilder einer Sequenz aus 32 Bildern, die der Berechnung des Multi-Mosaikbildes in Abbildung 6.21 zu Grunde lag.

## A.5 Bildsequenz „360-Grad-Scan“



(a) Bild 1 (b) Bild 31 (c) Bild 61 (d) Bild 91

(e) Bild 106 (f) Bild 121 (g) Bild 151 (h) Bild 166

**Abbildung A.5:** Acht Beispielbilder der Sequenz eines horizontalen Kamera-Scans. Die Sequenz, die insgesamt 360° umfasst, besteht aus 180 Bildern, die in 2°-Schritten aufgenommen wurden.

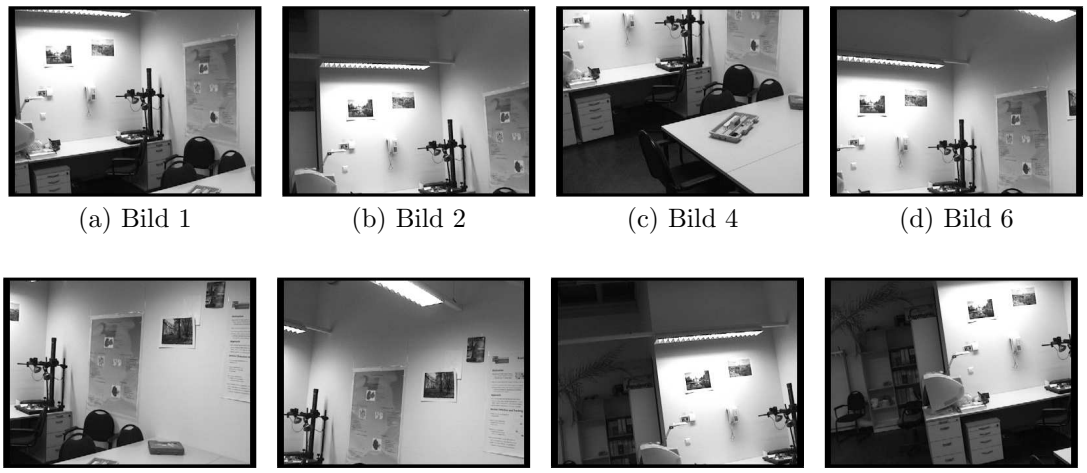
## A.6 Bildsequenz „Multi-Resolution“

(a) Bild 1 (b) Bild 10 (c) Bild 35 (d) Bild 66

(e) Bild 99 (f) Bild 139 (g) Bild 149 (h) Bild 168

**Abbildung A.6:** Acht Beispielbilder einer Sequenz aus 173 Bildern, die mit einer von Hand gesteuerten Kamera aufgenommen wurde und mehrere Auflösungsstufen umfasst (vgl. Abb. 6.23 und 6.24).

## A.7 Bildsequenz „Labor-Exploration“



(a) Bild 1 (b) Bild 2 (c) Bild 4 (d) Bild 6

(e) Bild 11 (f) Bild 12 (g) Bild 21 (h) Bild 23

**Abbildung A.7:** Acht Beispielbilder einer Sequenz, die im Rahmen einer aktiven Szenenexploration eines Raumes aufgenommen wurde (vgl. Abb. 7.5).

# B Geometrische Eigenschaften konvexer Polyeder

| Polyeder | Flächengrundformen | Anzahl | Innenwinkel der Flächen |
|---|---|---|---|
| Tetraeder | 3-eckig | 4 | $\langle 3 \times 60° \rangle$ |
| Hexaeder | 4-eckig | 6 | $\langle 4 \times 90° \rangle$ |
| Oktaeder | 3-eckig | 8 | $\langle 3 \times 60° \rangle$ |
| Dodekaeder | 5-eckig | 12 | $\langle 5 \times 108° \rangle$ |
| Ikosaeder | 3-eckig | 20 | $\langle 3 \times 60° \rangle$ |
| Kuboktaeder | 3-eckig | 8 | $\langle 3 \times 60° \rangle$ |
| | 4-eckig | 6 | $\langle 4 \times 90° \rangle$ |
| Ikosidodekaeder | 3-eckig | 20 | $\langle 3 \times 60° \rangle$ |
| | 5-eckig | 12 | $\langle 5 \times 108° \rangle$ |
| Rhombenkuboktaeder | 3-eckig | 8 | $\langle 3 \times 60° \rangle$ |
| | 4-eckig | 18 | $\langle 4 \times 90° \rangle$ |
| Ikosaederstumpf | 5-eckig | 12 | $\langle 5 \times 108° \rangle$ |
| | 6-eckig | 20 | $\langle 6 \times 120° \rangle$ |
| Rhombenikosidodekaeder | 3-eckig | 20 | $\langle 3 \times 60° \rangle$ |
| | 4-eckig | 30 | $\langle 4 \times 90° \rangle$ |
| | 5-eckig | 12 | $\langle 5 \times 108° \rangle$ |

**Tabelle B.1:** Flächenformen verschiedener konvexer Polyeder (nach [Mai03]).

| Polyeder | Eckengrad[1] | Winkel[2] der Körperecken | Flächenwinkel[3] der Körperecken |
|---|---|---|---|
| Tetraeder | 3 | $\langle 3 \times 60° \rangle$ | $\langle 3 \times 70,5288° \rangle$ |
| Hexaeder | 3 | $\langle 3 \times 90° \rangle$ | $\langle 3 \times 90° \rangle$ |
| Oktaeder | 4 | $\langle 4 \times 60° \rangle$ | $\langle 4 \times 109,4712° \rangle$ |
| Dodekaeder | 3 | $\langle 5 \times 108° \rangle$ | $\langle 3 \times 116,5651° \rangle$ |
| Ikosaeder | 5 | $\langle 5 \times 60° \rangle$ | $\langle 5 \times 138,1897° \rangle$ |
| Kuboktaeder | 4 | $\langle 60°/90°/60°/90° \rangle$ | $\langle 4 \times 125,2644° \rangle$ |
| Ikosidodekaeder | 4 | $\langle 60°/108°/60°/108° \rangle$ | $\langle 4 \times 142,6226° \rangle$ |
| Rhombenkuboktaeder | 4 | $\langle 60°/90°/90°/90° \rangle$ | $\langle 144,7356°/135°/135°/144,7356° \rangle$ |
| Ikosaederstumpf | 3 | $\langle 120°/108°/120° \rangle$ | $\langle 142,6226°/142,6226°/138,1897° \rangle$ |
| Rhombenikosidodekaeder | 4 | $\langle 60°/90°/108°/90° \rangle$ | $\langle 159,0948°/148,2825°/148,2825°/159,0948° \rangle$ |

**Tabelle B.2:** Winkeleigenschaften verschiedener konvexer Polyeder (nach [Mai03]).

[1]Der Grad einer Körperecke definiert sich über die Anzahl eingehender Kanten.

[2]Die Winkel einer Körperecke sind durch die Innenwinkel der angrenzenden Flächen gemäß der Reihenfolge ihrer Anordnung definiert.

[3]Die Flächenwinkel einer Körperecke beschreiben die Flächenwinkel zwischen je zwei aneinandergrenzenden Flächen an der Ecke in der Reihenfolge ihrer Anordnung.

# C Übersicht des Gesamtsystems

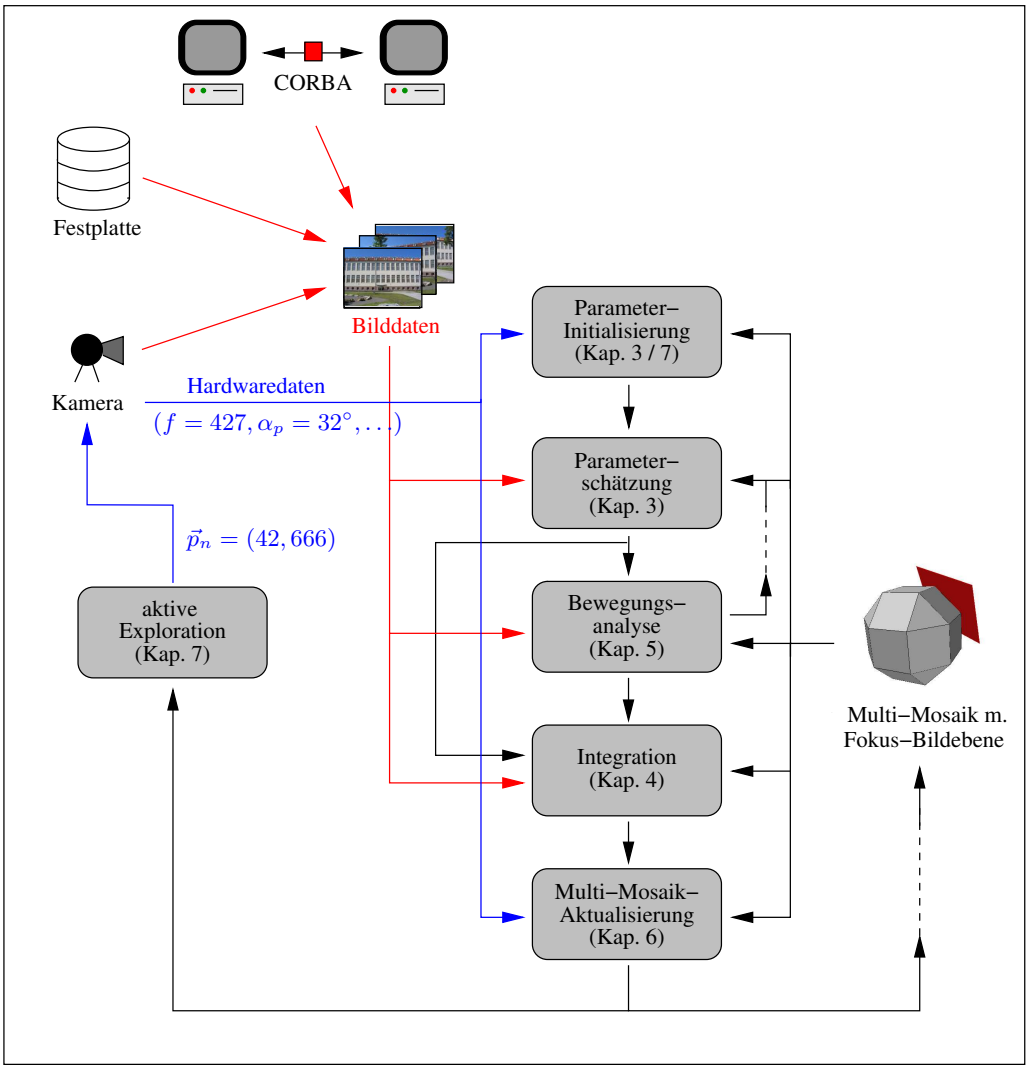



**Abbildung C.1:** Schema des Gesamtsystems zur Online-Berechnung von Multi-Mosaikbildern: Neben den grundlegenden Modulen zur Parameterschätzung und Bildintegration beinhaltet das System zusätzlich Komponenten zur expliziten Parameterinitialisierung, zur Bewegungsdetektion und -analyse sowie zur Aktualisierung der Multi-Mosaikdatenstruktur. Darüber hinaus kann bei Bedarf ein Modul zur aktiven Steuerung der Kamera im Rahmen einer aktiven Szenenexploration eingebunden werden. Das System verarbeitet derzeit Bilddaten, die direkt von der Festplatte oder von einem lokalen Grabber gelesen werden können. Außerdem steht eine CORBA-basierte Schnittstelle zur verteilten Bildaufnahme sowie perspektivisch zur Anbindung eines mobilen Roboters zur Verfügung [Wed05].

# Literaturverzeichnis


[Ahm91] S. Ahmad. *VISIT: An Efficient Computational Model of Human Attention*. Dissertation, Int. Comp. Science Institute, University of California, Berkeley, USA, 1991.

[Bal82] D. Ballard und C. Brown. *Computer Vision*. Prentice Hall, 1982.

[Bau04] C. Bauckhage, M. Hanheide, S. Wrede und G. Sagerer. A Cognitive Vision System for Action Recognition in Office Environments. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 2:827–832, Washington, USA, 2004.

[BE94] M. Ben-Ezra, S. Peleg und B. Rousso. Motion Segmentation Using Convergence Properties. In *Proc. of DARPA Image Understanding Workshop*, S. 1233–1235, Monterey, USA, Nov. 1994.

[BE98] M. Ben-Ezra, S. Peleg und M. Werman. Robust, Real-Time Motion Analysis. In *Proc. of DARPA Image Understanding Workshop*, S. 207–210, Monterey, USA, Nov. 1998.

[Ber92a] J.R. Bergen, P. Anandan, K.J. Hanna und R. Hingorani. Hierarchical Model-based Motion Estimation. In *Proc. of European Conf. on Comp. Vision*, S. 237–252, Santa Margherita Ligure, Italien, Mai 1992.

[Ber92b] J.R. Bergen, P.J. Burt, R. Hingorani und S. Peleg. A Three Frame Algorithm for Estimating Two-Component Image Motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 14:886–896, 1992.

[Ber94] D.F. Berman, J.T. Bartell und D.H. Salesin. Multiresolution Painting and Compositing. Technical Report UW-CSE-94-01-09b, Department of Comp. Science and Engineering, Univ. of Washington, Washington, USA, April 1994.

[Ber03] L.E. Berk. *Child Developement*. Allyn and Bacon, 6. Aufl., 2003.

[Bha00] K.S. Bhat, M. Saptharishi und P.K. Khosla. Motion Detection and Segmentation Using Image Mosaics. In *Proc. of Int. Conf. on Multimedia and Expo*, S. 1577–1580, New York, USA, 2000.

[Bie85] I. Biederman. Human Image Understanding: Recent Research and a Theory. *Computational Vision Graphics Image Processing*, S. 32:29–73, 1985.

[Bis95] G. Bishop und L. McMillan. Plenoptic Modeling: An Image-Based Rendering System. In *Proc. of SIGGRAPH, Annual Conf. on Comp. Graphics*, S. 39–46, Los Angeles, USA, Aug. 1995.

[Bob96] M. Bober, N. Georgis und J. Kittler. On Accurate and Robust Estimation of Fundamental Matrix. In *Proc. of British Machine Vision Conf.*, Edinburgh, Schottland, 1996.

[Böh03] H.J. Böhme, T. Wilhelm, T. Hempel, C. Schröter und H.-M. Gross. An Approach to Multimodal Human-Machine Interaction for Intelligent Service Robots. *Robotics and Autonomous Systems*, S. 44:83–96, 2003.

[Bol97] M. Bollmann, R. Hoischen und B. Mertsching. Integration of Static and Dynamic Scene Features Guiding Visual Attention. In E. Paulus und F.M. Wahl, Hrsg., *Mustererkennung*, S. 483–490. Springer, 1997.

[Bor98] H. Borotschnig, L. Paletta, M. Prantl und A. Pinz. Active Object Recognition in Parametric Eigenspace. In *Proc. of British Machine Vision Conf.*, S. 629–638, Southampton, England, 1998.

[Bou98a] E. Bourque und G. Dudek. Viewpoint Selection - An Autonomous Robotic System for Virtual Environment Creation. In *Proc. of Int. Conf. on Intelligent Robots and Systems*, S. 1:526–532, Victoria, Kanada, Okt. 1998.

[Bou98b] E. Bourque, G. Dudek und P. Ciaravola. Robotic Sightseeing - a Method for Automatically Creating Virtual Environments. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, S. 3186–3191, Leuven, Belgien, Mai 1998.

[Bra96] M. Brand. Understanding Manipulation in Video. In *Proc. of Int. Conf. on Face and Gesture Recognition*, S. 94–99, Killington, USA, Okt. 1996.

[Bre03] N. Bredeche, J.-D. Zucker und Z. Shi. Online Learning for Object Identification by a Mobile Robot. In *Proc. of the 5th Int. Symp. on Comp. Intelligence in Robotics and Automation*, S. 2:630–635, Kobe, Japan, 2003.

[Bro92] L. Gottesfeld Brown. A Survey of Image Registration Techniques. *ACM Comp. Surveys*, S. 24(4):325–376, 1992.

[Bur83a] P.J. Burt und E.H. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. on Communications*, S. COM-31(4):532–540, 1983.

[Bur83b] P.J. Burt und E.H. Adelson. A Multiresolution Spline with Application to Image Mosaics. *ACM Trans. on Graphics*, S. 2(4):217–236, 1983.

[Bur94] P.J. Burt und P. Anandan. Image Stabilization by Registration to a Reference Mosaic. In *Proc. of DARPA Image Understanding Workshop*, S. 1:425–434, Monterey, USA, Nov. 1994.

[Can86] J.F. Canny. A Computational Approach to Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 8(6):679–698, 1986.

[Cap04] D. Capel. *Image Mosaicing and Super-resolution*. Springer, 2004.

[Che95] S.E. Chen. QuickTime VR - An Image-Based Approach to Virtual Environment Navigation. In *Proc. of SIGGRAPH, Annual Conf. on Comp. Graphics*, S. 29–38, Los Angeles, USA, Aug. 1995.

[Che02] K. Cheoi und Y. Lee. A Feature-Driven Attention Module for an Active Vision System. In *Pattern Recognition, Proc. of DAGM Symp.*, LNCS 2449, S. 583–590, Zürich, Schweiz, Sep. 2002. Springer.

[Coh99] I. Cohen und G. Medioni. Detecting and Tracking Objects in Video Surveillance. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 2:2319–2325, Fort Collins, USA, Juni 1999.

[Coo00] S. Coorg und S. Teller. Spherical Mosaics with Quaternions and Dense Correlation. *Int. Journal of Comp. Vision*, S. 37(3):259–273, 2000.

[Cow97] N. Cowan. *Attention and Memory - An Integrated Framework*. Oxford University Press, 1997.

[Cri02] A. Criminisi. Single-View Metrology: Algorithms and Applications. In *Pattern Recognition, Proc. of DAGM Symp.*, LNCS 2449, S. 224–239, Zürich, Schweiz, Sep. 2002. Springer.

[dA99] L. de Agapito, R.I. Hartley und E. Hayman. Linear Self-Calibration of a Rotating and Zooming Camera. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 15–21, Fort Collins, USA, Juni 1999.

[dA00] L. de Agapito, E. Hayman und I. Reid. Self-Calibration of Rotating and Zooming Cameras. Technical Report OUEL 0225/00, Department of Engineering Science, University of Oxford, England, Sep. 2000.

[Dae99] F. Daellert, S. Thrun und C. Thorpe. Mosaicing a Large Number of Widely Dispersed, Noisy and Distorted Images: A Bayesian Approach. Technical Report CMU-RI-TR-99-34, Carnegie Mellon University, Pittsburgh, USA, 1999.

[Dar96] T. Darrell, P. Maes, B. Blumberg und A.P. Pentland. A Novel Environment for Situated Vision and Behavior. In M.S. Landy, L.T. Maloney und M. Pavel, Hrsg., *Exploratory Vision - The Active Eye*, Kap. 13. Springer, 1996.

[Dav98] J. Davis. Mosaics of Scenes with Moving Objects. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 1:97–100, Santa Barbara, USA, Juni 1998.

[Den02] J. Denzler und C.M. Brown. Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 24(2):145–157, Feb. 2002.

[Dor04] F. Dorsch, H. Häcker und K.-H. Stapf, Hrsg. *Psychologisches Wörterbuch.* Verlag Hans Huber, 2004. 14., vollst. überarb. und erw. Aufl.

[Edg02] S.Y. Edgerton. *Die Entdeckung der Perspektive.* Wilhelm Fink Verlag, 2002.

[Egn97] S. Egner und C. Scheier. Feature Binding Through Temporally Correlated Neural Activity in a Robot Model of Visual Perception. In *Proc. of Int. Conf. Artificial Neural Networks*, S. 703–708, Lausanne, Schweiz, 1997.

[Esk95] A.M. Eskicioglu und P.S. Fisher. Image Quality Measures and their Performance. *IEEE Trans. on Communications*, S. 43(12):2959–2965, 1995.

[Fau92] O.D. Faugeras, Q.-T. Luong und S.J. Maybank. Camera Self-Calibration: Theory and Experiments. In *Proc. of European Conf. on Comp. Vision*, S. 321–334, Santa Margherita Ligure, Italien, Mai 1992.

[Fel03] D. Feldman und A. Zomet. Least Distorted Mosaics. Technical Report 2003-71, Hebrew University, Jerusalem, Israel, 2003.

[Fin96] A. Finkelstein, C.E. Jacobs und D.H. Salesin. Multiresolution Video. *Computer Graphics*, S. 30:281–290, 1996. Annual Conference Series.

[Fis81] M.A. Fischler und R.C. Bolles. Random Sample Concensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Comm. of the Assoc. for Comp. Machinery (ACM)*, S. 24:381–395, 1981.

[Fri03] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G.A. Fink und G. Sagerer. Multi-Modal Anchoring for Human-Robot-Interaction. *Robotics and Autonomous Systems, Special Issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, S. 43(2-3):133–147, 2003.

[Gau97] P. Gaussier, C. Joulain, S. Zrehen, J.P. Banquet und A. Revel. Visual Navigation in an Open Environment without Map. In *Proc. of Int. Conf. on Intelligent Robots and Systems*, S. 545–550, Grenoble, Frankreich, Sep. 1997.

[Gel98] M. Gelgon und P. Bouthemy. Determining a Structured Spatio-Temporal Representation of Video Content for Efficient Visualization and Indexing. In *Proc. of European Conf. on Comp. Vision*, S. 1:595–609, Freiburg, Juni 1998.

[Gol02] E.B. Goldstein. *Wahrnehmungspsychologie*. Spektrum Verlag, 2. Aufl., 2002.

[Gon98] M.G. González, P. Holifield und M. Varley. Improved Video Mosaic Construction by Accumulated Alignment Error Distribution. In *Proc. of British Machine Vision Conf.*, S. 377–387, Southampton, England, 1998.

[Gra00] N. Gracias und J. Santos-Victor. Underwater Video Mosaics as Visual Navigation Maps. *Comp. Vision and Image Understanding*, S. 79(1):66–91, 2000.

[Gre86] N. Greene. Environment Mapping and Other Applications of World Projections. *IEEE Comp. Graphics and Applications*, S. 6(11):21–29, Nov. 1986.

[Güm96] Ş. Gümüştekin und R. Hall. Mosaic Image Generation on a Flattened Gaussian Sphere. In *Proc. of IEEE Workshop on Applications of Comp. Vision*, S. 50–55, Sarasota, USA, 1996.

[Haa03] A. Haasch. Extraktion von 3D-Objektinformationen aus korrelierten Kamera- und Laser-Daten. Diplomarbeit, Universität Bielefeld, 2003.

[Haa04] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede und G. Sagerer. BIRON – The Bielefeld Robot Companion. In E. Prassler, G. Lawitzky, P. Fiorini und M. Hägele, Hrsg., *Proc. of Int. Workshop on Advances in Service Robotics*, S. 27–32, Stuttgart, Mai 2004. Fraunhofer IRB Verlag.

[Han94] M. Hansen, P. Anandan, K. Dana, G. van der Wal und P. Burt. Real-time Scene Stabilization and Mosaic Construction. In *Proc. of DARPA Image Understanding Workshop*, S. 457–463, Monterey, USA, Nov. 1994.

[Har88] C.J. Harris und M. Stephens. A Combined Corner and Edge Detector. In *Proc. of Alvey Vision Conference*, S. 147–151, Manchester, England, Sep. 1988.

[Har99] R.I. Hartley, E. Hayman, L. de Agapito und I. Reid. Camera Calibration and the Search for Infinity. In *Proc. of Int. Conf. on Comp. Vision*, S. 1:510–517, Kerkyra, Griechenland, Sep. 1999.

[Har00] R. Hartley und A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[Hay02] M. Hayhoe, P. Aivar, A. Shrivastavah und R. Mruczek. Visual Short-term Memory and Motor Planning. In *Progress in Brain Research, Vol. 140*, S. 349–363. Elsevier Science B.V., 2002.

[Hof04] N. Hofemann, J. Fritsch und G. Sagerer. Recognition of Deictic Gestures with Context. In *Pattern Recognition, Proc. of DAGM Symp.*, LNCS 3175, S. 334–341, Tübingen, Aug./Sep. 2004. Springer.

[Hor80] B.K.P. Horn und B.G. Schunck. Determining Optical Flow. *Artificial Intelligence Memo No. 572*, April 1980.

[Hor98] T.S. Horowitz und J.M. Wolfe. Visual Search Has No Memory. *Nature*, S. 357:575–577, 1998.

[Ira94] M. Irani, B. Rousso und S. Peleg. Computing Occluding and Transparent Motions. *Int. Journal of Comp. Vision*, S. 12(1):5–16, 1994.

[Ira95] M. Irani, S. Hsu und P. Anandan. Video Compression Using Mosaic Representations. *Signal Processing: Image Communications.*, S. 7(4-6):529–552, 1995.

[Ira96] M. Irani, P. Anandan, J. Bergen, R. Kumar und S. Hsu. Efficient Representations of Video Sequences and their Applications. *Signal Processing: Image Communications*, S. 8:327–351, 1996.

[Ira98] M. Irani und P. Anandan. A Unified Approach to Moving Object Detection in 2D and 3D Scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 20(6):577–589, Juni 1998.

[Irw90] D.E. Irwin, J.L. Zacks und J.S. Brown. Visual Memory and the Perception of a Stable Visual Environment. *Perception & Psychophysics*, S. 47(1):35–46, 1990.

[Irw96] D.E. Irwin. Integrating Information Across Saccadic Eye Movements. *Current Directions in Psychological Science*, S. 5(3):94–100, Juni 1996.

[Ish94] H. Ishiguro, T. Maede, T. Miyashita und S. Tsuji. A Strategy for Acquiring an Environmental Model with Panoramic Sensing by a Mobile Robot. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, S. 724–729, San Diego, USA, Mai 1994.

[Ish96] H. Ishiguro und S. Tsuji. Image-Based Memory of Environment. In *Proc. of Int. Conf. on Intelligent Robots and Systems*, S. 634–639, Osaka, Japan, Nov. 1996.

[Itt01] L. Itti und C. Koch. Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, S. 2(3):194–203, März 2001.

[Itt03] L. Itti. Visual Attention. In M. A. Arbib, Hrsg., *The Handbook of Brain Theory and Neural Networks*, S. 1196–1201. MIT Press, 2. Aufl., 2003.

[Iva99] Y. Ivanov, C. Stauffer, A. Bobick und E. Grimson. Video Surveillance of Interactions. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition, Workshop on Visual Surveillance*, S. 82–89, Fort Collins, USA, Juni 1999.

[Jäh97] B. Jähne. *Digitale Bildverarbeitung*. Springer, 4. Aufl., 1997.

[Jet98] M. Jethwa, A. Zisserman und A. Fitzgibbon. Real-time Panoramic Mosaics and Augmented Reality. In *Proc. of British Machine Vision Conf.*, S. 852–862, Southampton, England, 1998.

[Jia05] J. Jia und C.-K. Tang. Tensor Voting for Image Correction by Global and Local Intensity Alignment. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 27(1):36–50, Jan. 2005.

[Jon01] M.J. Jones und P. Viola. Robust Real-time Object Detection. Technical Report CRL-2001-1, Cambridge Research Laboratory, Cambridge, USA, 2001.

[Jun98] N. Jungclaus, M. von der Heyde, H. Ritter und G. Sagerer. An Architecture for Distributed Visual Memory. *Zeitschrift für Naturforschung*, S. 53c:550–559, 1998.

[Kan99] K. Kanatani und N. Ohta. Accuracy Bounds and Optimal Computation of Homography for Image Mosaicing Applications. In *Proc. of Int. Conf. on Comp. Vision*, S. 1:73–78, Kerkyra, Griechenland, 1999.

[Kan00] E.-Y. Kang, I. Cohen und G. Medioni. A Graph-based Global Registration for 2D Mosaics. In *Proc. of Int. Conf. on Pattern Recognition*, S. 1:257–260, Barcelona, Spanien, 2000.

[Kap04] F. Kaplan und V.V. Hafner. The Challenges of Joint Attention. In *Proc. of the 4th Int. Workshop on Epigenetic Robotics*, S. 67–74, Genua, Italien, 2004.

[Kat94] H. Kattner. Using Attention As A Link Between Low-Level And High-Level Vision. Technical Report TUM-I9439, Institut für Informatik, Technische Universität München, 1994.

[Kes00] H. Kestler, S. Sablatnög, S. Simon, S. Enderle, A. Baune, G.K. Kraetzschmar, F. Schwenker und G. Palm. Concurrent Object Identification and Localization for a Mobile Robot. *Künstliche Intelligenz*, S. 23–29, April 2000.

[Kim00a] H. Kim und K.S. Hong. Soccer Video Mosaicing using Self-Calibration and Line Tracking. In *Proc. of Int. Conf. on Pattern Recognition*, S. 1:592–595, Barcelona, Spanien, 2000.

[Kim00b] H.-S. Kim, H.-C. Kim, W.-K. Lee und C.-H. Kim. Stitching Reliability for Estimating Camera Focal Length in Panoramic Image Mosaicing. In *Proc. of Int. Conf. on Pattern Recognition*, S. 1:596–599, Barcelona, Spanien, 2000.

[Kow95] E. Kowler. Eye Movements. In D.N. Osherson, Hrsg., *An Invitation to Cognitive Science (2): Visual Cognition*, Kap. 6, S. 215–265. MIT Press, 2. Aufl., 1995.

[Lan03] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G.A. Fink und G. Sagerer. Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot. In *Proc. of Int. Conf. on Multimodal Interfaces*, S. 28–35, Vancouver, Kanada, Nov. 2003. ACM.

[Lao00] T.-K. Lao, K.-H. Wong, K.-S. Lee und S.-H. Or. Creating Virtual Walkthrough Environment From Vertical Panoramic Mosaic. In *Proc. of Int. Conf. on Pattern Recognition*, S. 1:575–578, Barcelona, Spanien, 2000.

[Lip99] A.J. Lipton. Virtual Postman - Real-Time, Interactive Virtual Video. Technical Report CMU-RI-TR-99-12, Carnegie Mellon Univ., Pittsburgh, USA, 1999.

[Mac03] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[Mai03] P.H. Maier. *Körper und Raum - Die dritte Dimension.* Verlag Franzbecker, 2003.

[Man93] R. Manduchi und G.A. Mian. Accuracy Analysis for Correlation-based Image Registration Algorithms. In *IEEE Int. Symp. on Circuits and Systems*, S. 834–837, Chicago, USA, 1993.

[Man96] S. Mann und R.W. Picard. Video Orbits of the Projective Group: A New Perspective on Image Mosaicing. Technical Report 338, MIT Media Laboratory Perceptual Computing Section, Boston, USA, 1996.

[Mat92] M.W. Matlin und H.J. Foley. *Sensation and Perception.* Allyn and Bacon, 3. Aufl., 1992.

[Mat96] Y. Matsumoto, M. Inaba und H. Inoue. Visual Navigation using View-Sequenced Route Representation. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, S. 83–88, Minneapolis, USA, April 1996.

[McG76] C.D. McGillem und M. Svedlow. Image Registration Error Variance as a Measure of Overlay Quality. *IEEE Trans. on Geoscience Electronics*, S. 14(1):44–49, 1976.

[Med98] G. Medioni, R. Nevatia und I. Cohen. Event Detection and Analysis from Video Streams. In *Proc. of DARPA Image Understanding Workshop*, S. 63–72, Nov. 1998. Monterey, USA.

[Med01] G. Medioni, I. Cohen, F. Brémond, S. Hongeng und R. Nevatia. Event Detection and Analysis from Video Streams. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 23(8):873–889, Aug. 2001.

[Még99] R. Mégret und C. Saraceno. Building the Background Mosaic of an Image Sequence. Technical Report PRIP-TR-060, TU Wien, Österreich, 1999.

[Mil94] R. Milanese, T. Pun, S. Gil und J.-M. Bost. Exploiting Dynamic Aspects of Visual Perception for Object Recognition. In *Proc. of PerAc - From Perception to Action*, S. 193–205, Lausanne, Schweiz, Sep. 1994.

[Mit00] A. Mittal und D. Huttenlocher. Scene Modeling for Wide Area Surveillance and Image Synthesis. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 160–167, Hilton Head, USA, Juni 2000.

[Mod04] J. Modersitzki. *Numerical Methods for Image Registration*. Numerical Mathematics and Scientific Computation. Oxford Science Publications, 2004.

[Moh96] R. Mohr und B. Triggs. Projective Geometry for Image Analysis, Juli 1996. A Tutorial given at International Symposium of Photogrammetry and Remote Sensing, Wien, Österreich.

[Möl01a] B. Möller. Detektion von Bewegung bei der Berechnung von Mosaikbildern. Diplomarbeit, Universität Bielefeld, 2001.

[Möl01b] B. Möller und S. Posch. Detection and Tracking of Moving Objects for Mosaic Image Generation. In *Pattern Recognition, Proc. of DAGM Symp.*, LNCS 2191, S. 208–215, München, Sep. 2001. Springer.

[Möl02] B. Möller und S. Posch. Analysis of Object Interactions in Dynamic Scenes. In *Pattern Recognition, Proc. of DAGM Symp.*, LNCS 2449, S. 361–369, Zürich, Schweiz, Sep. 2002. Springer.

[Möl03] B. Möller, D. Williams und S. Posch. Robust Image Sequence Mosaicing. In *Pattern Recognition, Proc. of DAGM Symp.*, LNCS 2781, S. 386–393, Magdeburg, Sep. 2003. Springer.

[Möl04] B. Möller, D. Williams und S. Posch. Towards a Mosaic-based Visual Representation of Large Scenes. *Int. Journal on Pattern Recognition and Image Analysis, Spec. Issue*, S. 14(2):262–266, 2004.

[Möl05] B. Möller und S. Posch. A Mosaic-based Visual Memory with Applications to Active Scene Exploration. In *Proc. of Int. Conf. Mirage*, S. 117–125, INRIA Rocquencourt, Frankreich, März 2005.

[Nav02] V. Navalpakkam und L. Itti. A Goal Oriented Attention Guidance Model. In *Proc. of 2nd Workshop on Biologically Motivated Computer Vision*, S. 453–461, Tübingen, November 2002.

[Nay90] B. Naylor, J. Amanatides und W. Thibault. Merging BSP Trees Yields Polyhedral Set Operations. *Proc. of SIGGRAPH, Annual Conf. on Comp. Graphics*, S. 24(4):115–124, Aug. 1990.

[Pal00] L. Paletta und A. Pinz. Active Object Recognition by View Integration and Reinforcement Learning. *Robotics and Autonomous Systems*, S. 31(1-2):1–18, 2000.

[Par94] B. Parhami. Voting Algorithms. *IEEE Trans. on Reliability*, S. 43(4):617–629, Dez. 1994.

[Pas95] H. Pashler. Attention and Visual Perception: Analyzing Divided Attention. In Daniel N. Osherson, Hrsg., *An Invitation to Cognitive Science (2): Visual Cognition*, Kap. 2, S. 71–100. MIT Press, 2. Aufl., 1995.

[Pea78] P. Pearce und S. Pearce. *Polyhedra Primer*. Van Nostrand Reinhold Company, 1978.

[Pel97] S. Peleg und J. Herman. Panoramic Mosaics by Manifold Projection. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 338–343, San Juan, Puerto Rico, Juni 1997.

[Pel00] S. Peleg, B. Rousso, A. Rav-Acha und A. Zomet. Mosaicing on Adaptive Manifolds. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 22(10):1144–1154, Okt. 2000.

[Ple99] R. Pless, T. Brodský und Y. Aloimonos. Independent Motion: The Importance of History. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 2085–2091, Fort Collins, USA, Juni 1999.

[Pol99] M. Pollefeys, R. Koch und L. Van Gool. Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters. *Int. Journal of Comp. Vision*, S. 32(1):7–25, 1999.

[Pre92] W.H. Press, S.A. Teukolsky, W.T. Vetterling und B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2. Aufl., 1992.

[Rae01] R. Rae. *Gestikbasierte Mensch-Maschine-Kommunikation auf der Grundlage visueller Aufmerksamkeit und Adaptivität*. Dissertation, Technische Fakultät, Universität Bielefeld, 2001.

[Ras96] C. Rasmussen und G.D. Hager. Robot Navigation Using Image Sequences. In *Proc. of the 13th Nat. Conf. on Artificial Intelligence/8th Innovative Applications of Artificial Intelligence Conf.*, S. 2:938–943, Portland, USA, 1996.

[Ric96] G. Rickheit und I. Wachsmuth. Collaborative Research Centre "Situated Artificial Communicators" at the University of Bielefeld, Germany. In Paul McKevitt, Hrsg., *Integration of Natural Language and Vision Processing*, Jgg. IV, S. 11–16, Dordrecht, 1996. Kluwer.

[Rob03a] D. Robinson und P. Milanfar. Fundamental Performance Limits in Image Registration. *IEEE Trans. on Image Processing*, S. 13(9):1185–1199, 2003.

[Rob03b] J.A. Robinson. A Simplex-Based Projective Transform Estimator. *Visual Information Engineering (VIE)*, S. 290–293, Juli 2003.

[Roy04] S. Dutta Roy, S. Chaudhury und S. Banerjee. Active Recognition through Next View Planning: A Survey. *Pattern Recognition*, S. 37(3):429–446, 2004.

[Sag97] G. Sagerer und H. Niemann. *Semantic Networks for Understanding Scenes*. Plenum Press, 1997.

[Saw96] H.S. Sawhney und S. Ayer. Compact Representations of Videos Through Dominant and Multiple Motion Estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 18(8):814–830, Aug. 1996.

[Saw98] H. S. Sawhney, S. Hsu und R. Kumar. Robust Video Mosaicing Through Topology Inference and Local to Global Alignment. In *Proc. of European Conf. on Comp. Vision*, S. 103–119, Freiburg, 1998.

[Saw99] H.S. Sawhney und R. Kumar. True Multi-Image Alignment and its Application to Mosaicing and Lens Distortion Correction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 21(3):235–243, 1999.

[Sch90] H.R. Schiffmann. *Sensation and Perception - an Integrated Approach*. John Wiley & Sons, Inc., 3. Aufl., 1990.

[Sch95] K. Schill und C. Zetschke. A Model of Visual Spatio-temporal Memory: The Icon Revisited. *Psychological Research*, S. 57:88–102, 1995.

[Sch98a] W.X. Schneider. Visual-spatial Working Memory, Attention, and Scene Representation: A Neuro-cognitive Theory. *Psychological Research*, S. 62:220–236, 1998.

[Sch98b] W.X. Schneider und S. Maasen, Hrsg. *Mechanisms of Visual Attention - A Cognitive Neuroscience Perspective*. Psychology Press, 1998.

[Shu99] H.-Y. Shum und L.-W. He. Rendering with Concentric Mosaics. In *SIGGRAPH Comp. Graphics*, S. 33:299–306, 1999.

[Shu00] H.-Y. Shum und R. Szeliski. Systems and Experiment Paper: Construction of Panoramic Image Mosaics with Global and Local Alignment. *Int. Journal of Comp. Vision*, S. 36(2):101–130, Feb. 2000.

[Sla02] M. Slater, A. Steed und Y. Chrysanthou. *Computer Graphics and Virtual Environments*. Addison-Wesley, 2002.

[Smi97] S.M. Smith und J.M. Brady. SUSAN - A New Approach to Low Level Image Processing. *Int. Journal of Comp. Vision*, S. 23(1):45–78, 1997.

[Stu97] P. Sturm. Critical Motion Sequences for Monocular Self-Calibration and Uncalibrated Euclidean Reconstruction. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 1100–1105, San Juan, Puerto Rico, Juni 1997.

[Sue92] P. Suetens, P. Fua und A.J. Hanson. Computational Strategies for Object Recognition. *ACM Comp. Surveys*, S. 24(1):5–62, März 1992.

[Swa93] M.J. Swain und M. Stricker. Promising Directions in Active Vision. *Int. Journal of Comp. Vision*, S. 11(2):109–126, 1993.

[Syp99] D. Sypli und H. Tappe. Konstruktion von Mosaikbildern für die Bildanalyse. Diplomarbeit, Universität Bielefeld, 1999.

[Sze96] R. Szeliski. Video Mosaics for Virtual Environments. *IEEE Comp. Graphics and Applications*, S. 16(2):22–30, März 1996.

[Sze97] R. Szeliski und H.-Y. Shum. Creating Full View Panoramic Image Mosaics and Environment Maps. *Proc. of SIGGRAPH, Annual Conf. on Comp. Graphics*, S. 31:251–258, 1997.

[Tel98] S. Teller. Toward Urban Model Acquisition from Geo-Located Images. In *Proc. of Pacific Graphics*, S. 45–51, Singapur, Okt. 1998.

[Top04] I. Toptsis, S. Li, B. Wrede und G.A. Fink. A Multi-modal Dialog System for a Mobile Robot. In *Proc. of Int. Conf. on Spoken Language Processing*, S. 1:273–276, Jeju, Korea, 2004.

[Tor97] P.H.S. Torr und D.W. Murray. The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix. *Int. Journal of Comp. Vision*, S. 24(3):271–300, 1997.

[Tor00a] B. Tordoff und D. Murray. Violating Rotating Camera Geometry: The Effect of Radial Distortion on Self-Calibration. In *Proc. of Int. Conf. on Pattern Recognition*, S. 423–427, Barcelona, Spanien, 2000.

[Tor00b] P.H.S. Torr und A. Zisserman. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Comp. Vision and Image Understanding*, S. 78:138–156, 2000.

[Tor03] P.H.S. Torr und C. Davidson. IMPSAC: Synthesis of Importance Sampling and Random Sample Concensus. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 25(2):354–364, 2003.

[Tor04] B. Tordoff und D. Murray. Reactive Control of Zoom while Fixating Perspective and Affine Cameras. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 26(1):98–112, Jan. 2004.

[Tri00] B. Triggs, P. McLauchlan, R. Hartley und A. Fitzgibbon. Bundle Adjustment – A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, Lecture Notes in Computer Science, S. 298–375. Springer Verlag, 2000.

[Tsa86] R.Y. Tsai. An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 364–374, Miami Beach, USA, 1986.

[Tsu93] S. Tsuji und S. Li. Making Cognitive Map of Outdoor Environment. In *Proc. of Int. Joint Conf. on Artificial Intelligence*, S. 2:1632–1638, Chambéry, France, Aug. 1993.

[Tur91] M. Turk und A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, S. 3(1):71–86, 1991.

[Uyt01] M. Uyttendaele, A. Eden und R. Szeliski. Eliminating Ghosting and Exposure Artifacts in Image Mosaics. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 2:509–516, Hawaii, Dez. 2001.

[vdH04] A.H.C. van der Heijden. *Attention in Vision - Perception, Communication and Action*. Psychology Press, 2004.

[Vel96] B.M. Velichkovsky, M. Pomplun und H. Rieser. Attention and Communication: Eye-Movement-Based Research Paradigms. In W.H. Zangemeister, H.S. Stiehl und C. Freksa, Hrsg., *Visual Attention and Cognition*, Advances in Psychology (116), S. 125–154. Elsevier, Amsterdam, 1996.

[Vin95] J. Vince. *Virtual Reality Systems*. Addison-Wesley, 1995.

[Wan02] Z. Wang und A. C. Bovik. A Universal Image Quality Index. *IEEE Signal Processing Letters*, S. 9(3):81–84, März 2002.

[Wan03] Z. Wang, H. R. Sheikh und A. C. Bovik. Objective Video Quality Assessment. In B. Furht und O. Marqure, Hrsg., *The Handbook of Video Databases: Design and Applications*, Kap. 41, S. 1041–1078. CRC Press, Sep. 2003.

[Wan04] L. Wang, S.B. Kang, H.-Y. Shum und G. Xu. Error Analysis of Pure Rotation-based Self-Calibration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, S. 26(2):275–280, Feb. 2004.

[Web00] M. Weber, M. Welling und P. Perona. Towards Automatic Discovery of Object Categories. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 2:101–108, Hilton Head Island, USA, 2000.

[Wed05] M. Wedekind. Roboterunterstützte Generierung von Multiresolution-Mosaikbildern. Diplomarbeit, Martin-Luther-Universität Halle-Wittenberg, 2005.

[Wen93] J. Weng, T.S. Huang und N. Ahuja. *Motion and Structure from Image Sequences*. Springer-Verlag, 1993.

[Wil03] D. Williams, B. Möller und S. Posch. Integrated System for a Visual Memory Based on Mosaics. In H.R. Arabnia und Y. Mun, Hrsg., *Proc. of Int. Conf. on Imaging Science, Systems, and Technology*, S. 2:633–639, Las Vegas, USA, Juni 2003. CSREA Press.

[Wol00a] J.M. Wolfe. Visual Attention. In K.K. De Valois, Hrsg., *Seeing*, S. 335–386. Academic Press, San Diego, USA, 2. Aufl., 2000.

[Wol00b] J.M. Wolfe, N. Klempen und K. Dahlen. Post-attentive Vision. *The Journal of Experimental Psychology: Human Perception and Performance*, S. 26(2):693–716, 2000.

[Wol03] J.M. Wolfe, S.J. Butcher, C. Lee und M. Hyle. Changing Your Mind: On the Contributions of Top-Down and Bottom-Up Guidance in Visual Search for Feature Singletons. *The Journal of Experimental Psychology: Human Perception and Performance*, S. 29(2):483–502, 2003.

[Wol04] J.M. Wolfe und T.S. Horowitz. What Attribues Guide the Deployment of Visual Attention and How Do They Do It? *Nature Reviews, Neuroscience*, S. 5:1–7, Juni 2004.

[Xu94] W. Xu und G. Hauske. Picture Quality Evaluation Based on Error Segmentation. In *Proc. of the Int. Society for Optical Engineering*, S. 2308:1454–1465, 1994.

[Zan96] W.H. Zangemeister, H.S. Stiehl und C. Freksa, Hrsg. *Visual Attention and Cognition*. Advances in Psychology, 116. Elsevier North-Holland, 1996.

[Zei96] E. Zeidler, Hrsg. *Teubner - Taschenbuch der Mathematik*. B.G.Teubner Verlagsgesellschaft, Stuttgart/Leipzig, 1996.

[Zog97] I. Zoghlami, O. Faugeras und R. Deriche. Using Geometric Corners to Build a 2D Mosaic from a Set of Images. In *Proc. of Int. Conf. on Comp. Vision and Pattern Recognition*, S. 420–425, San Juan, Puerto Rico, Juni 1997.

[Zom00] A. Zomet und S. Peleg. Efficient Super-Resolution and Applications to Mosaics. In *Proc. of Int. Conf. on Pattern Recognition*, S. 1:579–583, Barcelona, Spanien, 2000.

## Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig und ohne fremde Hilfe verfasst habe. Ich habe keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt. Die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen sind als solche kenntlich gemacht worden. Ich habe mich bisher nicht um den Doktorgrad beworben.

Halle (Saale), den 10. Mai 2005 Birgit Möller

# Lebenslauf

## Persönliche Daten

| | |
|---|---|
| Name | Birgit Möller |
| geboren am | 20. September 1976 |
| in | Bielefeld |
| Staatsangehörigkeit | deutsch |
| Familienstand | ledig |

## Schulausbildung

| | |
|---|---|
| 8/1983 - 7/1987 | Eichendorff-Grundschule in Bielefeld |
| 8/1987 - 6/1996 | Max-Planck-Gymnasium in Bielefeld,<br>Abschluss Abitur |

## Universitätsausbildung

| | |
|---|---|
| 10/1996 - 3/2001 | Studium der naturwissenschaftlichen Informatik<br>an der Universität Bielefeld,<br>Abschluss Diplom-Informatikerin (Dipl.-Inform.) |
| 1/1999 - 11/2000 | studentische Hilfskraft an der Universität Bielefeld,<br>Arbeitsgruppe Angewandte Informatik |
| 4/2001 - 9/2002 | wissenschaftliche Mitarbeiterin an der Universität<br>Bielefeld, Arbeitsgruppe Angewandte Informatik |
| seit 10/2002 | wissenschaftliche Mitarbeiterin an der<br>Martin-Luther-Universität Halle-Wittenberg,<br>Arbeitsgruppe Mustererkennung & Bioinformatik |

Halle (Saale), Mai 2005 Birgit Möller