3-D Objekterkennung und Szeneninterpretation

Ein System zur multimodalen Beschreibung von Innenraumszenen

Michael Wünstel

Dissertation
zur Erlangung des Grades eines Doktors der Ingenieurwissenschaften
– Dr.-Ing. –

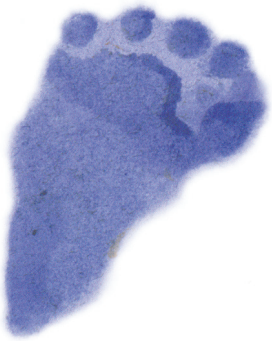
Vorgelegt im Fachbereich 3 (Mathematik und Informatik)
Universität Bremen

28. Januar 2009

Datum des Promotionskolloquiums: 7. Januar 2009

Gutachter

Prof. Dr. Bernd Krieg-Brückner (Universität Bremen)
Prof. Dr. Kerstin Schill (Universität Bremen)



Für Anton

Danksagung

Mein erster Dank gilt Bernd Krieg-Brückner und Kerstin Schill für die Überlassung des Themas, dem dabei gewährten Gestaltungsfreiraum und nicht zuletzt ihrer Fürsorge, auch für eine finanzielle Unterstützung Verantwortung zu tragen.

Ich danke Reinhard Moratz für die Einführung in die Thematik der Raumkognition. Mein besonderer Dank gilt Thomas Röfer für die fruchtbare Diskussionen vieler Aspekte des entwickelten Systems. In diesem Zusammenhang wurden wichtige Prozesse der 2-D und 3-D Registrierung und der form- und farbbasierten Objekterkennung in verschiedenen Diplomarbeiten behandelt, die auch für diese Arbeit wichtige Informationen liefern. Hier möchte ich insbesondere Andreas Meyer, Benjamin Schröder, Nils Breden, Christian Kanthak, Julia Finken und Istvan Lovas danken, die sehr engagiert, nicht selten bis spät in die Nacht, an der Entwicklung des Systems mitgearbeitet haben und erreichbar waren.

Speziell in Bezug auf die vorliegende schriftliche Arbeit möchte ich Kai Hübner danken. Trotz der örtlichen Distanz war er stets bereit mir vielfältige Hinweise und kostbare Anmerkungen zu geben.

Ich danke meinen (ehemaligen) Kollegen Kai Hübner, Christian Mandel und Tillmann Vierhuff für die vielen interessanten Gespräche und Diskussionen, die dazu beitrugen den Horizont in Bezug auf andere Projekte, aber auch weit darüber hinausgehend, zu erweitern.

Zuletzt danke ich meiner kleinen Familie, meiner Frau Ulrike und meiner Tochter Franziska, die mich einerseits direkt unterstützt hat und andererseits für einen Ausgleich für die Schreibtischarbeit gesorgt hat.

Inhaltsverzeichnis

iii

Abbildungsverzeichnis **ix**

Tabellenverzeichnis **xiii**

1 Einführung **1**

2 Kognitive Aspekte der Objekterkennung **5**

2.1 Objekterkennung 6
2.2 Visuelle Repräsentation des Menschen 8
2.3 Erkennung durch invariante Eigenschaften 10
2.4 Erkennung durch strukturelle Beschreibungen 14
2.5 Erkennung durch Ausrichtung und Transformation 18
2.6 Erkennung durch multiple Gedächtnistransformation 19
2.7 Struktur und Funktionalität 21
2.8 Cognitive Vision 25
2.9 Funktionalitätsbasierte Objekterkennungsverfahren 30
2.10 Mensch-Maschine Kommunikation 34

3 Komponenten des ORCC-Systems **39**

3.1 Sensordaten 42
3.2 Oberflächensegmentierung 43
3.2.1 Nachbarschaft über Distanz 47
3.2.2 Nachbarschaft über Index 48
3.3 Erkennung einfach strukturierter Objekte 50
3.3.1 Bayessches Netz als Objektmodell 51
3.3.2 Interaktives Lernmodul 55
3.3.3 Anwendung: Raumbegrenzungen 58

3.4 Registrierung 60
3.5 Strukturbasierte Objekterkennung 62
3.6 Segmentierung auf funktionalen Ebenen 68
3.7 Umgebungsboxberechnung 71
3.8 3-D ansichtenspezifische formbasierte Objekterkennung 73
3.9 2-D/3-D Systemkalibrierung 77
3.10 Texturausschnittsberechnung 82
3.11 Farbnormierung und Farbbeschreibung 86
3.12 Geometrische Szenenbeschreibung 90
3.13 Räumliche Relationen und Benutzerdialog 91

4 Anwendung **97**

5 Zusammenfassung und Ausblick **107**

A Anhang **111**
A.1 ORCC-System: Technische Beschreibung 113
A.2 Optische Wahrnehmung 115
A.3 Begriffe der Wahrscheinlichkeitsrechnung 119
A.4 Bayessche Netzwerke 125
A.5 Merkmalsparameter eines Stuhls 131

Literaturverzeichnis **133**

Abbildungsverzeichnis

1.1 Entwicklungsumgebung *ORCC* 2

2.1 Schwierigkeiten bei der Objekterkennung 6
2.2 Designerstühle 7
2.3 Merkmalbasierte Erkennung 11
2.4 Rekonstruktion eines Raumschiffs mit Zernike-Funktionen 14
2.5 Strukturbasierte Erkennung nach Marr 14
2.6 Hierarchisch organisiertes 3-D Modell nach Marr 16
2.7 Geon 17
2.8 Modellbasierte Erkennung 19
2.9 Modellbasierte Verfolgung eines Autos 20
2.10 Light Field Descriptor 20
2.11 Struktur und Funktion 25
2.12 Computerprahik und Computer Vision 27
2.13 Funktionale Objekterkennung: Kategorien-Definitions-Baum von Stühlen 32
2.14 Funktionales Modell eines Hammers 33
2.15 Objektauswahl durch einen Roboter des MIT 35
2.16 PeopleBot der Universität Ulm 36

3.1 Komponenten des ORCC-Systems 40
3.2 Verschattungspunkte an Tischkante 42
3.3 Segmentierte Tischoberfläche 43
3.4 Regionenwachstum: Koplanarität und Konormalität 46
3.5 Segmentierung über geometrische Nachbarschaft 49
3.6 Segmentierung über Indexnachbarschaft 50
3.7 Funktionalbasierte Erkennung einfach strukturierter Objekte 52
3.8 Lernen geometrischer Eigenschaften 56
3.9 Segmentierte Innenraumszenen 59

3.10 Merkmalsverteilung bei Wänden (links) und Tischen (rechts) 60
3.11 Registrierung mit Wandsegmenten 62
3.12 Strukturelles und funktionales Stuhlmodell 63
3.13 Segmentierte Stühle 64
3.14 Stühle innerhalb einer Szene 66
3.15 Dichtebasierte Segmentierung auf Basis ausgezeichneter Ebenen 70
3.16 Umgebungsbox eines Kaffeebechers 72
3.17 3-D formbasierte Objekterkennung 74
3.18 Erkennung teilverdeckter Objekte [Bre07] 77
3.19 Transformation zwischen Kamera- und Weltkoordinaten 78
3.20 Kalibrierung (Datensätze #277 und #281) 81
3.21 Kalibrierung: Rückprojektionsfehler 82
3.22 2-D/3-D Registrierung (Schreibtischszene) 83
3.23 2-D/3-D Registrierung (Tischszene) 84
3.24 Grenzen des Aufnahmesystems 85
3.25 Farbnormierung durch Approximierung der Referenzbedingung 88
3.26 Farbreduktion auf Standardfarbwerte 89
3.27 Dialogsystem 93

4.1 Einzelschritte hin zur Szenenbeschreibung und einem Mensch-Maschine Dialog 99
4.2 Szenenaufnahme und Segmentierungsergebnis 101
4.3 Externe Szenenaufnahme 101
4.4 Oberflächensegmentierung 101
4.5 Erkannte Wände, Decke, Fußboden, Tischoberfläche und extrahierte Punkte 101
4.6 Dichtebasierte 3-D Segmentierung 101
4.7 3-D Registrierung eines Laptops zur formbasierten Klassifizierung 103
4.8 Umgebungsboxberechnung zur Positionsbestimmung 103
4.9 Punktkorrespondenzen zur Berechnung der 3-D/2-D Transformation . . . 103
4.10 2-D Ausschnittsberechnung 103
4.11 Ergebnis der Farbkalibrierung 103
4.12 Farbbeschreibungsvektor 103
4.13 Sprachliche Referenzierungen der Tischobjekte 105

A.1 Client-Server Modell des *ORCC*-Systems 113
A.2 Bilder zu Gestaltgesetzen 115

A.3 Pintos von Bev Doolittle 116
A.4 Farbnachbild 117
A.5 Farbkontrast - Farbwahrnehmung, Farbtonauflösung 118
A.6 Simultaner Kontrasteffekt 118
A.7 Einfaches Bayessches Netzwerk 126

Tabellenverzeichnis

3.1 Einfach strukturierte Objekte 53
3.2 Merkmale von Raumbegrenzungen und Tisch 57
3.3 Erkennungsverhalten bei Variation der Parameter: Neigungswinkel gegen Flächeninhalt 68

A.1 CPDs für diskreten Kindknoten 129
A.2 CPDs für kontinuierlichen Kindknoten 130

So ist jeder organische Körper eines Lebewesens eine Art göttliche Maschine oder natürlicher Automat, der alle künstlichen Automaten unendlich übertrifft.

(G.W. Leibniz, 1646 – 1716)

Der Schematismus unseres Verstandes, in Ansehung der Erscheinungen und ihrer bloßen Form, ist eine verborgene Kunst in den Tiefen der menschlichen Seele, deren wahre Handgriffe wir der Natur schwerlich jemals abraten und sie unverdeckt vor Augen legen werden.

(I. Kant, 1724 – 1804)

1 Einführung

Auch zwei Jahrhunderte nach den Aussagen von Leibniz und Kant haben diese wenig an Aktualität und Kraft im Zusammenhang mit aktuellen künstlichen kognitiven Systemen verloren: Es ist immer noch eine große Herausforderung, Systeme zu entwickeln, die möglichst automatisch die Umwelt erkennen, sicher in ihr navigieren und sie über ihre ursprünglichen Grenzen hinweg selbst entdecken. Die Offenlegung von Mechanismen, wie wir Menschen von der Sinneswahrnehmung zur Erkenntnis gelangen - beispielsweise bei der räumlichen Einschätzung - ist notwendig, um zukünftig in eine offene Interaktion zwischen Mensch und Maschine treten zu können. Eine wesentliche Komponente eines autonomen Roboters ist die Perzeption, die Kombination aus Aufnahmesystem und Algorithmen zum Erkennen und Klassifizieren von Objekten.

Gegenstand dieser Arbeit ist die Entwicklung eines Systems[1] zur Erstellung einer auf optischen Sensordaten basierenden Szenenbeschreibung, die für eine Mensch-Maschine-Interaktion genutzt werden kann. Das dabei benutzte technische System besitzt einen Laserscanner und eine Farbbildkamera, die beide auf einer Schwenk-Neige-Einheit montiert sind. Dieses System ist in der Lage, 2.5-D Entfernungsaufnahmen und 2-D Farbbildaufnahmen zu erstellen. Bei den betrachteten Szenen handelt es sich um schlicht

[1] Der Name des Systems lautet *ORCC*. Das Akronym steht für *O*bject *R*ecognition using *C*ognitive *C*omputing

gestaltete, aber reale Innenraumszenen. Das System durchläuft hierbei einen vollständigen Erkennungszyklus von der Aufnahme über die Einzelobjekterkennung bis hin zur letztlichen Szeneninterpretation. Wenn möglich wird bei der Erkennung eine funktionale Interpretation angestrebt. Dies soll die Erkennung unabhängiger von speziellen Objektausprägungen machen. Konkret bedeutet diese angestrebte Generalisierung, dass beispielsweise die Erkennung eines Tisches unabhängig von der speziellen Tischform oder dem Tischgestell erfolgt. Dort, wo eine funktionale Bestimmung nicht möglich ist, da die Auflösung des Aufnahmesystems zu gering ist, kommen modellbasierte Verfahren zum Einsatz. Neben der funktionalen Klassifikation der Objekte erfolgt auch eine farbliche Beschreibung, die hier ausschließlich zur Objektidentifikation eingesetzt wird. Die Besonderheit des Systems besteht darin, dass die Segmentierung der unbekannten Objekte auf den 3-D Daten erfolgt, wodurch es möglich ist, auch farb- oder texturähnliche Objekte von ihrem Hintergrund zu lösen. Weiterführend werden die Ergebnisse der erhaltenen Szeneninterpretation genutzt, um innerhalb eines Mensch-Maschine-Dialogs die Objekte zu referenzieren.

Die Übersicht 1.1 zeigt verschiedene Aufgabenbereiche des Gesamtsystems *ORCC*. Es beinhaltet Methoden aus den Bereichen der klassischen Computer Vision, wie beispielsweise Kamerakalibrierung, und führt hin in den Bereich der Cognitive Vision, wie beispielsweise bei der Szeneninterpretation.

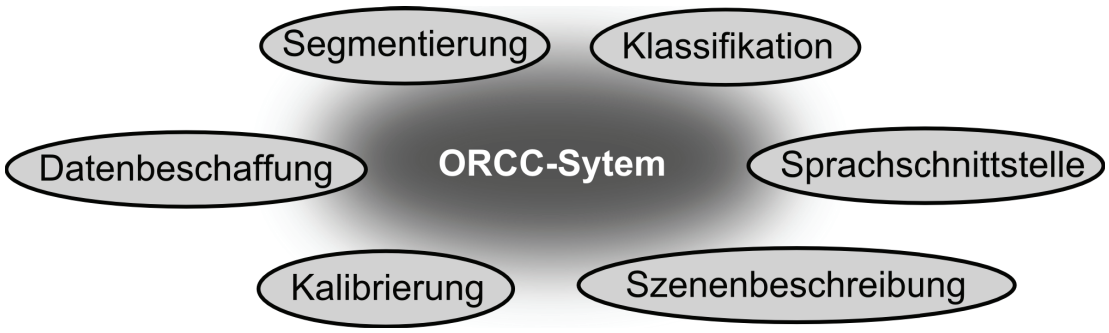


Abbildung 1.1: Entwicklungsumgebung *ORCC*

Diese Arbeit gliedert sich in drei große Bereiche: einen Teil, der die Grundlagen erläutert, einen weiteren Teil, der die einzelnen Module des *ORCC*-Systems beschreibt und einen dritten Teil über die Herstellung der Szenenbeschreibung.

Das Kapitel 2, *Kognitive Aspekte der Objekterkennung*, beinhaltet den Grundlagenteil, in dem die menschlichen Wahrnehmungsaspekte beleuchtet werden, die hier auch

Anwendung im künstlichen Erkennungsprozess finden. Es werden desweiteren die verschiedenen Erkennungsverfahren kategorisiert und Beispiele aufgeführt. Zugleich werden die für diese Arbeit besonders interessanten funktionalitätsbasierten Verfahren diskutiert.

Im Kapitel 3, *Komponenten des ORCC-Systems*, werden die einzelnen Module des *ORCC*-Systems vorgestellt. Eine Übersicht dazu findet sich auf Seite 40. Diese gliedern sich in vier unterschiedlich große Bereiche: die 3-D Verarbeitung, die 2-D Verarbeitung, die Verknüpfung beider Verarbeitungszweige durch ein Kalibriermodul und als letztes die resultierende Szenenbeschreibung und deren Anwendung in einem Mensch-Maschine Dialog. Den mächtigsten Bereich stellt die 3-D Verarbeitung dar, die, ausgehend von der Betrachtung der Scannercharakteristik, über die Oberflächensegmentierung hin zur Erkennung von sowohl einfach strukturierten Objekten, wie Tischen oder Wänden, als auch zur Erkennung von strukturstärkeren Objekten, wie Stühlen, auf Basis funktionaler Merkmale führt. Dabei wird auf die verwendete Modellierung mit Bayesschen Netzen eingegangen. Als weitere Nutzungsmöglichkeit dieser Erkennungsergebnisse wird die Registrierung von Innenraumszenen vorgestellt. Es folgen die Betrachtungen eines dichtebasierten Segmentierungsverfahrens, um beispielsweise Tischobjekte von ihrem Untergrund zu lösen, eines Umgebungsboxberechnungsverfahrens, das zur Positionsabschätzung herangezogen wird, und eines auf den zuvor erhaltenen Segmenten arbeitenden punktbasierten Erkennungsverfahrens. Die 2-D Bildverarbeitung umfasst im ersten Schritt die Texturausschnittsberechnung. Dabei werden die vom Kalibriermodul erlangten Information dazu herangezogen, basierend auf den zuvor erhaltenen 3-D Segmenten, die korresponierenden 2-D Segmente zu berechnen. Um die Abhängigkeit gegenüber bestimmten Lichtverhältnissen zu reduzieren, werden die Farbbildaufnahmen unter Zuhilfenahme eines Kalibriermusters normiert. Anschließend wird eine farbliche Beschreibung der erhaltenen Segmente berechnet. Die Ergebnisse münden in eine geometriebasierte Szenenbeschreibung, die wiederum innerhalb eines Mensch-Maschine-Dialogs über die räumlichen Relationen innerhalb dieser Szene zu Referenzierung von Objekten eingesetzt wird.

Das Kapitel 4, *Anwendung*, beschreibt den kompletten Ablauf des *ORCC*-Systems am Beispiel einer Innenraumszene. Ausgehend von den Rohaufnahmen, werden die Zwischenschritte bis hin zur Szeneninterpretation und dem folgenden Benutzerdialog aufgezeigt.

2 Kognitive Aspekte der Objekterkennung

Die Objekterkennung, als Teil der Szeneninterpretation, bildet eine zentrale Aufgabe des entwickelten *ORCC*-Systems. In diesem Kapitel werden kognitive Aspekte der Objekterkennung motiviert und untersucht. Es werden die unterschiedlichen Arten, wie der menschliche Wahrnehmungsprozess von Objekten interpretiert werden kann, dargestellt. In diesem Zusammenhang wird auf die Modelle über die interne Repräsentation von Objektmodellen Bezug genommen.

Für diese Arbeit von Interesse sind dabei besonders klassische struktur- und funktionsbasierte Verfahren. So zeigt die von Biederman erforschte Theorie der *Recognition by Components* (RBC), bei der *Geons* eine Art Grundbausteine der menschlichen Wahrnehmung darstellen, dass Menschen in der Lage sind, Objekte zu erkennen, auch wenn sie auf ihre Grundkörper reduziert sind [Bie87]. Eine andere Theorie der Wahrnehmung verwendet den Begriff der Affordanzen (engl. *Affordances*), die besagt, dass Menschen in der Lage sind, Objekte durch die mit ihnen mögliche Aktion oder Interaktion zu erkennen. Dieses durch Gibson [Gib82] entwickelte Modell der Wahrnehmung kommt ohne interne Repräsentation aus. Eine eher technische Deutung der Affordanz führt zum Begriff der *Funktionalität*. Diese spielt auch in dem von Minsky [Min85] entwickelten Gesamtkonzept „The Society of Mind“, welches symbolische und konnektionistische Vorstellungen aufnimmt [Sin03], ein Rolle. Daran schließt sich eine Betrachtung von realisierten funktionalitätsbasierten Systemen an.

Mit der Herausforderung, leistungsfähigere Erkennungssysteme zu entwickeln, beschäftigt sich der folgende Abschnitt mit der Betrachtung der sich recht neu herausgebildeten Disziplin *Cognitive Vision*. Diese hat sich zum Ziel gesetzt, mittels typisch kognitiver Ansätze wie Lernfähigkeit unter anderem die Robustheit von Vision-Systemen zu stärken.

Das Kapitel schließt mit der Studie von Systemen, die visuelle Erkennungssysteme mit Dialogsystemen koppeln, was auch das Anwendungsziel dieser Arbeit darstellt.

2.1 Objekterkennung

Der Erkennungsvorgang ist ein komplexer Prozess, auch wenn er für uns Menschen scheinbar einfach und recht zuverlässig funktioniert. Unter Erkennung fassen wir dabei mehrere Mechanismen zusammen wie Identifizierung, Kategorisierung oder Unterscheidung [LB98]. Die enorme Komplexität, die sich hinter dem Erkennungsvorgang bei uns Menschen verbirgt, wird nicht zuletzt dadurch deutlich, dass ein großer Teil des Kortex (*dt.* Hirnrinde) mit der visuellen Wahrnehmung, Erkennung und Kategorisierung beschäftigt ist. Unsere Wahrnehmung selbst registriert dabei nicht die messbare Realität, sondern eine biologisch gewichtete Form dieser. Dies wird deutlich, wenn man die den Gestaltgesetzen zugrunde liegenden Effekte untersucht, die vorrangig die Form betreffen, oder Effekte, die die Farbwahrnehmung betreffen (vgl. das Kapitel A.2, *Optische Wahrnehmung*). Zu welchen Leistungen wir dabei im Stande sind, zeigt beispielsweise die Abb. 2.1. Wir erkennen die dort abgebildeten Stühle unabhängig von der Skalierung bzw. Entfernung oder Blickrichtung, ob teilverdeckt wie am Schreibtisch oder hinter dem Paravent, ob zerbrochen oder in natürlich nicht vorkommender Position wie an der Decke, ob als Silhouette oder Lichtschatten oder unterschiedlich ausgeleuchtet [Gol02]. Doch diese Robustheit unserer Erkennung ist nur ein Aspekt. Wir sind auch in der Lage, Objekte richtig zu kategorisieren, trotz mitunter sehr unterschiedlicher Ausformung.

Abbildung 2.1: Schwierigkeiten bei der Objekterkennung [WB00]

(a) *B55* von Marcel Breuer (1928)

(b) *Slide* von Mattheo Thun (2006)

(c) *Panton Chair* von Verner Panton (1967)

Abbildung 2.2: Designerstühle[2]

Wie verschieden die Ausformungen von Objekten gleicher Kategorie sein können, die wir Menschen aber dennoch klar als der gleichen Kategorie zugeordnet erkennen, zeigt die Abb. 2.2. Diese Abbildung beinhaltet drei verschiedene Stühle aus unterschiedlichen kunsthistorischen Epochen. Während eine individuelle Erkennung, d.h. die Identifizierung anhand sehr individueller Merkmale wie der Farbe, offensichtlich ist, ist im Gegensatz dazu eine Erkennung auf Basisniveau[3] viel schwieriger. Eine gesamtheitliche strukturelle Beschreibung von Formmerkmalen scheitert für uns Menschen an der doch stark unterschiedlichen Gestaltung der einzelnen Stühle. Als Ausweg bietet sich die Ver-

[2] Bildquellen:
Marcel Breuer `http://de.wikipedia.org/wiki/Marcel_Breuer`,
Mattheo Thun `http://www.matteothun.com`,
Verner Panton `http://www.zimmerdahl.se`

[3] Man unterscheidet vier Stufen der Erkennung [Mal06], [RL78]:

Stufe, *dt.* Bezeichnung	**Stufe, *engl.* Bezeichnung**	**Beispiel**
übergeordnetes Niveau	superordinate level	Möbelstück
Basisniveau	basic or entry level	Stuhl
untergeordnetes Niveau	subordinate level	Bürostuhl
individuelle Erkennung		Mein Stuhl

knüpfung von strukturellen Elementen bzw. deren Merkmalen mit ihren funktionalen Eigenschaften an, um auf diese Weise nach dem Vorhandensein der für die Kategorie wesentlichen Teile zu suchen. Auf diesen Ansatz wird im Abschnitt 2.7, *Struktur und Funktionalität*, weiter eingegangen werden.

Untersucht man nun, welche Merkmale neben den bereits erwähnten strukturellen Eigenschaften beim Erkennungsprozess insgesamt eine Rolle spielen, ergibt sich ein differenziertes Bild. In [Sch04] werden vier verschiedene Ansätze zur Erklärung des Objekterkennungsprozesses aufgeführt [LB98], [BET95]:

1. Erkennung durch *invariante Eigenschaften*
   Verwendung von Farbe, Textur, einfachen geometrischen Merkmalen (Parallelität oder Gekrümmtheit extrahierter Kanten)

2. Erkennung durch *strukturelle Beschreibungen*
   Verwendung der räumlichen Anordnung der Teile

3. Erkennung durch *Ausrichtung und Transformation*
   Vergleich des registrierten Stimulus-Bildes mit dem Gedächtnisbild

4. Erkennung durch *multiple Gedächtnistransformation*
   Vergleich des Stimulus-Bildes mit den verschiedenen Ansichten der Gedächtnisbilder

Auf diese einzelnen Ansätze wird in den folgenden Abschnitten separat eingegangen. Dazu werden die jeweiligen technischen Umsetzungen untersucht. Zuvor wird im folgenden Abschnitt der Frage nachgegangen, wie sich die visuelle Repräsentation beim Menschen gestaltet. Ein wichtiger Aspekt dabei ist Kopplung zwischen Betrachter und Objekt, die in ansichtenabhängig bzw. ansichtenspezifisch oder ansichteninvariant unterschieden werden kann.

2.2 Visuelle Repräsentation des Menschen

Der mit dem Erkennungsprozess einhergehenden wichtigen Frage, wie dreidimensionale Objekte in unserem Gehirn modelliert werden[4], widmen sich Bülthoff, Edelman und Tarr in [BET95]. Sie konzentrieren die Unterscheidbarkeit der verschiedenen Verfahren

[4] In [Ede99] werden der Erkennungsprozess und die Formrepräsentation als Zwillingsprobleme bezeichnet.

auf die Frage, ob deren Objektrepräsentationen ansichtenabhängig oder ansichteninvariant sind. Als Mittel dient die Untersuchung von Antwortzeiten und Fehlerraten bei der Erkennung, die dann entweder unabhängig von der Betrachterposition sein oder mit dem Grad der Blickpunktsänderung korrespondieren sollten. Benutzt werden dabei realistisch gerenderte Bilder von CAD-Modellen. Die Autoren postulieren, dass bei einer ansichteninvarianten Repräsentation weder die Erkennungszeit noch deren Güte letztlich vom Blickpunkt des Betrachters abhängen sollte. Umgekehrt sollte bei einer blickpunktabhängigen Repräsentation die Erkennungszeit und die Erkennungsgüte systematisch vom Blickpunkt des Betrachters abhängen.

Beachtet man zusätzlich die Unterscheidbarkeit von 2-D und 3-D Repräsentationen, ergeben sich insgesamt drei potentielle Repräsentationsformen. Bei der *ansichteninvarianten 3-D* Repräsentation wird aus dem wahrgenommenen Bild eine 3-D Repräsentation erstellt, die dann mit denen im Gedächtnis abgeglichen wird[5]. Bei der *ansichtenspezifischen 3-D* Repräsentation wird zum Vergleich die aus dem Bild erhaltene 3-D Repräsentation auf die jeweiligen blickpunktspezifischen 3-D Repräsentationen im Gedächtnis transformiert bzw. „normalisiert“. Bei der *ansichtenspezifischen 2-D* Repräsentation wird das 2-D Bild mit einer Menge von aus verschiedenen Blickwinkeln gespeicherten Gedächtnisbildern verglichen[6], wobei bei beiden ansichtenspezifischen Verfahren die Anzahl von Stützaufnahmen stark variieren kann.

Bei ihren Untersuchungen kommen die Autoren von [BET95] unter anderen zu folgenden Ergebnissen:

- Die ansichtenspezifische Repräsentation spielt bei der menschlichen Wahrnehmung eine große Rolle. Dies gilt insbesondere für die Erkennung auf untergeordnetem Niveau[3], bei der gegenüber der Erkennung auf Basisniveau eine feinere Auflösung

[5] Gegenüber der 3-D ansichteninvarianten Erkennung, wie sie beispielsweise von Marr oder Biederman verfolgt wurden (vgl. Abschnitt 2.4, *Erkennung durch strukturelle Beschreibungen*), sprechen sich die Autoren kritisch aus. Die Ergebnisse bei Messungen der menschlichen Erkennungsleistung variieren stark mit der Perspektive zwischen Objekt und Betrachter und widersprechen somit in gewissem Maß einem ansichteninvarianten Modell.

[6] Einer rein 2-D basierten Erkennung bzw. Repräsentation dreidimensionaler Objekte widersprechen die Autoren, da dies eine quasi kontinuierliche Speicherung verschiedenster Ansichten bzw. Stützaufnahmen zur Voraussetzung machte. Als Deutungsmodell für diese Gruppe von Repräsentationen führen die Autoren deshalb das sogenannte MVPT-Modell (Multiple View plus Transformation) an. Beim MVPT-Modell besteht die Gedächtnisrepräsentation eines Objekts aus einer Reihe ansichtenspezifischer 2-D Bilder, gegenüber denen das Eingangsbild abgeglichen wird. Zu solchen ansichtenspezifischen 2-D Methoden gehören die *Linear Combination of Views* von Ullman und Basri [UB91] oder die *Interpolation by Basis Functions* von Poggio und Edelman [PE90]. Diese beiden Verfahren sind der 4. Gruppe der auf Seite 8 aufgeführten Klassifizierung einzuordnen.

erreicht werden muss. Reine ansichteninvariante Modelle scheinen wenig vorherrschend zu sein.

- 3-D Objekte werden vorrangig durch verschiedene blickpunktspezifische 2-D Ansichten repräsentiert.
- Bei der Erkennung auf untergeordnetem Niveau[3] findet offenbar ein Normalisierungsprozess statt, um Objekte aus einer ungewohnten Perspektive in eine bekannte zu transformieren.

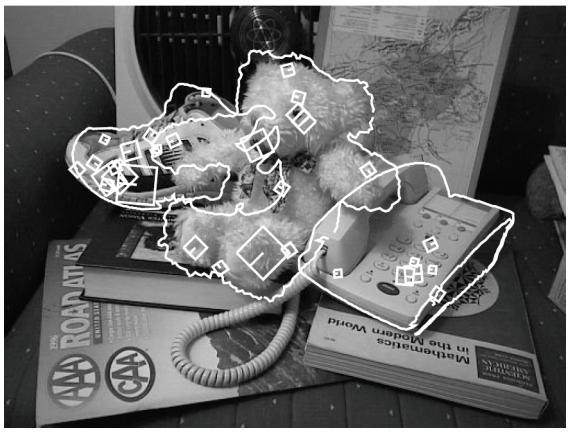
2.3 Erkennung durch invariante Eigenschaften: Verwendung von Farbe, Textur und geometrischen Merkmalen

Die Verwendung von aus dem Bild abgeleiteten Merkmalen stellt bei der Objekterkennung eine sehr bedeutsame Gruppe dar. Der Grund hierfür liegt darin, dass ein wesentlicher Teil der Erkennungsaufgaben nicht die Kategorisierung impliziert, sondern im Sinne einer Identifizierung oder Unterscheidung zu verstehen ist. Diese etwas andere Orientierung der Aufgabe ermöglicht dann die Verwendung von möglichst markanten, häufig lokalen, Merkmalen[7]. Bei den Merkmalen lassen sich verschiedene Kategorien unterscheiden, die den Wirkungsbereich der Merkmale beschreiben [Hüb06]. Dies reicht von lokalen Merkmalen (beispielsweise Kanten) über regionale (beispielsweise Symmetrien) bis hin zu globalen (beispielsweise Histogrammen). In der Regel wird dabei nicht nur ein Merkmal als Kriterium verwendet. Man fasst mehrere Merkmale zusammen und bezeichnet deren Gesamtheit als Merkmalsvektor. Jedoch auch dann bleibt die Schwierigkeit der Auswahl der für das Problem passenden Merkmale. Für diese Aufgabe eignen sich sog. Boosting-Algorithmen [BK99], die aus einem Merkmalspool die jeweils markantesten ihrer Stärke nach identifizieren und gewichten. Diese Boosting-Verfahren scheinen eine recht gute Generalisierungsleistung zu erbringen und damit relativ unanfällig gegenüber Überanpassung zu sein [FS99].

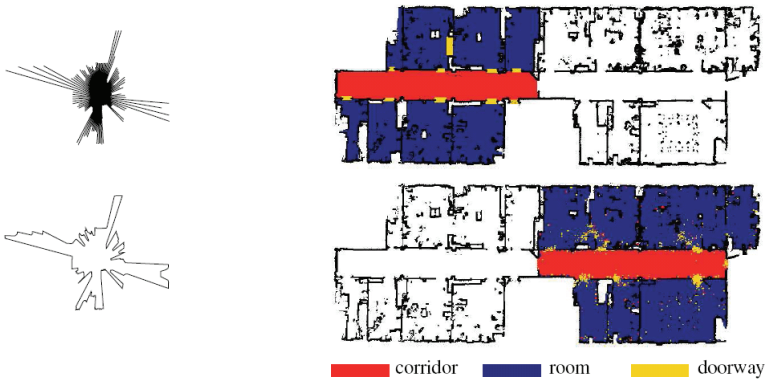
Die Abbildung 2.3a zeigt ein Schaubild zur erfolgreichen Anwendung eines Symmetrieoperators innerhalb einer Büroszene [HZ06]. Im Bild werden die Stellen identifiziert, an denen horizontale und vertikale Spiegelsymmetrien vorhanden sind und deren Ausdehnung zur Beschreibung und Klassifikation von Objekten genutzt wird. Die Abbildung

[7] Einfache Merkmale wären beispielsweise ein bestimmter Farbbereich oder der Objektumfang.

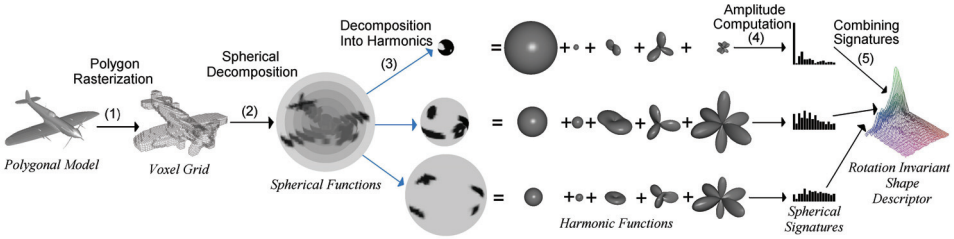
(a) Extrahierte Symmetrie-Merkmale [Hüb06]



(b) Objekterkennung mittels SIFT-Merkmal [Low99]



(c) Objekterkennung mit einer Menge von Merkmalen und AdaBoost [MSB05]
links: Entfernungsdaten oben und abgeleitetes Polygon unten
rechts: Trainingsdaten oben und Testdaten unten



(d) Kugelflächenfunktionsbasierter Form-Deskriptor [FMK+03]

Abbildung 2.3: Merkmalbasierte Erkennung

2.3b zeigt die Anwendung des sehr populären sog. SIFT-Operators (*engl.* scale invariant feature transform) [Low99]. Dieser ist sehr robust gegenüber Skalierung, Rotation

und Beleuchtungsvariationen und wurde in dem Beispiel zur Klassifizierung dreier, teilweise verdeckter, Objekte benutzt. Es konnte gezeigt werden, dass bei Primaten orts-, skalierungs- und beleuchtungsinvariante Merkmale eine Rolle spielen [Tan97], so dass dieser Ansatz auch eine biologische Motivation besitzt.

Die Abbildung 2.3c zeigt die Anwendung des Boosting-Verfahrens AdaBoost bei einer Ortsklassifizierungsaufgabe im Innenraumbereich [MSB05]. Das obere Teilbild zeigt die Trainingsdaten, bei denen die unterschiedlichen Orte entsprechend ihrer Klassenzugehörigkeit (Flur, Raum oder Türöffnung) farblich markiert sind. Als Merkmale stehen hier eine ganze Reihe von Kandidaten zur Verfügung, die in zwei Gruppen unterschieden werden können. Solche, die auf den rohen Entfernungsdaten arbeiten, und solche, die auf dem durch die Entfernungsdaten gebildeten Polygon arbeiten. Zu den ersten zählen beispielsweise die durchschnittliche Differenz der Längen zweier aufeinander folgenden Strahlen oder der euklidische Abstand zwischen den zwei Punkten, die den zwei kleinsten lokalen Entfernungsminima entsprechen. Zur Gruppe der polygonbasierten Merkmale gehören beispielsweise die Fläche des Polygons oder der Durchmesser des Polygons. Die Autoren testeten zwei Versionen des Boosting-Algorithmus. Zum einen eine Multi-Klassen-Version und zum anderen eine sequentielle Binär-Klassen-Version, bei der die Klassenzugehörigkeitsberechnung in einer bestimmten Reihenfolge berechnet wird, und sobald eine Zuordnung erfolgt ist, abgebrochen wird. Für letzteres Verfahren wurden, abhängig von der gewählten Klassifizierungsreihenfolge, Erkennungsraten von knapp 94% erreicht. Eine Anwendung des Boosting auf 2-D Daten findet sich in [VJ01]. Auch hier dienen relativ schwache Merkmale (*engl.* weak features), die man als Haar-Features bezeichnen kann, als Ausgangspunkt, um sehr erfolgreich Gesichter zu erkennen. Diese Merkmale bestehen beispielsweise aus zwei benachbarten Rechtecken, bei denen die Summen der Grauwerte voneinander subtrahiert werden. Trotz dieser Einfachheit gelingt es, sehr robust Gesichter zu erkennen.

Die Abbildung 2.3d zeigt das auch auf großen Datenmengen sehr erfolgreiche Erkennungssystem von Funkhouser [FMK+03]. Bei diesem Verfahren wird das 3-D Modell in eine Summe von formbeschreibenden, auf konzentrischen Sphären arbeitenden Funktionen zerlegt. Die Verwendung von Kugelflächenfunktionen erlaubt es, eine orientierungsunabhängige Beschreibung des 3-D Modells zu erhalten, aus der sich hier allerdings nicht mehr das Ausgangsmodell rekonstruieren lässt. Im ersten Schritt wird das Polygonnetz des Modells zentriert, skaliert, in einer festen Auflösung gerastert und letztlich binari-

siert. Hiernach wird das erhaltene Voxelmodell[8] in konzentrische Sphären zerlegt, bei denen jede Sphäre einer binären reelwertigen Funktion entspricht. Diese Sphären lassen sich als quadratintegrable Funktionen, als eine Reihe von Kugelflächenfunktionen (*engl.* Spherical Harmonics), entwickeln[9]. Diese Reihenelement-Funktionen bilden den Ausgangspunkt des Deskriptors, indem ihre L_2-Norm zu einem Formbeschreibungsvektor (zunächst für die jeweilige Schale) zusammengefasst wird. Somit ergibt sich für jede Sphäre ein Satz von orientierungsunabhängigen Betragswerten, die, zusammengefasst für alle Sphären, den letztlichen Beschreibungsvektor liefern. Die Ähnlichkeitsberechnung zweier Beschreibungsvektoren wird durch Berechnung des euklidischen Abstands realisiert. Dieses 3-D basierte Erkennungsverfahren ist Teil der webbasierten 3-D Suchmaschine[10] der Princeton Shape Retrieval and Analysis Group, die auch 2-D basierte Skizzen zur Eingabe zulässt.

Ein zu dem von Funkhouser verwandtes Verfahren ist das von Novotni und Klein [NK03]. Hier werden die Modelle allerdings nicht in Sphären diskretisiert, sondern Radialpolynome verwendet, um die Kugelflächenfunktionen dreidimensional zu modulieren. Als Funktionensystem werden die sog. 3-D Zernike-Funktionen, die auf den Kugelflächenfunktionen aufgebaut sind, zugrunde gelegt[11]. Der sog. 3-D Zernike-Deskriptor ergibt sich als Norm des aus den einzelnen Momenten (den Elementen der Reihenentwicklung) aufgebauten Vektors. Die Abbildung 2.4 zeigt, wie abhängig von der Anzahl der verwendeten

[8] Ein Voxel ist das auf drei Dimensionen ausgedehnte Äquivalent eines Pixels.

[9] Für jede über der Einheitskugel quadratintegrable Funktion f, d.h. $\int |f(\theta,\phi)|^2 \mathrm{d}\Omega < \infty$ gilt, dass die Reihe $f(\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} Y_{l,m}(\theta,\phi) C_{l,m}$ im Mittel konvergiert [Gra69], [Eck02]. Die Funktionen $Y_{l,m}$ sind die sog. Kugel(flächen)funktionen, die Koeffizienten $C_{l,m}$ lassen aus der komplex konjugierten Kugelfunktion berechnen: $C_{l,m} = \int Y^*_{l,m}(\theta,\phi) f(\theta,\phi) \mathrm{d}\Omega$. Man nennt diese Projektionen der Funktion f auf die komplex konjugierten Funktionen des Funktionensystems, nach dem entwickelt wird, die sog. Momente.
Die Kugelfunktionen sind insbesondere für die Quantenphysik von großer Bedeutung, da sie Lösungen der stationären drei-dimensionalen Schrödingergleichung darstellen.

[10] Auswahl aus einer Liste von 3D-Suchmaschinen [Min04]:
- Princeton Shape Retrieval and Analysis Group, 3D Model Search Engine:
  `http://shape.cs.princeton.edu/search.html`
- Ogden IV:
  `http://ship.nime.ac.jp/~motofumi/Ogden/`
- 3D Object Retrieval System from National Taiwan University:
  `http://3d.csie.ntu.edu.tw/~dynamic/cgi-bin/DatabaseII_v1.8/index.html`

[11] Entsprechend der Entwicklung einer quadratintegrablen Funktion f mit Kugelflächenfunktionen kann eine solche Funktion mit Zernike-Funktionen angenähert werden [NK04]: $\hat{f}(\mathbf{x}) = \sum_{n=0}^{N} \sum_{l=0}^{n} \sum_{m=-l}^{l} \Omega^m_{nl} Z^m_{nl}(\mathbf{x})$, $(n-l)$ gerade. Die 3-D Zernike-Funktionen Z^m_{nl} sind von der Form: $Z^m_{nl}(\mathbf{x}) = R_{nl}(r) Y^m_l(\vartheta,\phi)$, die R_{nl} sind die Radialpolynome und die Ω^m_{nl} sind die durch f definierten Momente.

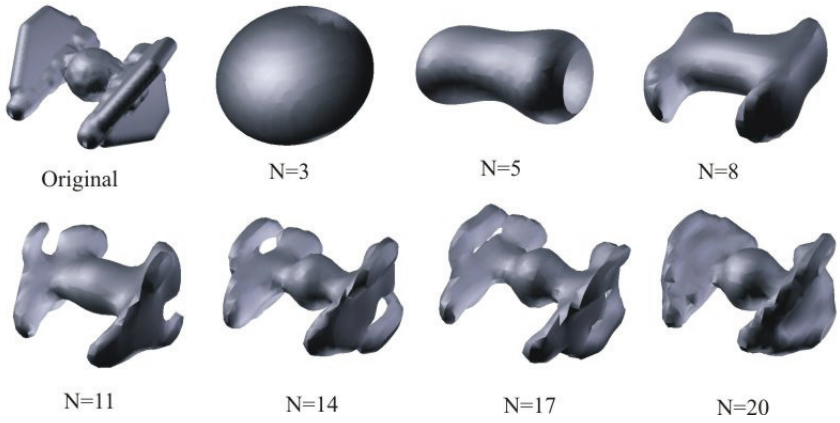


Abbildung 2.4: Rekonstruktion eines Raumschiffes in Abhängigkeit der Anzahl der verwendeten Zernike-Momente [NK04]

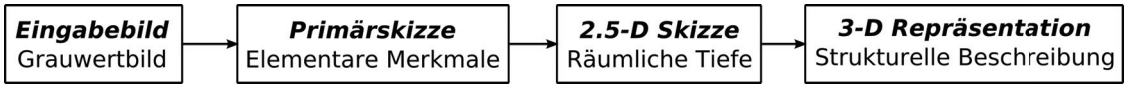


Abbildung 2.5: Strukturbasierte Erkennung nach Marr [Mar82]

Momente das Originalmodell (ein Raumschiff) dargestellt werden kann.

2.4 Erkennung durch strukturelle Beschreibungen: Verwendung der räumlichen Anordnung der Teile

Die Abb. 2.5 zeigt den Ablauf des Systems zur Objekterkennung bei David Marr [Mar82], [Sch04], [Gol02]. Dieser stellt einen der sehr frühen und einen als klassisch zu bezeichnenden Ansatz dar, bei dem eine 3-dimensionale, objektzentrierte Struktur des Objekts die Basis für den Objektvergleich darstellt. Im ersten Schritt werden Bildmerkmale wie Zero-Crossings[12], also Grauwertkanten, dazu genutzt, um elementare Bildmerkmale wie Balken (Objekte mit hinreichend parallelen Begrenzungen) und Blobs[13] (Objekte mit

[12] Mit Zero-Crossings wird eine spezielle Form der Kantenextraktion bezeichnet (vgl. [Jäh02]). Bei sogenannten Ableitungsfiltern wird die Ableitung dazu verwendet die Kante zu bestimmen, wobei die erste Ableitung an diesen Stellen ein Extremum und die zweite Ableitung einen Nulldurchgang (*engl.* Zero-Crossing) besitzt. Marr benutzte den *Laplacian-of-a-Gaussian*-Operator (LOG), $\nabla^2 G$, wobei die vorgeschobene Gaussfunktion das Bild glättet.

[13] Abweichend zu der Verwendung bei Marr steht die Abkürzung für *engl.* binary large object und damit ein von der Form unabhängiges Segment.

doppelt hinreichend parallelen Begrenzungen), also Segmente, mit Eigenschaften wie Position, Länge, Breite, Orientierung und Kontrast zu erhalten. Aus diesen elementaren Merkmalen wird dann das sog. 2.5-D[14] Bild berechnet. Als Methoden kommen für Marr unter anderen folgende Verfahren in Betracht:

- *Stereosehen*
  Beim Stereosehen werden zunächst über korrespondierende Punkte die Disparitäten[15] aus den verschiedenen Aufnahmepositionen bestimmt, aus denen dann trigonometrisch der Abstand bestimmt werden kann.

- *Richtungsselektivität*
  Hierbei wird zunächst die lokale Bewegungsrichtung bestimmt, um daraus den Vordergrund vom Hintergrund zu trennen.

- *Struktur aus Scheinbewegung*
  Die Korrespondenzen zwischen mehreren aufeinanderfolgenden Bildern ermöglichen die jeweiligen Verschiebungen zwischen den elementaren Merkmalen zu detektieren. Hieraus wird dann die dreidimensionale Struktur bestimmt.

- *Optischer Fluß*
  Bei diesem Ansatz wird die Tiefeninformation aus dem durch die Bewegung des Beobachters entstehenden retinalen Geschwindigkeitsfeld gewonnen.

- *Oberflächenausrichtung aus Oberflächenkontur oder Textur*
  Von den Konturen einer glatten Fläche oder der Oberflächentextur wird auf deren 3-dimensionale Form geschlossen.

- *Gestalt aus Schattierung*
  Es wird bei bekannter Situation der Oberflächenreflexions- und Beleuchtungseigenschaften aus den Schattierungen auf die Oberflächenform geschlossen.

Ergebnis dieser zweiten Phase ist eine segmentierte Oberfläche, deren Orientierung und Tiefe berechnet wurden. Ausgehend von diesen Segmenten wird ein modulares 3-D Modell erstellt, bei dem zunächst die kanonische Achse des Modells berechnet wird,

[14] Ein 2.5-D Bild bezeichnet eine Aufnahme, bei der kein vollständiges 3-D Modell vorhanden ist, sondern nur eine Art Höhenprofil, dessen Vollständigkeit zudem von der gewählten Perspektive abhängig ist.

[15] Disparitäten beschreiben den Versatz in den Aufnahmen eines aus zwei unterschiedlichen Positionen aufgenommenen realen Punktes.

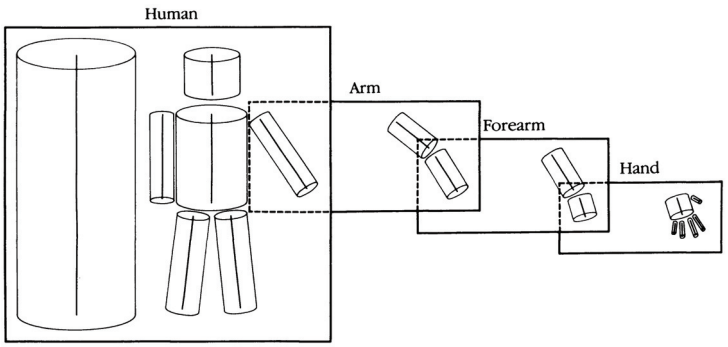


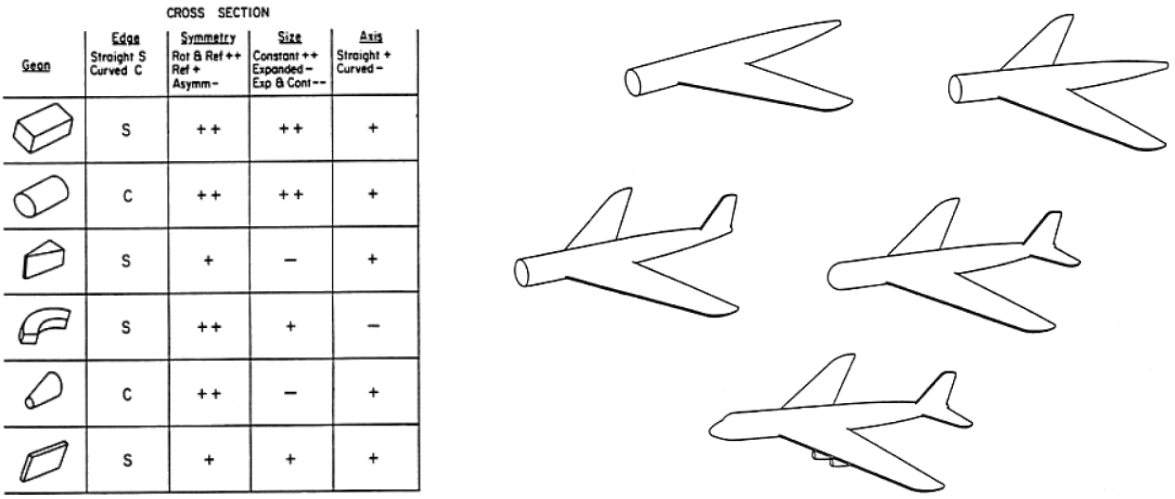
Abbildung 2.6: Hierarchisch organisiertes Modell nach Marr [Mar82]

vgl. Abb. 2.6. Dann wird das Objekt hierarchisch in Teilobjekte zerlegt, die allesamt aus einfachen Objekten wie Zylindern aufgebaut sind, und ihre Anordnung berechnet. Diese Objektbeschreibung wird dann mit den Gedächtnismodellen verglichen. Die Idee, ein Objektmodell durch die Anordnung einfacher geometrischer Körper zu realisieren, wurde von Biederman weiterentwickelt.

Die von Biederman eingeführte Theorie der *Recognition by Components* bzw. *Geon Structural Description* geht davon aus, dass jeder beliebige Körper aus einer Reihe unterscheidbarer Grundkörper, den Geonen[16], aufgebaut werden kann und deren Anordnung von relativ unterschiedlichen Positionen aus erkennbar sind [Bie87]. Diese Methode zählt damit ebenfalls zur Gruppe der 3-D strukturbasierten ansichteninvarianten Erkennungsverfahren.

Biederman sieht im Objekterkennungsprozess bzw. im Aufbau von Objekten eine Analogie zum Spracherkennungsprozess bzw. zum Aufbau von Sprache. Während bei der Sprache die Worte bzw. Silben aus Phonemen aufgebaut sind, entsprechen bei der visuellen Wahrnehmung die Teile eines Objekts den Geonen. Die Geone sind dadurch charakterisiert, dass sie konvex und volumetrisch sind. Die Trennungsstellen zwischen den Geonen eines Objekts ergeben sich aus Stellen deutlicher Konkavität. Im Gegensatz zur sequentiellen Anordnung der Phoneme bei der Sprache ist die Anordnung bei der Kombination der Geone natürlich vielfältiger. Entsprechend enthalten die Relationen zwischen den Geonen Angaben zu den relativen Größen, der Orientierung und den Verbindungspunkten der jeweiligen Partner. Aufbauend auf diesen Grundkörpern, ergeben

[16] Geon steht für *geometrical ion*, gemeint ist damit eine Grundform.



	CROSS SECTION			
Geon	Edge Straight S Curved C	Symmetry Rot & Ref ++ Ref + Asymm−	Size Constant ++ Expanded− Exp & Cont−−	Axis Straight + Curved−
	S	++	++	+
	C	++	++	+
	S	+	−	+
	S	++	+	−
	C	++	−	+
	S	+	+	+

Abbildung 2.7: Geon

sich Formbeschreibungen, die wir Menschen noch eindeutig zuordnen können. In der Abb. 2.7 ist ein Flugzeug zu sehen, das mit bis zu 9 Geonen dargestellt ist.

Ausgangspunkt für die Erkennung der volumetrischen Körper sind die detektierten Kanten im zweidimensionalen Bild. Es wird davon ausgegangen, dass diese Kanten des zweidimensionalen Bildes den volumetrischen Körper hinreichend robust beschreiben. Der sich anschließende Segmentierungsschritt beinhaltet zwei Kriterien. Erstes Segmentierungskriterium sind Orte starker Konkavität, die insbesondere bei Spitzen vorliegen. Konkavitäten finden sich paarweise immer an Stellen, an denen zwei konvexe Körper verbunden werden. Die zweite Gruppe von Segmentierungskriterien sind solche, die mit der Achsenbestimmung der Komponenten einhergehen, wie Maximierung der Achsenlänge oder Konstanz der Querschnittsgröße.

Basis für die entwickelten Geone sind sogenannte nicht-zufällige (*engl.* non-accidental) Merkmale, da sie im Allgemeinen nicht durch eine zufällige Anordnung von Blickpunkt und Bildmerkmalen bei der Abbildung auf die Netzhaut erreicht werden können. Dazu gehören Krümmung, Kollinearität, Symmetrie, Parallelität und Kotermination. Biederman entwirft daraus eine Menge von 36 Grundkörpern, die aus 4 Parametern beschrieben werden (6 Geone sind in Abb. 2.7). Drei der Parameter beziehen sich auf den Querschnitt und beschreiben dessen Kante (gekrümmt ↔ gerade), Symmetrieeigenschaft (spiegelsymmetrisch ↔ rotationssymmetrisch), Größe (konstant ↔ expandierend). Der vierte Parameter beschreibt die Achse, die gerade oder gekrümmt sein kann.

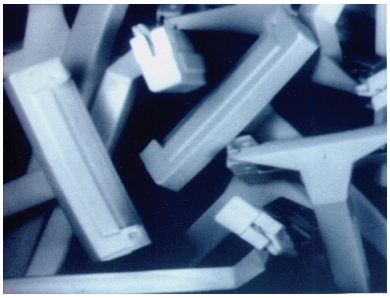
Wie in Abschnitt 2.2, *Visuelle Repräsentation des Menschen*, beschrieben, scheinen diese Verfahren aus wahrnehmungspsychologischer Sicht eher untergeordnet. Was sie für die Anwendung so schwer handhabbar macht, sind die Robustheitsunzulänglichkeiten bei der Kantenerkennung, auf der das ganze Modell basiert. Zu störend sind die Einflüsse beispielsweise durch Reflexionen, sich ändernde Lichtverhältnisse oder die technisch-physikalischen Grenzen, hervorgerufen beispielsweise durch das sog. Photonenrauschen.

2.5 Erkennung durch Ausrichtung und Transformation: Vergleich des registrierten Stimulus-Bildes mit dem Gedächtnisbild

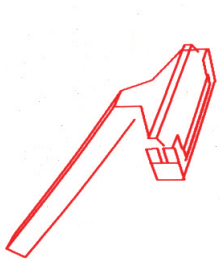
Das von Lowe [Low87] entwickelte System SCERPO aus den späten 80er Jahren ist in der Lage, festgelegte dreidimensionale Objekte (speziell Einwegrasierer, vgl. Abb. 2.8) aus beliebiger Blickrichtung in Grauwertbildern zu erkennen. Hierzu wird nach Strukturen gesucht, die signifikante Merkmale im zwei-dimensionalen Bild ergeben und ein entsprechendes Pendant im 3-dimensionalen Modell haben, wie beispielsweise parallele Kanten. In dieser Veröffentlichung weist Lowe darauf hin, dass der Einfluss von Dreidimensionalität beim menschlichen Erkennungsprozess zuweilen überschätzt werde (vgl. Abschnitt 2.2, Visuelle Repräsentation des Menschen).

Im Einzelnen werden bei dem Verfahren von Lowe folgende Schritte durchlaufen: Im ersten Schritt findet eine Kantenerkennung statt, wobei auch wie bei Marr der Laplace-of-Gaussian-Filter zum Einsatz kommt. Im nächsten Schritt werden die Kanten segmentiert, d.h. die darin enthaltenen signifikanten Linien berechnet. Auf diesen Liniensegmenten wird hiernach nach Relationen wie Kollinearität, Endpunktnähe oder Parallelität gesucht. Aufbauend wird danach nach komplexeren Strukturen gesucht. Diese werden dann dahingehend überprüft, ob sie durch das dreidimensionale Modell unter entsprechender, d.h. zu berechnender Perspektive, hervorgerufen werden können. „Das Erkennungssystem von Lowe kombiniert also zwei Grundsätze: Erkennung mittels Transformation und Ausrichtung sowie Erkennung mittels invarianter Eigenschaften“ [Sch04].

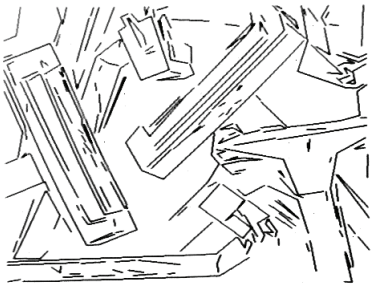
Die Verwendung von 3-D Modellen eignet sich nicht nur für die Erkennung, sondern auch für die Verfolgung von Objekten [GG00]. In Abbildung 2.9 ist eine Szene zu sehen,



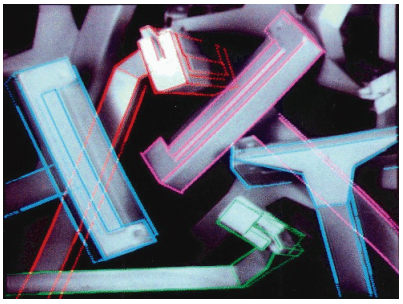
(a) Grauwertbild von Einwegrasierern



(b) 3-D Modell eines Einwegrasierers



(c) Extrahierte Kantensegmente



(d) Erfolgreich gematchte 3-D Modelle

Abbildung 2.8: Modellbasierte Erkennung durch Ausrichtung [Low87]

bei der sich ein Auto bewegt und ebenfalls die Kameraposition verändert wird. Dem System gelingt es durchgehend (bei einer Sequenz von 161 Bildern), das Auto korrekt zu verfolgen, was durch die Überlagerung des Grauwertbildes mit dem Drahtgitter des Automodells angezeigt wird.

2.6 Erkennung durch multiple Gedächtnistransformation: Vergleich des Stimulus-Bildes mit den verschiedenen Ansichten der Gedächtnisbilder

Ein Vertreter dieser Art von Objekterkennungsverfahren ist der sog. Light Field Descriptor [ZL02], [CTSO03]. Nach Vergleichsuntersuchungen von [TV04] ist der LFD der zur Zeit erfolgreichste Formen-Deskriptor. Die recht einfache, aber erfolgreiche Idee ist, dass zwei 3-D Modelle sich genau dann ähnlich sind, wenn alle ihre Projektionen aus

Abbildung 2.9: Modellbasierte Verfolgung (*engl.* Tracking) eines fahrenden Autos [GG00]

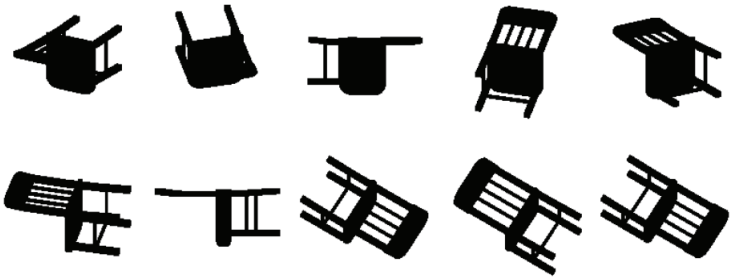


Abbildung 2.10: Light Field Descriptor: Silhouetten eines Stuhls [CTSO03]

beliebigen Blickwinkeln gleich sind (vgl. Abb. 2.10). Zur Formbeschreibung werden sog. Fourier-Deskriptoren und Zernike-Deskriptoren verwendet. Der Fourier-Deskriptor verwendet die Eigenschaft, dass sich periodische Funktionen f mit der Periode T als Summe von Kosinus und Sinus-Funktionen annähern lassen [BS89][17]. Durch weitere Operationen kann neben der Translationsinvarianz auch die Rotations-, Skaleninvarianz der erhaltenen Formparameter erreicht werden [ZL02]. Der Zernike-Deskriptor besteht entsprechend aus den Momenten, die von der Bildfunktion und den zugrunde liegenden Zernike-Polynomen aufgebaut sind. Auch die Zernike-Polynome bilden wiederum ein vollständiges orthogonales Funktionensystem, was bedeutet, dass die Bildfunktion nach den Zernike-Polynomen entwickelt werden können[18]. Da die Zernikefunktionen auf der

[17] In komplexer Schreibweise bedeutet die Fourierentwicklung, dass die Reihe $\hat{f}(x) = \sum_{n=-N}^{n=N} a_n \exp^{\imath knx}$ quadratisch im Mittel gegen die Funktion f konvergiert. Die Fourierkoeffzienten a_n berechnen sich zu $a_n = \frac{1}{N} \sum_{t=0}^{N-1} r(t) \exp^{\left(\frac{-\imath 2\pi nt}{N}\right)}$, $n = 0, 1, \ldots, N-1$ und N die Anzahl der Randpunkte der untersuchten Form. Die Funktion $r(t)$ entspricht hier dem Abstand der Konturpunkte vom Zentrum.

[18] Mit den Zernike-Polynomen $V_{nm}(x, y) = R_{nm}(\rho, \theta) \exp^{\imath m\theta}$, wobei für m und n gilt, dass $n - |m|$ gerade und $|m| < n$, ergeben sich die Zernike-Momente als $A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x, y) V^*_{nm}(x, y)$, wobei $(x, y) \leq 1$.

Einheitssphäre operieren, müssen die Formen entsprechend skaliert werden, wobei das Zentrum mit dem Formenschwerpunkt gleichgesetzt wird. Dadurch wird eine Skalen- und Translationsinvarianz erreicht. Die Rotationsinvarianz des Deskriptors wird erreicht, indem nur der Betrag der Momente betrachtet wird. Insgesamt stehen damit zwei verschiedene Deskriptoren zur Verfügung, wobei der Fourier-Deskriptor nur die Randform und der Zernike-Deskriptor sowohl Kontur als auch lokale Formen beschreibt. Zum Einsatz kommen 35 Parameter, die auf den Zernike-Momenten aufgebaut sind, und 10 Parameter, die auf den Fourier-Momenten aufgebaut sind.

Neben der einfachen Anwendung des LFDs gibt es auch Versuche und Untersuchungen, den LFD mit anderen Form-Deskriptoren zu verknüpfen, um damit die Gesamtleistung nochmals zu verbessern [OH06].

2.7 Struktur und Funktionalität

Die Theorie der Affordanzen (*engl. affordances*) besagt, dass Objekte gewisse Funktionen besitzen, die ihre Aktion mit dem Menschen bestimmen [Gib82]. Ein Stuhl beispielsweise ermöglicht das *Darauf-Sitzen*, eine Türklinke *das Drücken* derselbigen und damit letztlich das Öffnen oder Schließen der Tür. Diese Affordanzen sind in dieser Theorie wesentlich für die Wahrnehmung des Objekts. Dem psychologisch beheimateten Begriff der Affordanzen steht der technisch orientierte Begriff der Funktionalität gegenüber [Mal06]. Dabei verspricht man sich durch die vorrangige Repräsentation einer Objektkategorie über die Funktionalität eine bessere Generalisierbarkeit[19] [SB91].

Die visuelle Wahrnehmung[20] nach Gibson ist ein direkter Prozess, bei dem keine internen Repräsentationen oder Symbole notwendig sind. Die Wahrnehmung entspricht damit einer Art unbewussten Inferenz. Die Fähigkeit zu sehen entspricht in dieser Theorie, „auf bestimmte invariante Eigenschaften in der optischen Anordnung eingestellt zu

[19] Zum Problem der Generalisierbarkeit durch strukturelle Beschreibungen am Beispiel von Stühlen siehe Abb. 2.2.

[20] Die Begriffe Sinnesempfindung (*engl.* sensation) und Wahrnehmung (*engl.* perception) stehen nach [GZ05] in folgendem Zusammenhang: Wahrnehmung bedeutet die Erfassung von Objekten als übergeordneter Prozess und umfasst die Schritte von der sinnlichen Empfindung, über mögliche mentale Zwischenschritte bis hin zur Identifizierung und Klassifizierung des Objekts. Die Sinnesempfindung als erster Teilschritt der Wahrnehmung entspricht der sensorischen Rezeption von internen und externen Stimulationen.

sein und bei entsprechender Anregung bereit zu sein, in bestimmter Weise mit ihr zu handeln“[21] [Ede92]. Damit steht dieser Ansatz anderen Theorien entgegen, die, wie beispielsweise der klassische Ansatz von Marr [Mar82], zunächst eine interne Repräsentation erstellen, auf der dann erst der Erkennungsprozess durchgeführt wird.

Eine historische Einordnung der Arbeiten von Gibson liefern Bickhard und Richie in [BR83]. Ausgangspunkt der Arbeiten Gibsons ist eine Situation, in der klassische, sog. sinnesempfindungsbasierte, Ansätze[22] den sogenannten gestaltbasierten Ansätzen gegenüberstehen. Beim sinnesempfindungsbasierten Ansatz werden die Retinabilder oder gewisse darauf beruhende Basiseindrücke (Farbsegmente, Linien, etc.) in einem mentalen Prozess verarbeitet. Durch Vergleiche mit Gedächtnissinnesempfindungen, Folgerungen über gewisse Reize innerhalb der Sinnesempfindung und deren Beurteilung der Ursache des äußeren Stimulus wird das gesehene Objekte erkannt bzw. wahrgenommen. Der gestaltbasierte Ansatz hingegen geht davon aus, dass die ”Wahrnehmung ein Prozess relativ spontaner sensorischer Organisation ist”[23,24] [Gib50], bei dem die Erkennung nicht auf die Inferenz einzelner Komponenten zurückführbar ist, sondern bei dem der Gegenstand eher als ein Ganzes erkannt wird[25].

Gibson geht bei seiner Theorie davon aus, dass, wenn die Reize alle für die Wahrnehmung notwendigen Informationen enthalten, sowohl eine spontane sensorische Organisation wie bei den gestaltbasierten Modellen als auch eine mentale Verarbeitung wie bei den sinnesempfindungsbasierten Modellen überflüssig sei[26]. Er lehnt auch ab, den Betrachter als passives Wesen anzusehen, wie es bei den sinnesempfindungsbasierten Modellen der Fall ist. Gerade der sich normalerweise bewegende Betrachter erhält durch

[21] *engl.* „to be attuned to certain invariant qualities present in the optic array, and, potentially, to become disposed to act in certain ways, given the appropriate stimulation“

[22] *engl.* sensation or retinal-image-based theories

[23] *engl.* „... sensory organization appears as a primary fact which arises from the elementary dynamics of the nervous system.“ (W. Köhler, Gestalt Psychology) [Jan96]

[24] *engl.* The process of perception „was one of a relatively spontaneous sensory organization“. (Gibson) [BR83]

[25] Diese Eigenschaft wird als Emergenz bezeichnet und entspricht der Grundregel „Das Ganze ist mehr als die Summe seiner Teile“. Hinweise zu den Gestaltfaktoren bzw. Gestaltgesetzen finden sich im Anhang in Kapitel A.2, *Optische Wahrnehmung*. Dort ist in der Abb. A.2 eine Darstellung zu sehen, die nicht mit Wahrnehmungsmodellen erklärt werden kann, bei denen sich die Wahrnehmung aus der Verknüpfung einzelner Sinnesempfindungen zusammensetzt. Denn das dort sichtbare Dreieck existiert physikalisch nicht.

[26] Dies bedeutet auch, dass die Trennung zwischen Sinnesempfindung (*engl.* sensation) und Wahrnehmung (*engl.* perception) für Gibson so nicht bestehen bleiben kann.

die Bewegungsparallaxe einen sehr nützlichen Tiefeneindruck. Der Erkennungsprozess kann also nicht unabhängig und isoliert von der Umwelt betrachtet werden. Ein weiteres Argument gegen einen mentalen Prozess wie bei den sinnesempfindungsbasierten Modellen ist das sogenannte Homunculus-Problem: „... begreift man das Auge als Kamera am Ende eines Nervenkabels, das das Bild zum Gehirn überträgt, dann muss es einen kleinen Mann geben, den sog. Homunculus, der im Gehirn sitzt und auf das physiologische Bild schaut. Dieser kleine Mann bräuchte, um das Bild zu sehen, natürlich ein Auge mit einem kleinen Netzhautbild, das mit einem kleinen Gehirn verbunden ist, und entsprechend hat man mit dieser Theorie keine wirkliche Erklärung gefunden.“[27,28] Das Modell, das Gibson entwickelte, entsprach zunächst hingegen einem umweltdirektkodierenden[29,30] Modell. Ausgehend von einem aktiven Betrachter, ergibt sich aber das Problem, dass sich das Netzhautbild zu stark verändert, als dass es zu den konstanten Wahrnehmungen passen würde. Gibson reagierte darauf, indem er eine andere Quelle als das Netzhautbild für die Wahrnehmung postulierte: „In meinem Buch, The Perception of the Visual World (1950), betrachtete ich das Netzhautbild als Stimulus für ein Auge. In diesem Buch gehe ich davon aus, dass es nur der Stimulus für die Netzhaut ist und dass das Umgebungslicht[31] der Stimulus für das visuelle System ist.“[32] Die Erkennung entspricht damit jetzt dem Finden von örtlichen und zeitlichen Mustern im Umgebungslicht und wird als interaktiver Prozess verstanden[33]. Es ist dabei das Muster dieser Interaktion, welches die Klassifizierung bestimmt. Es bleibt aber die Frage, wie

[27] *engl.* „I call this the *little man in the brain* theory of the retinal image, which conceives the eye as a camera at the end of a nerve cable that transmits the image to the brain. Then there has to be a little man, a homunculus, seated in the brain who looks at the physiological image. The little man would have to have an eye to see it with, of course, a little eye with a little retinal image connected to a little brain, and so we have explained nothing by this theory.“ (Gibson) [BR83]

[28] Ein von Marvin Minsky in seiner Society of Minds beschriebener Agent bzw. Prozess (der damit einen Teilprozess des menschlichen Verstandes darstellt) sind die sog. „B-Brains“. Danach kann das Gehirn als eine Kaskade von Teilgehirnen angesehen werden, wobei jedes Teilgehirn nur Zugriff auf das vorgeschaltete Teilgehirn hat.

[29] *engl.* ecological direct encoding

[30] *engl.* „When I assert that the perception of the environment is direct, I mean that it is not mediated by *retinal* pictures, *neural* pictures, or *mental* pictures. Direct perception is the activity of getting information from the ambient array of light. I call this a process of information pickup.“ (Gibson) [BR83]

[31] Umgebungslicht (*engl.* ambient light) entspricht „Licht, das an einem Punkt im umgebenden Medium ankommt. Nur dieses Licht bestimmt die Lichtreizenergie für einen dort befindlichen Beobachter. Es ist somit *Umgebungslicht*, bezogen auf einen Beobachtungsort.“ [Gib82]

[32] *engl.* „In my book, The Perception of the Visual World (1950), I took the retinal image to be the stimulus for the eye. In this book I will assume that it is only the stimulus for a retina and that *ambient light* is the stimulus for the visual system.“(Gibson) [BR83]

[33] *engl.* „there is a loop from response to stimulus to response again“(Gibson) [BR83]

die Wahrnehmung von nützlichen Dingen erreicht wird. Die Antwort sieht Gibson wiederum in den Interaktionsmustern beim Erkennen der Muster im Umgebungslicht. Sie sind funktionale Gradmesser und zeigen, welche Aktionen mit dem dahinterliegenden Objekt möglich sind. Dies führt zum Konzept der Affordanzen: „Die Affordanz eines beliebigen Gegenstands ist eine spezifische Vereinigung von Eigenschaften seines Materials und seiner Oberflächen in Bezug auf ein Lebewesen. Affordanzen sind Dinge, welche die Umgebung einem Lebewesen anbietet, die sie bereitstellt oder mit dem sie es ausstattet, entweder zum Guten oder zum Bösen.“[34]

Auf die Notwendigkeit des Zusammenhangs von Funktionalität und Struktur geht Minsky in „Society of Mind“ [Min85] ein. Sein Modell menschlichen Verstandes geht davon aus, dass der Verstand aus verschiedenen jeweils relativ unintelligenten kleinen Prozessen, den sogenannten (mentalen) Agenten, aufgebaut ist. Ihr Zusammenspiel bildet die sog. „Society of Mind“. Als motivierendes Beispiel nennt Minsky die Aufforderung *„Hole eine Tasse Tee“*. Diese vereine einen Greif-Agenten, um die Tasse festzuhalten, einen Balancier-Agenten, um den Tee vor dem Auslaufen zu hindern, einen Durst-Agenten, um das Verlangen nach Teegenuss hervorzurufen und den Bewegungs-Agenten, um die Tasse zu den Lippen zu führen. Jeder dieser Agenten macht seine festgelegte Arbeit, für die er spezialisiert ist. Alle zusammen, als Kollektiv, erreichen dann die übergeordnete Absicht, hier speziell die des Teetrinkens. Ein für diese Arbeit interessanter Agent bzw. Prozess ist derjenige, welcher Struktur mit Funktion in Verbindung bringt. Minsky geht dabei davon aus, dass jede „Welt-Idee“ mit einer Vielzahl von Aspekten in Zusammenhang steht, wie Ursache, Handlung, Zweck oder Erläuterung. Eine entsprechende Verknüpfung, als Folge des Lernvorgangs, existiert auch zwischen *Struktur* und *Funktionalität*. So ist beispielsweise die Struktur eines Bogens verknüpft mit der Funktionalität einer Brücke, die Struktur von Beinen mit der Funktionalität des Stützens. Wie notwendig die gemeinsame Betrachtung von Struktur und Funktionalität bei der Sicht auf Dinge ist, zeigt Minsky am Beispiel eines Stuhls. So trifft die strukturelle Beschreibung „Ein Stuhl ist ein Ding mit Beinen und einer Rückenlehne und einem Sitz“ zwar auf viele Stühle zu, aber eben nicht auf alle. Die funktionale Beschreibung „Ein Stuhl ist etwas, auf dem man sitzen kann“ trifft zwar für alle Stühle zu, ist allerdings zu unspezifisch. Es ist also notwendig, die funktionale Betrachtung auch auf die Teilobjekte des Ganzen auszudeh-

[34] *engl.* „The affordance of anything is a specific combination of the properties of its substance and its surfaces taken with reference to an animal.“ Affordances are those things the environment „offers the animal, what it provides or furnishes either for good or ill.“(Gibson) [BR83]

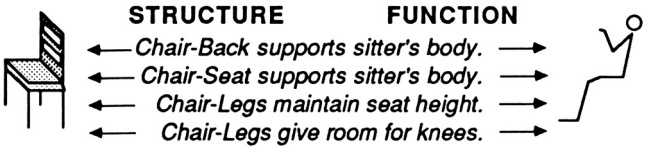


Abbildung 2.11: Zusammenhang von Struktur und Funktion bei einem Stuhl

nen, um die Generalisierbarkeit, die mit der funktionalen Beschreibung einhergeht, mit der Spezifität der strukturellen Beschreibung zu verknüpfen (vgl. Abbildung 2.11). Ein neuerer Artikel zu Minskys Society of Mind findet sich in [Sin03].

Über sein neues Buch [Min06] sagt Minsky in einem Interview: „Es ist eher eine Top-Down als eine Bottom-Up Theorie. Sie besitzt deshalb viel mehr Theorien darüber, welcher Art die höheren Organisationsformen sind ... . Ich denke immer noch, dass die Low-Level Theorien aus *The Society of Mind* nützlich sind. Das Buch startet mit einer Art von Freudschem Bild des Verstands als einem Sandwich von Grundinstinkten und High-Level Zielen und Konflikten. Es stellt Theorien auf, wie das High-Level Wissen repräsentiert wird, und geht runter bis zu dem Level, wo *The Society of Mind* aufhörte“[35] [Sch06].

Eine Umsetzung von funktionalen Aspekten im Bereich maschineller Objekterkennung erfolgte durch Stark, Bowyer, Woods [SB96], [WCH+95] , Rivlin [RDR95] oder Bogoni [BB95] und wird im Abschnitt 2.9, *Funktionalitätsbasierte Objekterkennungsverfahren*, behandelt.

2.8 Cognitive Vision

Der Bereich Cognitive Vision basiert auf der Disziplin Computer Vision, deren Aufgabe es ist, basierend auf Sensoreingaben die Umwelt dreidimensional zu rekonstruieren. Damit stellt Computer Vision (kurz: Vision) gewissermaßen die, wenn auch nicht eindeutige, Umkehrung des bei der Computergrafik stattfindenden Prozesses dar. Bei Computergra-

[35] *engl.* „It is a top down theory rather then bottom up. So it has much more theories of what the highest levels of organization are ... . I still think the low level theories of *The Society of Mind* are useful. So it starts with a sort of Freudian picture of the mind as a sandwich of basic instincts and high level goals and conflicts. And then it makes theories of how the high level knowledge is represented and goes down to about the level where *The Society of Mind* stopped.“

fik wird aus einem gegebenen 3-D Modell und bekannten Beleuchtungsverhältnissen, Kameraparametern sowie gegebener Betrachterposition die 2-D Ansicht berechnet. Dabei kann der Grad der modellierten physikalischen Oberflächeneigenschaften sehr aufwändig werden, wenn es wie beispielsweise in [MMK+06] darum geht, möglichst realistisches Virtual Prototyping im Bereich Fahrzeuginnenbereich zu betreiben (vgl. Abb. 2.12(a)). Bei Vision wird umgekehrt versucht, aus einer 2-D Aufnahme auf das verursachende Modell zu schließen. Dazu müssen zunächst die die Abbildung bestimmenden Parameter des Aufnahmesystems, wie Brennweite der Kamera, bestimmt werden. Dann wird beispielsweise durch die Bestimmung der Disparitäten korrespondierender Bildpunkte eines stereoskopischen Aufnahmesystems auf den Punktabstand geschlossen. Dies funktioniert jedoch nur für sehr texturreiche Aufnahmen robust. Bei konstant hellen Regionen oder kontinuierlichen Übergängen, kurz gesagt in kantenarmen Bildern, können korrespondierende Punkte der verschiedenen Kameras bzw. Kamerapositionen nicht zuverlässig bestimmt werden. Um sich diese Kanten künstlich zu verschaffen, kann man ein definiertes hell-dunkel oder farbkodiertes Streifenmuster projizieren. Durch Variation der Streifenbreite kann somit mit nur einer Kamera eine sehr gute Auflösung des rekonstruierten Modells erreicht werden. In Abb. 2.12(b) ist im linken Bild die Streifenprojektion auf einer Büste zu erkennen. Im rechten Bild sieht man das daraus erhaltene 3-D Modell. Um ein möglichst komplettes Modell zu erhalten, wurde die Büste aus zwei verschiedenen Perspektiven aufgenommen und diese Aufnahmen zueinander registriert, was an den unterschiedlichen Farben rot und grün im gerenderten Bild zu erkennen ist. Ziel solcher Rekonstruktionen kann das Modell an sich sein, beispielsweise zur Sicherung und Untersuchung archäologischer Funde, aber auch das Modell als Ausgangspunkt für eine angeschlossene Objekterkennung oder Szeneninterpretation. Als ultimatives und sehr visionäres Ziel von Vision wird in [MST00] aufgeführt, „Systeme zu erstellen, die ihre visuelle Umgebung autonom interpretieren können und zwar unter nahezu beliebiger Bedingung, was bedeutet, die erstaunliche Leistung der menschlichen visuellen Wahrnehmung nachzubilden“.

Der Begriff Cognitive Vision verbindet den Begriff Vision mit dem Begriff Kognition, der häufig mit den fortgeschrittenen Fähigkeiten insbesondere von Primaten verknüpft wird [CN06]. Drei Aspekte spielen dabei eine wesentliche Rolle: weite Anwendbarkeit, Robustheit und Geschwindigkeit. Eine weite Anwendbarkeit bedeutet, dass das kognitive System einfach an sich ändernde Umgebungsbedingungen angepasst werden kann. Dies soll das System durch einen Lernvorgang möglichst eigenständig erreichen. Robust-

(a) Gerenderte Szene eines Wageninnenraummodells [MMK+06]

(b) Streifenprojektionsmuster auf einer Büste und berechnetes 3-D Modell [Neu08]

Abbildung 2.12: Computerprahik und Computer Vision

heit bedeutet, dass das System nicht in chaotischer Weise, sondern hinreichend stabil auf kleine Änderungen der Eingabedaten reagiert. Betrachtet man aber real existierende Systeme, werden in der Regel nur zwei dieser Ziele wirklich erreicht. So bedeutet beispielsweise eine hohe Geschwindigkeit bei gleichzeitig hoher Robustheit in der Regel eine eingeschränkte Anwendbarkeit, wie beispielsweise bei typischen Machine-Vision Aufgaben in der Qualitätskontrolle.

Der Bereich Kognition lässt sich, wie schon zuvor erwähnt, in zwei verschiedene Lager aufspalten [Ver06b]:

- Kognitivistischer Ansatz: Symbolische (Wissens-)Repräsentation
- Emergenter Ansatz: konnektionistische, dynamische oder enaktive Systeme

Bei kognitivistischen Systemen ist „Kognition eine Art Berechnung, die auf symboli-

schen Repräsentationen arbeitet. Kognitive Systeme instanziieren solche Repräsentationen physikalisch als kognitive Codes und ... ihr Verhalten ist eine kausale Folge von Operationen, die auf diesen Codes ausgeführt werden“ [Ver06b], [Pyl84]. Diese These geht zurück auf Newell und Simon [NS76]. Dem gegenüber stehen die sog. emergenten Ansätze. „Kognition ist hier ein selbstorganisierender Prozess, bei dem sich das System ständig in Echtzeit neu aufbaut, um seine Funktionsfähigkeit durch Moderation von wechselseitiger Interaktion mit der Systemumgebung und Mitbestimmung über diese zu erhalten. Mitbestimmung impliziert, dass der kognitive Agent durch seine Umgebung bestimmt wird und zur gleichen Zeit der kognitive Prozess bestimmt, was für den Agenten real oder bedeutungsvoll ist. In gewisser Hinsicht bedeutet Mitbestimmung, dass der Agent seine Realität (seine Welt) als Ergebnis seiner Handlung in dieser Welt konstruiert“[36] [V+05]. Bedeutende Vertreter dieser Gruppe sind konnektionistische Verfahren. Diese haben ihren Ursprung in der Entwicklung von künstlichen neuronalen Netzen. Hier wird kein a-priori Modell vorgegeben und entsprechend finden symbolische Modelle keine Verwendung. Die parallelen Verbindungen zwischen den einzelnen neuronalen Einheiten definieren über ihre Gewichtung den Zustand des Systems. Kohonen Netzwerke beispielsweise erlauben es, ohne Vorgaben mehrdimensionale Daten zu clustern, d.h. nach gewissen Ähnlichkeitskriterien zu ordnen. Ein weiterer Vertreter konnektionistischer Verfahren ist der Bereich Genetischer Algorithmen.

Zur Entwicklung der Disziplin Cognitive Vision schreibt Vernon „Der Ausdruck Cognitive Vision wurde in den letzten Jahren eingeführt, um den Versuch zu umschreiben, robustere, stabilere und anpassungsfähigere Vision Systeme durch die Ausstattung mit kognitiven Fähigkeiten zu erreichen: die Fähigkeit zu lernen, sich anzupassen, das Abwägen von alternativen Lösungen und die Anwendung neuer Strategien für die Auswertung und Deutung“[37] [Ver06a]. Eine explizite Definition von Cognitive Vision lautet: „Cognitive Vision befasst sich mit der Einbindung und Steuerung von Vision Systemen durch Verwendung von expliziten, aber nicht notwendigerweise symbolischen Model-

[36] *engl.* „Here, cognition is a process of self-organization whereby the system is continually reconstituting itself in real-time to maintain its operational identity through moderation of mutual system-environment interactions and co-determination [MV87]. Co-determination implies that the cognitive agent is specified by its environment and at the same time that the cognitive process determines what is real or meaningful for the agent. In a sense, co-determination means that the agent constructs its reality (its world) as a result of its operation in that world.“ [V+05]

[37] „The term cognitive vision has been introduced in the past few of years to encapsulate an attempt to achieve more robust, resilient, and adaptable computer vision systems by endowing them with a cognitive faculty: the ability to learn, adapt, weigh alternative solutions, and develop new strategies for analysis and interpretation.“ [Ver06a]

len über Kontext, Situation und zielgerichtetem Verhalten. Cognitive Vision impliziert Funktionalitäten für Wissensrepräsentation, Lernen, Schlussfolgern über Ereignisse und Strukturen, Erkennung und Kategorisierung, und Zielanforderungen, die sich alle mit der Semantik der Beziehung zwischen dem visuellen Agenten und seiner Umgebung beschäftigen“[38] [ECV03]. Vernon führt zu dieser eher architekturbezogenen Definition noch einige generelle Funktionalitäten und nichtfunktionale Eigenschaften von Cognitive Vision Systemen auf [Ver05]. Danach existieren vier Stufen genereller Funktionalitäten eines Cognitive Vision Systems in Bezug auf die Erkennung, die von der Wahrnehmung des Vorhandenseins eines Objekts auf der untersten Stufe bis hin zu dessen Klassifikation in Kategorien auf der obersten Stufe reichen:

- *Entdeckung* von Objekten oder Ereignissen
- *Lokalisierung* von der Lage und der Ausdehnung von Objekten
- *Erkennung* eines lokalisierten Objekts im Sinne einer Kennzeichnung
- *Kategorisierung* eines Objekts

Daneben besitzt das System allgemeine Eigenschaften, wie

- *Zielgerichtetes Verhalten*
- *Anpassungsfähigkeit* gegenüber unerwarteten visuellen Änderungen: Das System soll in der Lage sein, mit Situationen klarzukommen, für die es so nicht explizit konstruiert wurde.
- *Vorausschätzung* für das Vorkommen eines Objekts oder das Auftreten eines Ereignisses: Das System soll also mehr erreichen als ein rein reflexives Reiz-Antwort-Verhalten.

Erreicht werden können diese Anforderungen durch folgende charakteristische Fähigkeiten, die sich mit der Aneignung, Speicherung und Verwendung von Wissen befassen:

[38] „Cognitive Computer vision is concerned with integration and control of vision systems using explicit but not necessarily symbolic models of context, situation and goal-directed behaviour. Cognitive vision implies functionalities for knowledge representation, learning, reasoning about events and structures, recognition and categorization, and goal specification, all of which are concerned with the semantics of the relationship between the visual agent and its environment.“ [ECV03]

- Die Fähigkeit, semantisches Wissen durch die Verknüpfung von Form und Funktion zu erlernen und entsprechende Wahrnehmungsstrategien zu entwickeln.
- Die Speicherung von Wissen über die Umgebung, das System und deren Beziehung.
- Die Überlegung[39] oder das Nachdenken über Objekte und Ereignisse in der Umgebung als auch des kognitiven Systems[40].

2.9 Funktionalitätsbasierte Objekterkennungsverfahren

Der Ansatz von Woods, Stark, Bowyer et al. [WCH+95], [SB96] ist ein bedeutender Vertreter für eine Kategorie von Objekterkennungsstrategien, bei der eine funktionalbasierte Identifikation der Objekte verfolgt wird. Objekte werden hierbei vorrangig dahingehend untersucht, inwieweit sie die Funktionalität einer möglichen Objektkategorie unterstützen. Ein Objekt wird dabei i. d. R. durch eine Menge unterschiedlicher funktionaler Eigenschaften definiert. In der ursprünglichen Version des Systems GRUFF (*engl.* Generic Representation Using Form and Function) wurde das Wissen über Objekte durch Experten vorgegeben. Über Zugehörigkeitsfunktionen, welche den Grad der Übereinstimmung bestimmter Eigenschaften angeben, wird die Zuordnung zu einer bestimmten Objektkategorie ermittelt. Im weiterentwickelten System wurde dafür ein Lernmodul entwickelt, das die Zugehörigkeitsfunktionen anhand vorgegeben markierter und kategorisierter Objekte erlernt.

Jede Objektklasse bzw. -kategorie wird durch ihre charakteristischen Funktionalitäten definiert. Für einen Stuhl mit gerader Rückenlehne sind diese:

- *bietet sitzfähige Oberfläche*
- *bietet Standfestigkeit*
- *bietet Rücken Abstützung*

[39] *engl.* deliberation

[40] Es werden drei Stufen des Denkvermögens kognitiver Agenten unterschieden: nicht-denkende Agenten mit reflexivem Verhalten, denkende Agenten mit vorgreifend anpassungsfähigem Verhalten und über Denken denkende Agenten, die über den kognitiven Prozess nachsinnen.

Diese Funktionalitäten werden ihrerseits durch bestimmte Eigenschaften, die mit der Form und den physikalischen Eigenschaften zusammenhängen, definiert. Dies bedeutet im Falle von *bietet sitzfähige Oberfläche*, dass die entsprechenden Werte von Breite und Tiefe der Sitzfläche und Höhe über der abstützenden Ebene gewissen Randbedingungen entsprechen müssen. Um diese kontinuierlichen Werte zu beschreiben und zu verarbeiten, werden beim GRUFF-System Fuzzy-Mengen verwendet.

Die unterschiedlichen Objektausprägungen gleicher Kategorie stehen über einen Kategorien-Definitions-Baum zueinander in Beziehung (vgl. Abb. 2.13). Die Wurzel beschreibt dort die Objektkategorie in der höchsten Generalisierung, bei diesem Beispiel einen Stuhl. Jeder Knoten, mit Ausnahme der Wurzel, hat zwei verschiedene Arten von Nachfolgern. Der eine Typ, der beim Wurzelknoten nicht vorhanden ist, ist vom Typ *funktionale Eigenschaft* und definiert jeweils die funktionale Eigenschaft der entsprechenden Kategorie. Der andere Nachfolger ist vom Typ *Unterkategorie* und definiert eine spezialisierte Kategorie der übergeordneten Elternkategorie. Die Subkategorie besitzt eigene funktionale Eigenschaften, die zusätzlich zu denen der Elternkategorie erfüllt sein müssen. Während ein gewöhnlicher Stuhl (*engl.* conventional chair) beispielsweise in dieser Kategorisierung nur die funktionalen Eigenschaften *bietet sitzfähige Oberfläche* und *bietet Standfestigkeit* hinreichend erfüllen muss, gilt für den Stuhl mit gerader Rückenlehne (*engl.* straightback chair) zusätzlich die Eigenschaft *bietet Rücken Abstützung*. Die physikalischen Primitiven, mit denen diese Eigenschaften bei Stühlen beschrieben werden, sind die *relative Orientierung* zweier Oberflächen, die *Ausdehnung*, *Lagebeschreibung* wie über oder unter, *Freiraum*, um unversperrten Raum zu beschreiben, und *Stabilität*, um die Standfestigkeit zu prüfen.

Neben CAD Daten wurde das Verfahren auch auf Laser Range Daten erfolgreich angewendet [SHGB93], [HGB95].

Ein anderer Ansatz, der ebenfalls funktionale Aspekte bei der Objekterkennung integriert, ist der von Rivlin et al. [RDR95]. Dieser Ansatz stellt eine Erweiterung eines strukturbasierten Vorgehens dar. Die das Objekt beschreibenden Formprimitiven zusammen mit ihren räumlichen Relationen sind hier erweitert durch funktionale Grundformen (*engl.* primitives) und entsprechende funktionale Relationen. Es werden vereinfachend solche Objektmodelle in Betracht gezogen, die eine eins-zu-eins Beziehung zwischen Formprimitiven und funktionalen Primitiven haben. Im allgemeinsten Fall wäre eine m-zu-n Beziehung möglich. Die *räumlichen Relationen* zwischen den Formprimiti-

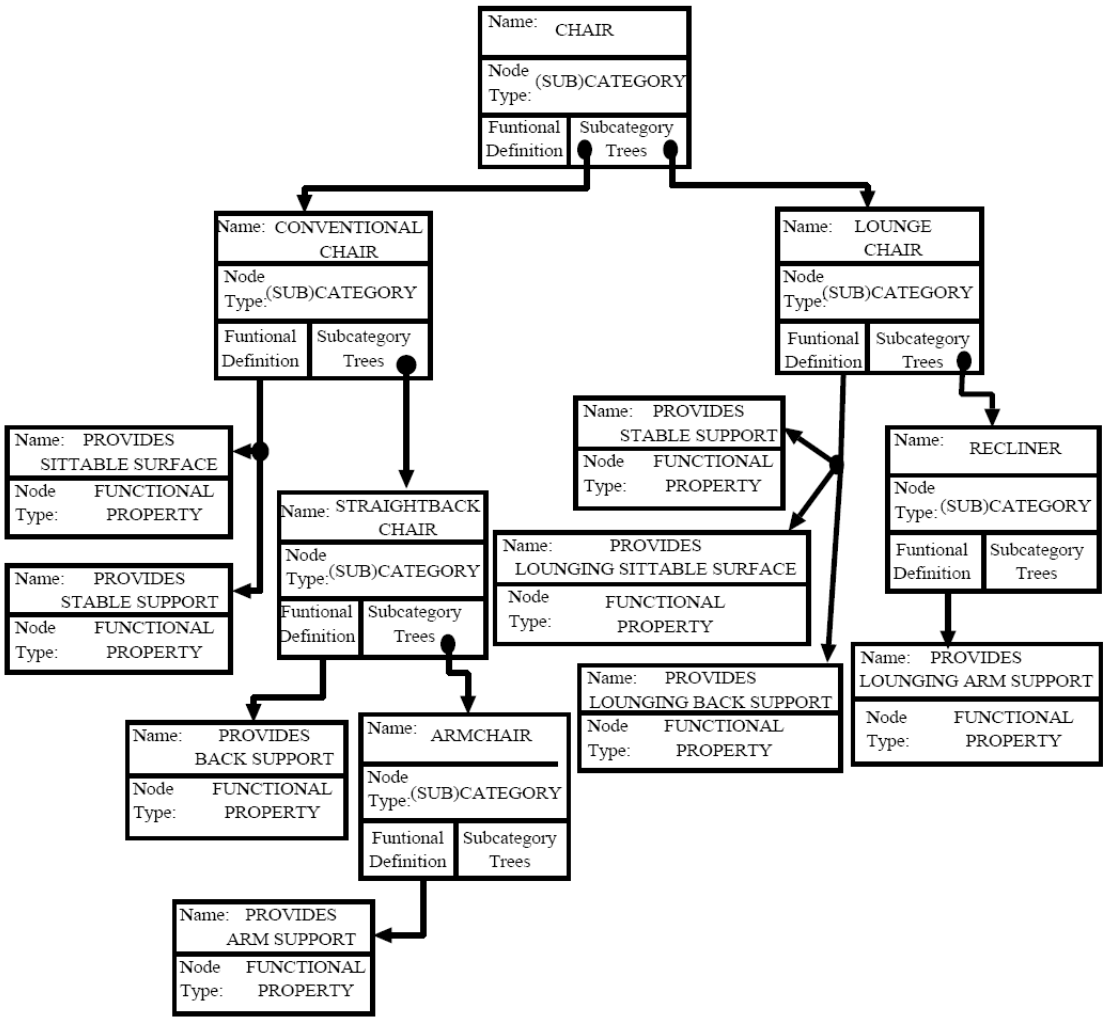


Abbildung 2.13: Funktionale Objekterkennung: Kategorien-Definitions-Baum von Stühlen [WCH+95]

ven werden durch Eigenschaften wie die beteiligten Verbindungsoberflächen der beteiligten Primitiven, durch ihre gegenseitige Lageposition und ihr relatives Größenverhältnis definiert. Die *funktionale Relation* beschreibt die Aufgabe der Interaktion zwischen den funktionalen Grundformen. Wie in Abb. 2.14 veranschaulicht, korrespondieren die Grundformen und Relationen aus der funktionalen Betrachtungsebene mit denen aus der formbasierten Betrachtungsebene.

Als Eingabe des Systems dienen 2-D oder 3-D Bilder. Im Segmentierungsschritt wer-

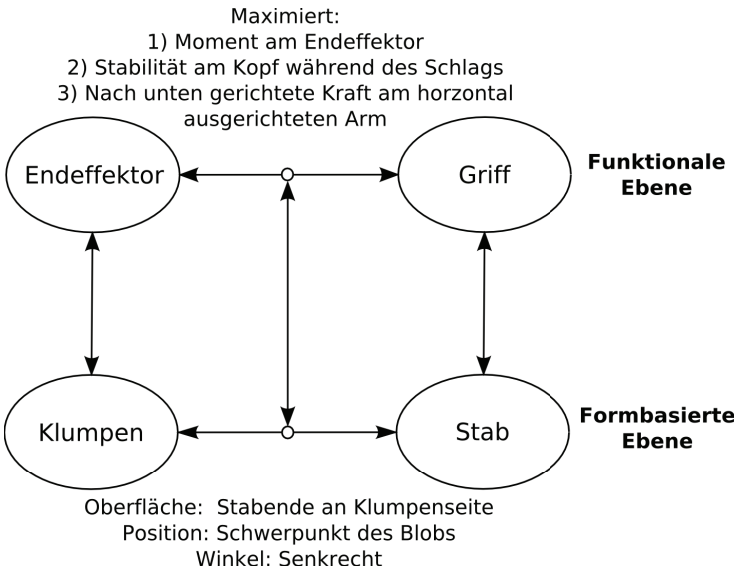


Abbildung 2.14: Funktionales Model eines Hammers [RDR95]

den 3-D Teile extrahiert, an die verformbare Superquadriken[41] angenähert werden. Dies führt zu einer der vier Grundformen: Stab, Streifen, Platte oder Klumpen, die über ihre relative Ausdehnung von Höhe, Breite und Tiefe unterschieden werden. Die räumlichen Relationen resultieren, wie oben beschrieben, aus der Art der räumlichen Verbindung oder der Orientierung zwischen den segmentierten Formen. Die räumlichen Primitiven und ihre Beziehungen werden dazu genutzt, um daraus funktionale Primitiven und Relationen abzuleiten. Diese funktionalen Beschreibungen wiederum werden mit funktionalen Referenzobjektmodellen verglichen.

Als Beispiel für die Anwendung des Verfahrens werden Handwerkzeuge betrachtet. Dort können die funktionalen Grundformen ein Endeffektor oder ein Griff sein. Der Endeffektor ist der Teil eines Objekts, der die entsprechende Aktion liefert, der Griff liefert die Kopplung zwischen der handelnden Person und dem Endeffektor. In dem gegebenen Beispiel bedeutet die funktionale Relation, dass „der Griff benutzt wird, den Endeffektor

[41] „Eine Superquadrik ist eine Gleichung einer bestimmten Form, deren Lösung eine geschlossene Oberfläche definiert. Indem die zwei Gleichungsparameter variiert werden, verformt sich die Oberfläche über ein Spanne von Objekten, welche Würfel, Diamanten, Pyramiden und glatte Zwischenformen enthält. Sechs weitere Parameter werden hinzugefügt, um die Achsenlängen, Krümmung längs zweier Achsen, Verjüngung entlang einer Achse zu bestimmen. Es ergibt sich daraus eine Familie von Objekten, die ausdrucksstärker sind als generalisierte Kegel oder Zylinder. Da jedoch so viele Parameter beteiligt sind, hat es sich als schwierig herausgestellt, Superquadriken zuverlässig zu erkennen, auch aus Entfernungsdaten heraus.“ [Pop94]

so in eine Richtung zu schwingen, dass die Kraft tangential zum Schwingbogen maximal wird, wobei die Schlagstabilität ebenfalls maximiert wird “. Zuerst werden die Formprimitiven aus dem Bild extrahiert, ein Stab und ein Klumpen, sowie deren räumliche Relationen. Danach werden die Formprimitiven auf die funktionalen Primitiven abgebildet, also auf Griff und Endeffektor. Danach werden die räumlichen Relationen auf funktionale Relationen abgebildet, also nach Relationen gesucht, die zu dieser Konfiguration zwischen Stab und Klumpen passen. Diese funktionalen Merkmale werden dann den entsprechend passenden funktionalen Modellen zugeordnet.

2.10 Mensch-Maschine Kommunikation

Ging es in den vorherigen Abschnitten dieses Kapitels um das Thema Objekterkennung, soll hier zum Abschluss des Kapitels auf die mögliche Anwendung einer so erhaltenen Szenenbeschreibung innerhalb einer Mensch-Maschine-Interaktion eingegangen werden. Auch das *ORCC*-System bietet die Möglichkeit, basierend auf der in der Szene erkannten Objekte in einen Dialog über diese Objekte zu treten.

Die Kommunikation zwischen Menschen und Maschinen oder auch zwischen Menschen selbst erfolgt zum Austausch von Informationen [Dah06]. Vorrangige Medien, die für diesen Informationsaustausch verwendet werden, sind Sprache, Gesten, Schrift oder Bilder. Entsprechend versucht die KI, Sprach- oder Gestenerkennung in künstlichen Systemen zu implementieren. Im Bereich von Auskunftssystemen, bei der die Syntax stark eingeschränkt ist, kann somit bereits erfolgreich mit Sprache kommuniziert werden. Öffnet man den Sprachumfang, kann es jedoch zunehmend zu Problemen kommen. Gründe dafür sind beispielsweise, dass die Syntax zu nicht eindeutigen Aussagen führen muss, „Zwei Drittel aller Männer halten Frauen für egoistisch“, die in unterschiedlicher Richtung interpretiert werden kann (vgl. Disambiguierung [Wah02]). Des weiteren können Wörter oder Gesten unterschiedliche Bedeutungen haben. Dieser Sachverhalt spielt auch bei der Einschätzung von räumlichen Relationen eine Rolle (vgl. Abschnitt 3.13, *Räumliche Relationen und Benutzerdialog*). Dies kann auch die Erfordernis von Kulturwissen zur richtigen Interpretation der Situation notwendig machen. Überlagert werden diese grundsätzlichen Probleme von der Schwierigkeit, akustische oder visuelle Zeichen hinreichend gut zu segmentieren und zu erkennen.

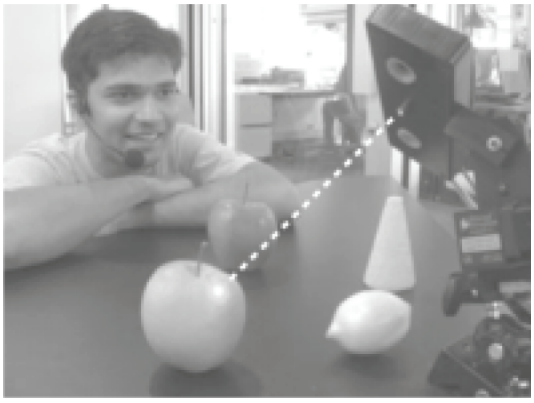


Abbildung 2.15: Objektauswahl durch einen Roboter des MIT [RGMJ02]

Befindet sich der Erkennungsprozeß alleine bereits im Spannungsfeld (vgl. Abschnitt 2.8, *Cognitive Vision*) zwischen Themenbereich bzw. Erkennungsumfang, Robustheit und Geschwindigkeit, vergrößert sich die Herausforderung, wenn die Erkennungskomponente um eine Schnittstelle zur sprachlichen Interaktion mit Menschen erweitert wird. Im Folgenden werden kurz zwei unterschiedliche Systeme vorgestellt, die sich dieser Aufgabe stellen.

Bei dem am MIT entwickelten System handelt es sich um ein sprachlernfähiges System, das gesprochene Sprache mit Objekterkennung kombiniert [RGMJ02]. Das System verwendet ausschließlich Farbbilder und einen Laser, mit dem es in der Anwendungsphase als Reaktion auf Spracheingaben auf Objekte zeigen kann. Die Kameras und der Laser sind auf einer Schwenk-Neige-Einheit mit zwei Freiheitsgraden montiert. Das System lernt mit dem Spracherkennungsystem basierend auf den visuellen Daten, Referenzierungen von erkannten Objekten zu verarbeiten. Dazu verfolgt das System verschiedene Objekte und berechnet ihre Eigenschaften, wie Farbe, Form, Höhe und Breite, und ihre räumlichen Beziehungen in Ausdrücken wie „above“, oder „to the left to“ [Reg96]. Während der Trainingsphase markiert das System ein segmentiertes Objekt mittels eines Lasers und erfasst dazu die Lautäußerung des Trainers. Diese Daten werden dann dazu genutzt, die Wörter und die Grammatik, basierend auf der visuellen Basis, zu erlernen. Nach dem Training kann der Benutzer Sätze sprechen wie „The green apple to the left of the cup“, die interpretiert werden. Das System markiert als Ergebnis auf diesen Satz das entsprechende Objekt mit dem Laser (vgl. Abb. 2.15). Von der Anwendung her gesehen, ist dieses System sehr ähnlich dem hier entwickelten *ORCC*-System. Die Umsetzung ist

Abbildung 2.16: PeopleBot der Universität Ulm [FKK+05]

allerdings ziemlich verschieden: Während das MIT-System ausschließlich Farbinformation verwendet und deshalb einen großen Vorder-Hintergrund-Kontrast als Voraussetzung für eine erfolgreiche Segmentierung benötigt, verwendet das *ORCC*-System primär 3-D Daten.

Ein weiteres System, das Objektidentifizierung in Kombination mit Benutzerinteraktion bietet, ist das System der Universität Ulm [FKK+05]. Technisch ist dies auf der Baureihe PeopleBot™ der Firma MobileRobots Inc.[42] aufgesetzt. Die entwickelte Softwarearchitektur hat den Anspruch, neurobiologisch plausibel zu sein, was zur Folge, hat dass etliche Teilsysteme als künstliche neuronale Netzwerke (*engl.* Artificial Neural Network - ANN) oder neuronale Assoziativspeicher (*engl.* Neural Associative Memory - NAM) implementiert sind. Die Aufgabe des Roboters ist es, auf gesprochene Sätze wie „bot show plum“ zu reagieren. Das System ist also in der Lage, Objekte zu erkennen, und auf die Spracheingabe zu reagieren. Im ersten Schritt der Objekterkennung werden die Aufnahmen segmentiert, beispielsweise in Bezug auf einen bestimmten Farbbereich. Die Segmente werden dann bezüglich ihrer Orientierung und Farbe charakterisiert. Dazu werden unter anderem eine Durchschnittsfarbe und ein Orientierungshistogramm [Fre95]

[42] MobileRobots Inc. (ehemals ActivMedia Robotics):
`http://www.activrobots.com/ROBOTS/peoplebot.html`

berechnet. Zur Klassifizierung dieser Merkmale werden hierarchisch angeordnete Radial-Basis-Funktions Netzwerke (*engl.* Radial Basis Function - RBF) verwendet. Dabei wird die eigentliche Klassifikationsaufgabe in eine Kaskade von Unterscheidungsklassifikationen zerlegt. Trainiert werden die Netzwerke durch ein überwachtes Lernverfahren. In der Anwendungsphase werden die Objekte, auf die sprachlich referenziert wird, mit einem Fadenkreuz gezeigt. Der beschriebene Testlauf umfasste sieben verschiedene Obstobjekte.

Die hier beschriebenen Systeme verwenden ausschließlich farbbasierte Informationen zur Klassifikation der Objekte. Für die Segmentierung der Objekte sind deshalb günstige Lichtverhältnisse und gewisse Einschränkungen bzgl. des Hintergrundes wichtig. Bei dem hier entwickelten *ORCC*-System wird deshalb eine dreidimensionale Segmentierung verwendet. Auf deren Grundlage können auch unbekannte Objekte unter schwierigen Verhältnissen segmentiert werden (vgl. Abschnitt 3.10, *Texturausschnittsberechnung*).

Kapitelzusammenfassung

Dieses Kapitel hat seinen Schwerpunkt auf der Betrachtung von Methoden und kognitiven Aspekten im Bereich Objekterkennung. Zu Beginn wird die Erkennungsaufgabe in ihren unterschiedlichen Zielanfragen betrachtet und anschließend auf die visuelle Repräsentation beim Menschen eingegangen, bei der ansichtenspezifische 2-D Ansichten eine große Rolle spielen. Es folgten die Darstellungen von vier kognitiv motivierten Erkennungsansätzen und jeweiliger Vertreter davon: der Erkennung durch Erfassung invarianter Merkmale wie Farbe oder Geometrie, der Erkennung durch Entdecken der räumlichen Struktur der Einzelteile, der Erkennung durch Vergleich des Stimulus-Bildes mit dem Gedächtnisbild nach erfolgter Registrierung der beiden und der Erkennung durch Korrelation mit einer Vielzahl von Gedächtnisbildern. Der darauf folgende Abschnitt erweitert die Betrachtung der bisher auf visuellen Eigenschaften basierenden Erkennungsverfahren um den Aspekt der Funktionalität. Dieser ist von großer Bedeutung, da die Funktionalität auch für uns Menschen bei der Kategorisierung von Objekten eine große Rolle spielt. Eine der spezifischen Methodik übergeordnete Betrachtung des Themas erfolgt in dem anschließenden Abschnitt Cognitive Vision. Diese Disziplin hat sich gebildet, um mittels kognitiv inspirierter Aspekte Ideen zu liefern, welche die Robustheit von Erkennungssystemen verbessern. Abgeschlossen wird das Kapitel mit der Vorstellung einiger Mensch-Maschinen Dialog-Systeme, deren Wissen über die Umgebung auf den Ergebnissen eines Erkennungsystems basiert.

Es folgt mit dem Kapitel 3, *Komponenten des ORCC-Systems*, die Vorstellung des eigens entwickelten Systems *ORCC*, bei dem hier erläuterte Aspekte, wie die funktionalbasierte Erkennung, eine Rolle spielen.

3 Komponenten des ORCC-Systems

Das zentrale Ergebnis dieser Arbeit ist das System *ORCC*, mit dem Objekte in einer Szene erkannt und in einem sprachbasierten Mensch-Maschine-Dialog genutzt werden können. In diesem Kapitel beginnt mit der Erkennung einfach strukturierter Objekte der Einstieg in die Betrachtung der eigens entwickelten Module dieses Systems. In den darauf folgenden Abschnitten wird auf die Segmentierung als Vorstufe für den Erkennungsprozess eingegangen und die Funktionsweise des hier verwendeten 3-D Segmentierungsalgorithmus vorgestellt. Anschließend wird das für die Erkennung der einfach strukturierten Objekte verwendete Modell, das auf Bayesschen Netzen aufbaut, vorgestellt. Die Fähigkeit des Systems, die Merkmale und Strukturverbindungen der Objekte zu erlernen, wird gezeigt. Die experimentelle Umsetzung wird an den Beispielen von Raumbegrenzungen und funktional-strukturell einfachen Objekten, wie einem Tisch, vorgestellt. Im nächsten Abschnitt wird gezeigt, dass sich die hier erkannten Objekte auch für eine Registrierung, d.h. Integration mehrerer Ansichten, verwenden lassen. Danach folgt die funktionalbasierte Erkennung strukturierter Objekte am Beispiel von Stühlen. Im Anschluss daran wird mit der Vorstellung eines 3-D modellbasierten Verfahrens das Repertoire der Erkennungsverfahren des *ORCC*-Systems erweitert. Dieses erlaubt es, auch Objekte, von denen keine hinreichenden funktionalen Merkmale erschlossen werden können, zu klassifizieren. Als Eingabe verlangt dieses Modul dreidimensional segmentierte Objekte. Um diese zu berechnen, wird ein punktdichte- und ebenenbasiertes Segmentierungsverfahren vorgestellt. Für die sprachliche Interaktion des *ORCC*-Systems mit dem Benutzer, die Anwendung der hier angestrebten Szenenbeschreibung ist, spielen neben der Kategorie auch texturorientierte Eigenschaften ein Rolle. Um diese Verbindung zwischen 3-D und 2-D Informationen zu ermöglichen, wird ein Verfahren vorgestellt, mit dem das *ORCC*-System die Ansichten des Laserscanners mit denen einer Farbbildkamera kombiniert. Damit können dann zu den kategorisierten 3-D Objekten die entsprechenden Texturausschnitte berechnet werden. Diese werden dann dazu genutzt, eine farbliche Beschreibung dieser Objekte zu erlangen. Insgesamt ergibt sich damit eine multimodale Szenenbeschreibung, die Grundlage für die folgende Benutzerinteraktion ist. Im letzten Abschnitt

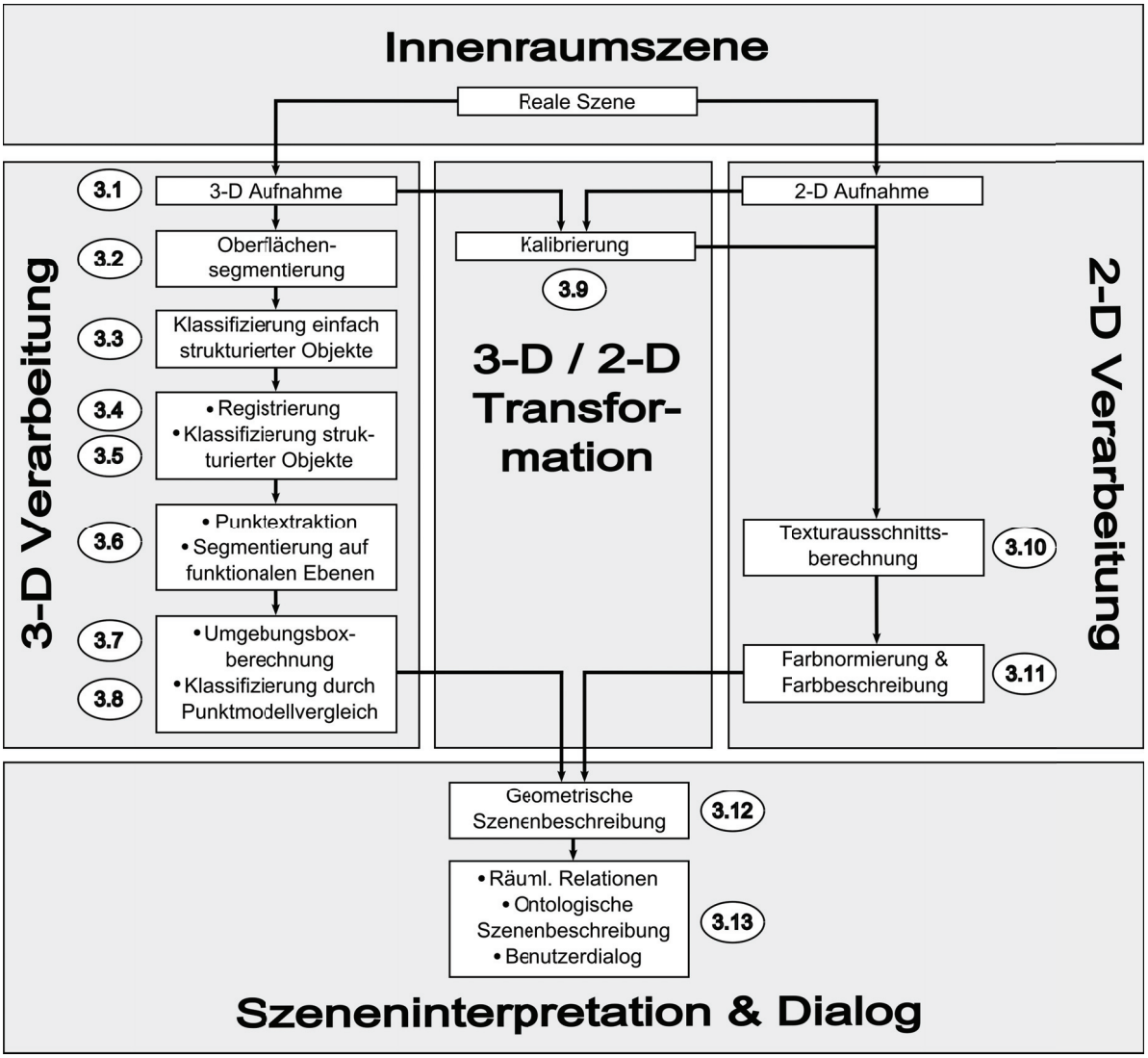


Abbildung 3.1: Komponenten des *ORCC*-Systems

wird diese Benutzerinteraktion, bei der Objekte per Spracheingabe räumlich referenziert werden können, vorgestellt.

Die Abbildung 3.1 gibt einen Überblick über die in diesem Kapitel beschriebenen Methoden. Ausgehend von der realen Szene, gibt es auf der linken Seite einen Zweig, der sich mit der Verarbeitung der Tiefeninformation beschäftigt. Auf der rechten Seite sind Themen dargestellt, die sich mit der Farbbildverarbeitung beschäftigen. Verknüpft sind die beiden Bereiche durch die berechnete 2-D/3-D Transformation. Zu allen Verfahren

oder Methoden ist jeweils die Kapitelnummer angegeben. Beide Zweige münden in einer Szenenbeschreibung, die Grundlage für einen Mensch-Maschine Dialog ist.

Eine zusammenhängende Gesamtdemonstration des Systems folgt im letzten Kapitel 4, *Anwendung*, dieser Arbeit. Dort findet sich auch eine Ablaufübersicht mit dargestellten Zwischenergebnissen.

3.1 Sensordaten

Die dreidimensionalen Sensordaten entstammen einem auf einer Pan-Tilt Einheit montierten 2-D Laserscanner. Der systematische Fehler der Entfernungsmessung wird mit ± 20 mm angegeben, der statistische mit 5 mm. Der statistische Fehler der Pan-Tilt Einheit entspricht beim Neigevorgang mit 0.02° bei zwei Meter Abstand einem Positionierungsfehler von kleiner 1 mm.

Neben diesen spezifizierten Abweichungen zeigt der Scanner im Betrieb leider eine weitere störende Eigenschaft. An dreidimensionalen Kanten kommt es im Verschattungsbereich zu Scheinpunkten. In Abb. 3.2 ist dies am Beispiel einer realen Tischkante aus zwei Ansichten gezeigt. Ursache für dieses Verhalten ist, dass an Kanten der aufgeweitete Laserstrahl zwei verschiedene zeitliche Reflexionsmaxima erzeugt, eines das von dem

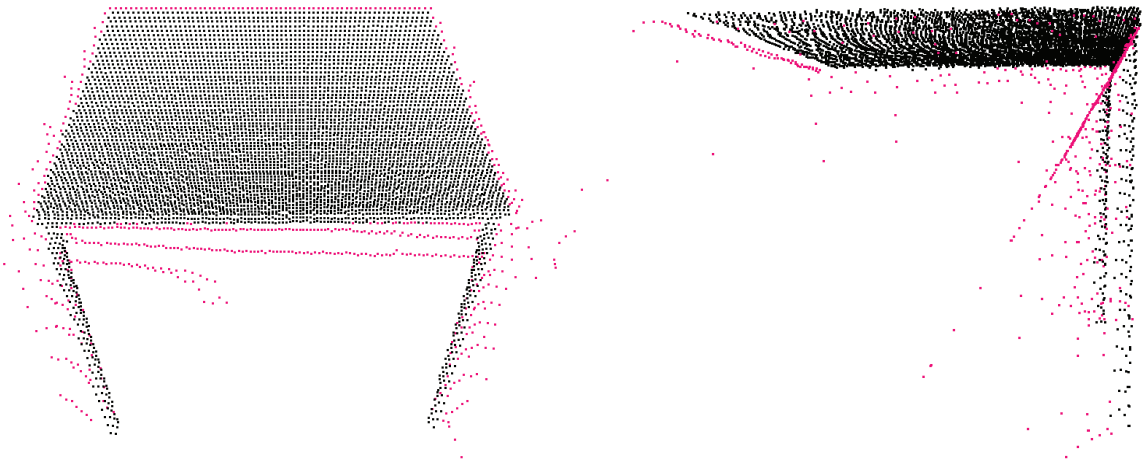


Abbildung 3.2: Verschattungspunkte an einer Tischkante in Vogelperspektive und Seitenansicht

Kantenpunkt ausgeht und eines, das von der im Hintergrund getroffenen Stelle ausgeht. Die Auswertungselektronik des Scanners interpretiert dieses Signal dann als einen Punkt, der je nach Verhältnis der Reflexionswerte näher am Kantenpunkt oder dem Punkt im Hintergrund liegt. Für die Segmentierung der Oberflächen über die Normalenausrichtung, wie sie in den vorherigen Abschnitten beschrieben wurde, ist dies ebenso problematisch, wie für die im Abschnitt 3.6, *Segmentierung auf funktionalen Ebenen*, beschriebene dichtebasierte Segmentierung. Bei der normalenbasierten Segmentierung kann es dazu führen, dass bei geraden Kanten ein gleichmäßiger Verschattungvorhang

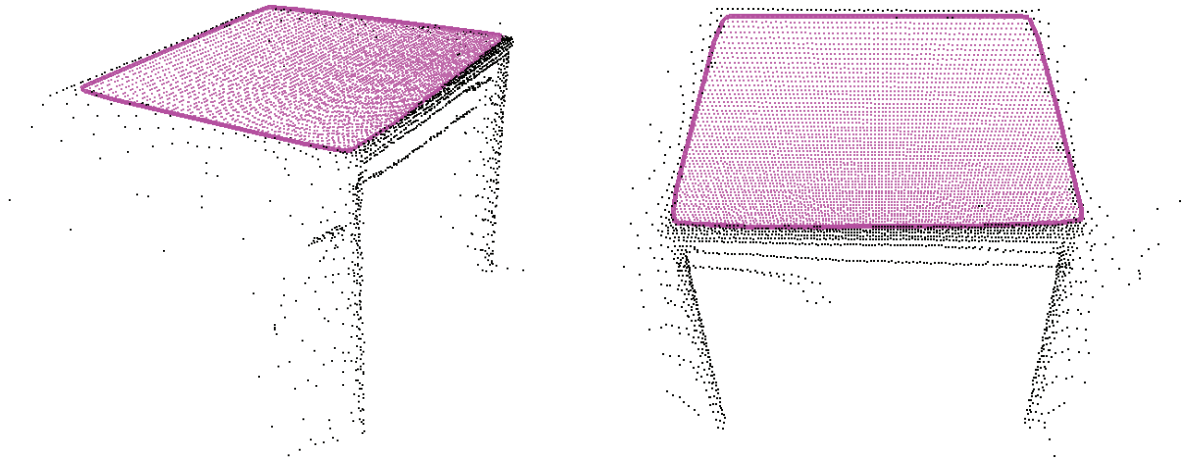


Abbildung 3.3: Segmentierte Tischoberfläche

entsteht, der dann als Scheinfläche interpretiert wird. Die dichtebasierte Segmentierung wird angewendet, um beispielsweise Objekte auf einem Tisch zu isolieren. Die Verschattungspunkte können dort dazu führen, dass sie die Übereinstimmung der Punktwolke mit der wahren Form mitunter stark verfälschen. Des weiteren kann es dazu kommen, dass nahe beieinander stehende Objekte miteinander verschmelzen und nicht mehr erfolgreich getrennt, d. h. segmentiert, werden können.

Welche Auswirkungen das auf die Berechnung von Merkmalen hat, soll durch die Berechnung der Größe einer Tischoberfläche eines quadratischen Tisches der Kantenlänge 80 cm gezeigt werden. Wie in Abb. 3.3 zu sehen, sind an den Kanten starke Verfälschungen der Tiefenwerte zu erkennen. Die aus der gemessenen Fläche abgeleitete Kantenlänge von knapp 77 cm bedeutet, dass im Durchschnitt auf jeder Seite des Tisches ca. 1.5 cm breite Streifen aufgrund der verfälschten Messpunkte verloren gehen. Bei den hier verwendeten auf der Ebenenextraktion basierenden Objekterkennungsverfahren fällt diese Abweichung jedoch nicht weiter ins Gewicht.

3.2 Oberflächensegmentierung

Unter Segmentierung versteht man die Teilung der Daten in zusammengehörige Gruppen. Damit stellt die Segmentierung insbesondere bei Bottom-Up orientierter Vorgehensweise in der Regel eine Vorstufe der Klassifikation dar. Es werden dabei gewisse

Merkmale im Bild berechnet, die dann als Grundlage für die Berechnung der Ähnlichkeit benachbarter Bildpunkte und der darauf folgenden Gruppierung eingesetzt werden. Die im Bild segmentierten Regionen werden dann der weiteren Verarbeitung zugeführt. Eine formale Definition lautet in Anlehnung an [GW92]: Sei R die Gesamtregion eines Bildes und seien $R_1, \ldots, R_n$ die segmentierten Teilregionen des Bildes, dann gilt:

$$R = \bigcup_{i=1}^{n} R_i$$
$$R_i \cap R_j = \emptyset \quad \forall\ i,\ j,\ i \neq j$$

Bei 2-D Bildern kann die Segmentierung beispielsweise die Suche nach spektralen Auffälligkeiten im Bild bedeuten. Dabei werden die Werte isolierter Pixel betrachtet, um danach Regionen bestimmter Eigenschaften zu finden. Entsprechendes gilt auch für 3-D Aufnahmen. Hier könnte in einer einfachen Applikation die Segmentierungsaufgabe darin bestehen, isoliert die Zugehörigkeiten zu Entfernungszonen zu berechnen, beispielsweise mit dem Ziel, Schutzzonenverletzungen festzustellen. Segmentierung muss aber nicht auf lokal isolierte Werte beschränkt sein. Es kann auch bedeuten, dass Umgebungspunkte dazu benutzt werden, Verläufe zu berechnen, um letztlich Kantenpixel zu bestimmen. Dies nennt man dementsprechend kantenbasierte Segmentierung [Jäh02]. Bei der modellbasierten Segmentierung hingegen wird nach bestimmten Formen im Bild gesucht. Ein Beispiel dafür wäre die sog. Hough-Transformation zum Finden gerader Kanten. Die Ergebnisse der Segmentierung sind häufig von Parametern oder Schwellwerten abhängig. Sind diese ungünstig gewählt, kann es dazu führen, dass das Bild in sehr viele kleine Regionen zerfällt (Übersegmentierung) oder sich in eine zu kleine Anzahl von Regionen aufteilt (Untersegmentierung).

In dieser Arbeit besteht die Segmentierungsaufgabe darin, zusammengehörige Oberflächenstücke zu detektieren. Eine Sammlung von auf 3-D Daten arbeitenden Segmentierungsverfahren findet sich in [HJBJ+96]. Umgesetzt wurde in dieser Arbeit ein auch zur Registrierung von 3-D Aufnahmen [SA00] erfolgreich eingesetztes Region-Growing Verfahren, bei dem lokal anhand von Umgebungspunkten die Ausrichtung des Pixels bestimmt wird. Im Folgenden wird dieses Verfahren vorgestellt. In den folgenden Unterkapiteln werden dann, unterschieden nach der Auswahl der Nachbarschaft - per Index der Aufnahmematrix oder räumlicher Distanz - die Ergebnisse für eine Ground Truth Aufnahme gezeigt.

Der Algorithmus zur Oberflächensegmentierung besteht aus zwei Schritten und ver-

wendet als Kriterium die Ausrichtung eines Punktes, die durch dessen Nachbarschaftspunkte gegeben ist:

1. *Normalenberechnung* unter Verwendung benachbarter Punkte durch Singulärwertzerlegung[43]

2. Bestimmung zusammenhängender Regionen durch *Regionenwachstum* (*engl.* Region-Growing) basierend auf den Kriterien

   - *Konormalität*
     Die Orientierung benachbarter Punkte ist hinreichend ähnlich

   - *Koplanarität*
     Die benachbarten Punkte liegen hinreichend nahe auf einer gemeinsamen Ebene

Im ersten Schritt wird für jeden Punkt anhand der ihn umgebenden Pixel seine Orientierung im Raum berechnet. Dazu wird eine Ebene so im Raum positioniert, dass die Summe der quadratischen Abweichungen bzw. Abstände von der Ebene minimal wird. Mathematisch lässt sich diese Ausgleichsrechnung mittels der sog. Singulärwertzerlegung[44] lösen [PTVF92]. Die das Ausgleichsproblem beschreibende $m \times n$ Matrix $\mathbf{A}$ wird in ein Produkt aus einer $m \times n$ Spalten-orthogonalen Matrix $\mathbf{U}$, einer $n \times n$ Diagonalmatrix $\mathbf{W}$, deren positiv- oder nullwertigen Diagonaleinträge die Singulärwerte w_i sind, und einer transponierten $n \times n$ Orthogonalmatrix $\mathbf{V}$ zerlegt:

$$\begin{pmatrix} & \\ & \mathbf{A} \\ & \end{pmatrix} = \begin{pmatrix} & \\ & \mathbf{U} \\ & \end{pmatrix} \begin{pmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_3 \end{pmatrix} \begin{pmatrix} & \\ & \mathbf{V}^T \\ & \end{pmatrix}$$

Das Problem kann also wie folgt formuliert werden: Sei P der Punkt, für den die Ausrichtung berechnet werden soll, und P_i, $i = 1, \ldots, n$, die verwendeten Umgebungspunkte und $\mathbf{m}$ der Schwerpunkt aller Punkte. Die Abstände b der Punkte P und P_i zu einer

[43] *engl.* Singular Value Decomposition (SVD)
[44] In [SA00] wird die Aufgabe der Ebenenannäherung als Eigenwert-Problem behandelt. Zur Beziehung beider Methoden siehe [Gon02].

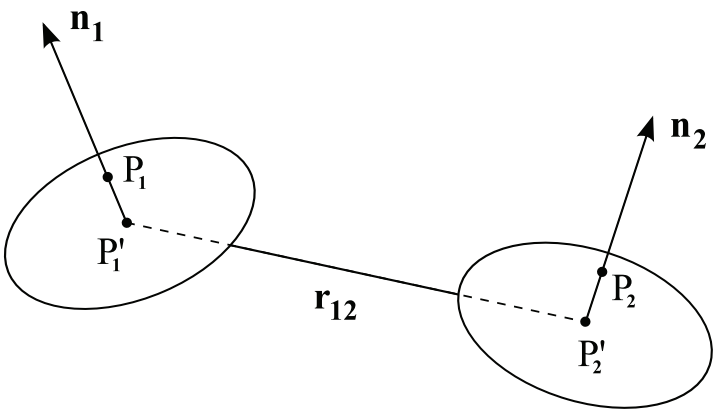


Abbildung 3.4: Kriterien für Regionenwachstum: Koplanarität und Konormalität benachbarter Punkte [SA00]

durch $\mathbf{m}$ gelegten Ebene mit der Normalen $\mathbf{n}$ ergeben sich damit jeweils zu:

$$b = (\mathbf{x} - \mathbf{m})\mathbf{n}$$

wobei $\mathbf{x} = \mathbf{x}(Punkt)$ der zu den Punkten P und P_i gehörende Ortsvektor ist. Der quadratische Gesamtfehler χ^2, den es zu minimieren gilt, ergibt sich damit zu

$$\chi^2 = \sum_{\substack{\mathbf{x}(i) \\ i \in \{P, P_1, \ldots, P_n\}}} ((\mathbf{x} - \mathbf{m})\mathbf{n})^2 \rightarrow \min$$

Die $(\mathbf{x} - \mathbf{m})$ ergeben die Parameter eines linearen und überbestimmten Gleichungssystems, welches oben mit der Matrix $\mathbf{A}$ beschrieben wird. Die Lösung des Gleichungssystems, d.h. die Bestimmung des gesuchten Normalenvektors, ergibt sich durch den zum kleinsten Singulärwert w_i gehörenden Spaltenvektor der Matrix $\mathbf{V}$ infolge der Singulärwertzerlegung. Dieser Singulärwert kann auch als ein Maß für die Güte der angepassten Ebene genutzt werden.

Aufbauend auf den berechneten Normalenvektoren werden die Segmente, basierend auf den beiden Kriterien Konormalität und Koplanarität, durch Regionenwachstum berechnet. Die Abbildung 3.4 zeigt die Situation zweier benachbarter Punkte P_1 und P_2. Ihre Projektion auf ihre lokal berechnete Ebene entspricht den Punkten P_1' und P_2'. Für das Kriterium der Konormalität wird die Abweichung der Ausrichtungen der beiden Normalenvektoren untersucht, die sich über das Skalarvektorprodukt einfach berechnen lässt. Für den Öffnungswinkel α, der durch die beiden Normalenvektoren aufgespannt

wird, gilt:

$$\alpha = \arccos\left(\mathbf{n}_1\mathbf{n}_2\right)$$

Für das Kriterium der Koplanarität werden die Abstände der auf die Ebenen projizierten Punkte P_1' und P_2' zu der durch den jeweils anderen projizierten Punkt und Normalenvektor gegebenen Ebene bestimmt. Das Maximum beider Werte ergibt dann den Abstand d, der unter einem festgelegten Schwellwert liegen muss:

$$d = \max\left(\left|\mathbf{r}_{12}\mathbf{n}_1\right|, \left|\mathbf{r}_{12}\mathbf{n}_2\right|\right)$$

Die Schwellwerte, die hier für Konormalität und Koplanarität, aber auch für die Normalenberechnung notwendig sind, sind stark abhängig von der Güte der vorgegebenen 3-D Daten. Je verrauschter diese sind, desto weicher sind besagte Parameter einzustellen, um zusammengehörige Segmente zu erhalten. Eine zu großzügige Einstellung kann jedoch dazu führen, dass eine Untersegmentierung stattfindet, d.h. eigentlich geometrisch nicht zueinander gehörige Teile miteinander verschmelzen.

Neben den Segmenten, d.h. den zusammengehörigen Punktmengen, können in diesem Verarbeitungsschritt noch weitere für die Erkennung wertvolle Merkmale bestimmt werden, nämlich die Position, Orientierung und auch der Flächeninhalt der segmentierten Fläche. Weiteres zu diesem Punkt folgt in den beiden nächsten Abschnitten.

3.2.1 Nachbarschaft über Distanz

Neben den Einstellungen der Parameter für die Ebenenberechnung und den Region-Growing Algorithmus stellt auch die Bestimmung der Nachbarschaftspixel eine variable Größe dar. Es besteht einerseits die Möglichkeit, die Nachbarschaft auf die Position des Punktes in der Aufnahmematrix[45] zu beziehen, oder andererseits den geometrischen Abstand im Raum zu betrachten. Der in diesem Abschnitt betrachtete letztere Fall ist der zunächst näherliegende, da er die wirkliche Position der Punkte im Raum berücksichtigt. Allerdings erkauft man sich diesen Vorteil mit einem Aufpreis, denn es müssen für jeden Punkt die Nachbarpunkte extra bestimmt werden. Um diese aufwändige

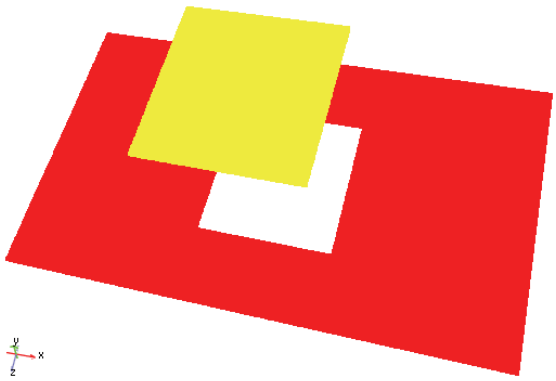
[45] Der Scanner liefert zunächst nur ein eindimensionales Array von Tiefeninformationen. Durch Schwenken entsteht dann ein als Matrix angeordnetes Raster von Tiefeninformationen, das hier Aufnahmematrix genannt wird.

Operation möglichst effizient zu halten, wird der sog. kd-Baum (*engl.* kd-tree) eingesetzt [CGA07], [Kle05].

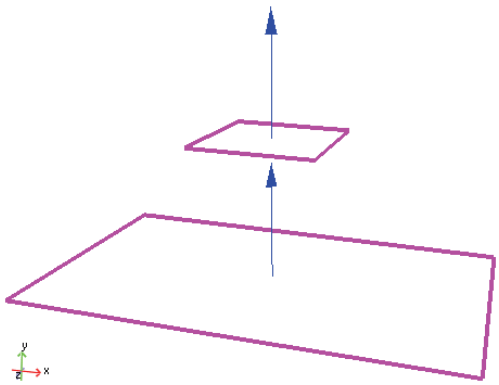
Um die Eigenschaften näher untersuchen zu können, wird ein 3-D Groundtruth Bild erstellt, das semantisch etwa einer von oben aufgenommenen Tischplatte der Kantenlänge 1×1 m entsprechen könnte. Das Segmentierungsresultat zeigt sich in Abb. 3.5(a). Das Ergebnis entspricht idealerweise der Vorstellung von zwei Segmenten, einem für den Boden und einem für die Tischplatte. Es gibt keinerlei Fehlklassifikation. Die berechnete Orientierung entspricht mit dem Wert $n = (0, 1, 0)$ für den Boden und den Tisch in idealer Weise den trigonometrischen Erwartungen. Diese Orientierungswerte werden erhalten, indem zu den berechneten Segmenten erneut eine Ebene angefittet wird. Zur Berechnung der Segmentgröße werden die Segmentpunkte auf die angefittete Ebene projiziert. Für diese Punkte in der Ebene wird dann das sie konvex umschreibende Polygon berechnet und dessen Flächeninhalt als Größenangabe für das Segment herangezogen. Es zeigt sich, dass mit einem Wert von $f = 1.00\ m^2$ der Flächeninhalt des Tisches exakt berechnet wird, gleiches gilt für den Boden. In Abb. 3.5(b) sind diese Polygone dargestellt. Das konvexe Polygon beschreibt je nach Datensegment dieses nicht exakt, insbesondere dann, wenn es starke Einbuchtungen hat. Da die 3-D Datenpunkte aber gerastert und nicht hinreichend kontinuierlich verteilt sind, kann in der Regel nicht unterschieden werden, wann es sich um eine Einbuchtung des Gegenstands oder einen Mangel an Messpunkten handelt.

3.2.2 Nachbarschaft über Index

Bei dieser Nachbarschaftsbeziehung wird nicht die tatsächliche Position im Raum, sondern die Position in der Aufnahmematrix zur Nachbarschaftsberechnung herangezogen. Dies hat deutliche Performanzvorteile, da über die Aufnahmeindizes direkt auf die Nachbarn zugegriffen werden kann. Da damit aber an Kanten auch sehr weit entfernte Punkte in die Ebenenberechnung einfließen, ist in solchen Bereichen die Normalenberechnung notwendigerweise gestört. Das Bild 3.6(a) zeigt die erhaltenen Segmente. Gegenüber dem Verfahren mit geometrisch basierten Nachbarn zeigen sich deutliche Artefakte an Sprungkanten. Die Auswahl der Nachbarschaft basierend auf den Indizes führt dazu, dass an den Randpunkten des stilisierten Tisches auch Hintergrundpixel des Bodens verwendet werden. Bei der Normalenberechnung kann es dazu führen, dass der Punkt mitunter



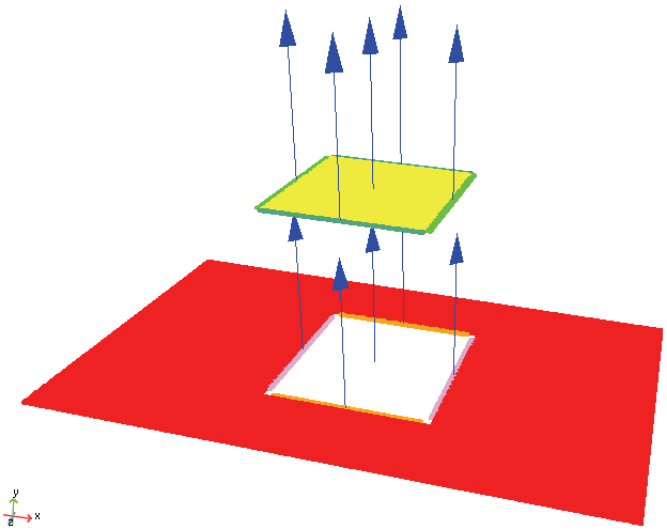
(a) Segmentierte Flächen



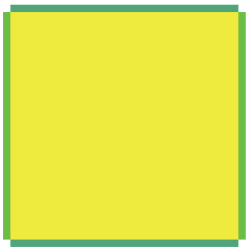
(b) Umfang beschreibende Polygone und Normalenvektor

Abbildung 3.5: Segmentierung über geometrische Nachbarschaft

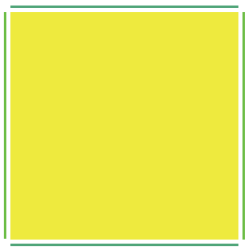
verworfen wird, wenn der entsprechende Parameter streng gewählt wird. Im Segmentierungsschritt kann es bedeuten, dass die Normalenvektorrichtung zu stark differiert und deshalb diese Punkte unterschiedlichen Segmenten zugeordnet werden. In Bild 3.6(b) wurde der Schwellwert für die Normalenberechnung großzügig, d.h. groß gewählt. Der Wert bei der Segmentierung ist entsprechend tolerant gewählt, so dass an den Rändern jeweils ein breiter Streifen entsteht, der durch Hintergrundpixel noch stark beeinflusst wird. Entsprechend sind die Normalen dort nach außen orientiert und verschmelzen deshalb nicht mit dem zentralen gelben Bereich. Die in Bild 3.6(a) gezeigte vertikale Ausrichtung dieser Segmente rührt daher, dass zum Abschluss die Normalen mit allen Segmentpunkten neu berechnet werden. In Bild 3.6(c) wird das Segmentierungsergebnis für einen strenger gewählten Schwellwert bei der Normalenberechnung gezeigt.



(a) Segmentierungsergebnis mit Normalenvektoren der detektierten Oberflächsensegmente



(b) Permissiv gewählter Schwellwert bei der lokalen Ebenenberechnung



(c) Streng gewählter Schwellwert bei der lokalen Ebenenberechnung

Abbildung 3.6: Segmentierung über Indexnachbarschaft

3.3 Erkennung einfach strukturierter Objekte

In diesem Abschnitt wird die funktionalbasierte Erkennung für einfach strukturierte Objekte vorgestellt. Diese beschränkt sich in der Anwendung auf Objekte, die für die noch folgende dichtebasierte Segmentierung (vgl. Abschnitt 3.6, *Segmentierung auf funktionalen Ebenen*) von Belang ist: *Fußboden*, *Wand*, *Decke* und *Tisch*. Die einfache Strukturiertheit bedeutet, dass nur einsegmentige Objekte untersucht werden. Strukturspezifisch komplexere Objekte werden im Abschnitt 3.5, *Strukturbasierte Objekterkennung*, behandelt. Aber auch bei einfach strukturierten Objekten spielen Beziehungen zu anderen Objekten ein Rolle. Die Tabelle 3.1 zeigt die auf diese Weise modellierten Objekte. In der ersten Spalte steht die Objektkategorie, in der zweiten die Affordanz, oder anders

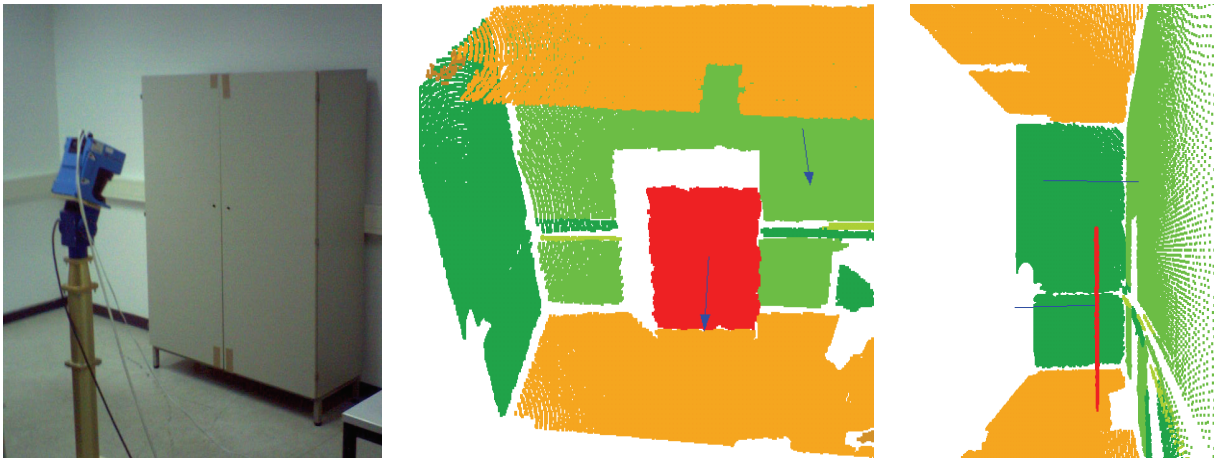
ausgedrückt, deren funktionale Beschreibung. In der letzten Spalte stehen die aus der funktionalen Beschreibung abgeleiteten physikalischen Merkmale. Für die Objekte *Fußboden*, *Wand*, *Decke* und *Tisch* sind diese direkt aus der funktionalen Beschreibung herleitbar. Anders verhält es sich beim Schrank, denn dessen wesentliche Funktion, nämlich die der Aufbewahrung, lässt sich ohne Interaktion so nicht eindeutig feststellen. Insofern bilden in diesem Fall die Merkmale in besonderer Art keine hinreichenden, sondern nur notwendige Hinweise für die Kategorie Schrank. Dies deckt sich jedoch mit den Erfahrungen, die auch wir Menschen machen, wenn wir beispielsweise Schranktüren in ungewohnter Umgebung untersuchen. Bevor wir beispielsweise bei einer Schrankwand nicht die Tür öffnen, bleibt unklar, ob sich dahinter der erwartete Regalboden, die Kleiderstange oder ein verblendeter Wartungsschacht befindet. Die Form spielt an dieser Stelle eine stark untergeordnete Rolle. Durch den Segmentierungsschritt wird sichergestellt, dass es sich um hinreichend flache Elemente handelt. In den Abbildungen 3.7 wird gezeigt, wie die beschriebenen Merkmale von dem System wahrgenommen werden. In Abb. 3.7(a) werden rechts die segmentierten Raumbegrenzer der links gezeigten Szene aufgeführt. Beim Boden ist zu sehen, dass aufgrund von Verdeckungen eine starke Zerklüftung stattfindet. Bei den Wänden ist festzuhalten, dass das Rollo nicht von den anderen Wänden unterschieden werden kann. Bei der Decke sind die Aussparungen für Lampe bzw. Rauchmelder deutlich zu erkennen. Beim Tisch in Abb. 3.7(b) sind ebenfalls viele Löcher aufgrund von Verdeckungen zu erkennen. Im Bereich der hinteren linken Tischkante führt dies dazu, dass über einen großen Bereich keine Punkte der Tischplatte zugeordnet werden konnten. Des weiteren fällt auf, dass aufgrund eines schleifenden Winkels zwischen Scanebene und Tischoberfläche der Abstand zwischen benachbarten Scanlinien im hinteren Teil stark anwächst. In der mittleren Abbildung ist der als Merkmal benutzte Normalenvektor des Tisches eingezeichnet. In der Abb. 3.7(c) ist gezeigt, wie die einfache Flächeninformation auch zur Detektion von Schränken genutzt werden könnte. Hier spielten dann die Ausrichtung von Schrank und dahinter befindlicher Wand eine Rolle.

3.3.1 Bayessches Netz als Objektmodell

Das Modell einer Objektkategorie wird durch kontinuierliche Bayessche Netze beschrieben, da diese sich besonders zum Modellieren von unsicherem Wissen eignen (vgl. Anhang A.4, *Bayessche Netzwerke*). Jede Objektklasse korrespondiert zu einem einzelnen

(a) Raumbegrenzungen

(b) Tisch



(c) Schrank

Abbildung 3.7: Funktionalbasierte Erkennung einfach strukturierter Objekte

Objekt-kategorie	**Affordanz**	**Physikalische Eigenschaften**
Fußboden	erlaubt das eben darauf gehen	• Horizontale Ausrichtung • Höhenbereich • Mindestfläche
Decke	schließt den Raum eben nach oben ab	• Horizontale Ausrichtung • Mindestabstand vom Boden • Mindestfläche
Wand	schließt den Raum zur Seite hin ab	• Vertikale Ausrichtung • Höchstabstand zur Decke • Mindestfläche
Tisch	erlaubt das darauf Stellen von Gegenständen	• Horizontale Ausrichtung • Höhenbereich • Mindestfläche
Schrank	erlaubt das Verstauen von Gegenständen in einem abgeschlossenen Raum	• Parallele Ausrichtung zur Wand • Mindestabstand zur Decke • Abstand zur Wand • Mindestfläche

Tabelle 3.1: Einfach strukturierte Objekte

Bayesschen Netz, das durch die Werte für Mittelwert und Varianz von spezifischen Merkmalen parametrisiert wird. Die für die Raumbegrenzungen benötigten physikalischen Attribute sind hierbei:

- Fläche f, $[f] = m^2$
- Höhe h, $[h] = m$
- Horizontale Ausrichtung ϕ, $[\phi] = °$
- Vertikale Ausrichtung ϕ, $[\phi] = °$
- Abstand zur Decke d, $[d] = m$

Diese Werte können entweder analytisch durch die Verwendung von Trainingsmengen bestimmt werden oder durch die Verwendung von Expertenwissen. Dies ist auch deshalb unkompliziert, da die Merkmale alle physikalisch direkt interpretierbaren Größen entsprechen. Da davon ausgegangen wird, dass die einzelnen Parameter stochastisch unabhängig voneinander sind, reicht es aus, einfach den Mittelwert und die Varianz der einzelnen Parameter zu berechnen. Letztlich besteht das Modell aus einer n-dimensionalen Gaussschen Dichtefunktion. Um die Wahrscheinlichkeitswerte zu erhalten, müsste eigentlich diese über das entsprechende Konfidenzintervall integriert werden. Dieses Vorgehen hätte jedoch einige Nachteile, wie beispielsweise einen sehr großen rechnerischen Aufwand für die numerische Integration oder die Bestimmung passender Intervallgrenzen. Ein weiterer Nachteil läge in der Tatsache, dass für kleine Wahrscheinlichkeiten die Multiplikation von Termen der n-dimensionalen Gaussdichte zu Wahrscheinlichkeitswerten von (nahezu) Null führen würde [Fin03]. Eine Lösung dieses Problems ist die Verwendung der sog. negativen logarithmischen Wahrscheinlichkeitsfunktion (*engl.* negative logarithmic likelihood function), wobei $\mathcal{N}$ die Normalverteilung darstellt:

$$\begin{aligned} nLL(x) &= -\ln \mathcal{N}(x|\mu,\sigma^2) \\ &= -\ln \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(x-\mu)^2}{2\sigma^2} \\ &= -\ln \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} - \ln \exp -\frac{(x-\mu)^2}{2\sigma^2} \\ &= -\ln 1 + \frac{1}{2}\ln(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \\ &= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \end{aligned}$$

Somit kann die n-dimensionale Dichtefunktion als Summe von Kosten interpretiert werden. Wenn das Modell einem Modell von n stochastisch unabhängigen normalverteilten eindimensionalen Zufallsvariablen entspricht (vgl. Anhang A.3.2 auf S. 121), führt dies zu:

$$\begin{aligned} nLL(\mathbf{x}) &= -\ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\sigma}^2) \\ &= -\ln \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} \\ &= \frac{1}{2}\sum_{i=1}^{n} \ln(2\pi\sigma_i^2) + \sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{2\sigma_i^2} \end{aligned}$$

Damit ist die negative logarithmische Wahrscheinlichkeitsfunktion weniger empfindlich gegen einzeln verschwindende Terme. Auf der anderen Seite führen kleine Wahrscheinlichkeitswerte zu sehr großen Summanden bei den Kosten. Um dies zu beschränken, wird eine obere Schranke für die Kosten eingeführt, dieses Vorgehen nennt man „flooring“. Somit können einzelne Ausreißer einzelner Features nicht zum sofortigen Ausschluss der Kategorie führen. Ein Problem entsteht jedoch dann, wenn die Modelle, die miteinander verglichen werden, unterschiedliche Dimension, d.h. Anzahl von Parametern, haben. Um dies zu bereinigen, bietet sich die Verwendung von Wahrscheinlichkeitsverhältnissen an. Dabei wird der Wahrscheinlichkeitswert ins Verhältnis gesetzt zur Wahrscheinlichkeit einer angenommenen Gleichverteilung. Übertragen auf die logarithmische Wahrscheinlichkeitsfunktion, bedeutet dies, dass der zur 50% Schranke gehörende nLL-Wert abgezogen wird.

$$r_{LL} = nLL(\mathbf{x}) - nLL(\mathbf{x}_{50\%})$$

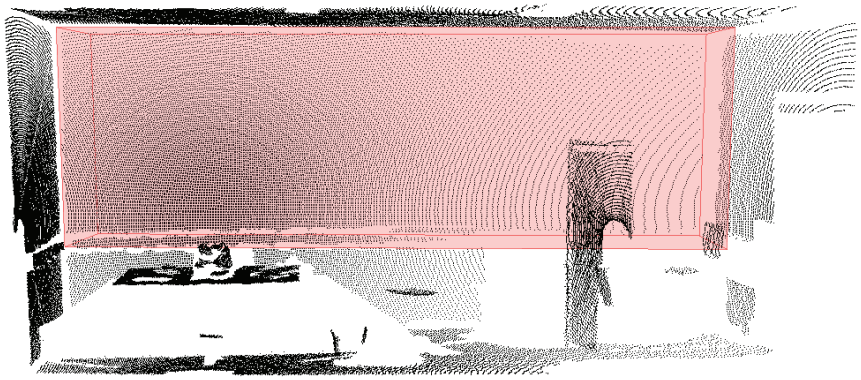
Diese standardisierten Werte sind dann besser und direkt miteinander vergleichbar. Um einen einfacher interpretierbaren numerischen Wert zu erhalten, werden die Werte für r_{LL} mittels einer sigmoidalen Funktion auf den Bereich [0-1] gespreizt.

$$val = \frac{1}{1 + \exp\left(-c \cdot r_{LL}\right)}\,, \qquad c \neq 0$$

Für den Sonderfall, dass ein Datenpunkt $\mathbf{x}$ gerade einem Punkt auf der 50%-Schranke entspricht, folgt wegen $nLL(\mathbf{x}) = nLL(\mathbf{x}_{50\%})$, dass $r_{LL} = 0$ und damit $val = 50\%$.

3.3.2 Interaktives Lernmodul

Wie im vorherigen Abschnitt bereits erwähnt, können die hier extrahierten Merkmale alle auch von einem Menschen direkt interpretiert werden. Somit ist es möglich, neue Objekte und die sie bestimmenden Merkmale durch einen Experten vorzugeben. Umgekehrt wünscht man sich von kognitiven Systemen aber, dass sie ihre Sicht auf die Welt erweitern beziehungsweise anpassen. Dies sollen die Systeme möglichst selbständig in einem Lernprozess erreichen. Es sind dabei grundsätzlich zwei Ansätze möglich: Modell lernen oder Parameter lernen. Beim Modell lernen muss gelernt werden, welche Parameter signifikant sind und wie sie wirken bzw. voneinander abhängen. Beim Parameter lernen ist das Modell, also die Menge der Merkmale und ihrer Beziehungen vorgegeben,



(a) Ausgewähltes und markiertes Wandsegment

	ID	Relation / Property	toID	Type	Subpart
1	491	deviation_horizontal	---	plane (size: 12.5333)	
2	491	size	---	plane (size: 12.5333)	
3	491	minimal_distance_ceiling	---	plane (size: 12.5333)	

(b) Zugeordnete Eigenschaften des Wandsegments

Abbildung 3.8: Lernen geometrischer Eigenschaften zur funktionalitätsbasierten Objekterkennung

und nur ihr Einfluss muss berechnet werden. Bei dem hier benutzten Lernansatz wird das Modell als gegeben vorausgesetzt. Damit behält der Experte die Kontrolle über das Modell, wird aber bei der Analyse der Werte unterstützt. Ein weiterer Vorteil dieses Vorgehens ist, dass weniger Trainingsdatensätze als beim vollständigen Lernansatz benötigt werden.

Das entwickelte Lernmodul arbeitet auf bereits segmentierten Oberflächen. Der Benutzer kann Segmente durch eine graphische Benutzerinteraktion markieren (vgl. Abb. 3.8(a)) und benennen[46], d.h. einer Klasse zuordnen, und gegebenenfalls Beziehungen zu anderen Objekten angeben, wie es bei mehrfach strukturierten Objekten notwendig wird. Die Bekanntmachung der Eigenschaften erfolgt über ein Eingabefeld, wie es in Abb. 3.8(b) für ein Wandsegment zu sehen ist. Die zur Modellierung der Raumbegrenzungen und des Tisches verwendeten abgeleiteten physikalischen Größen sind:

- ***d**eviation_**h**orizontal*
  Abweichung der Normalen von der Horizontalen

- ***d**eviation_**v**ertical*

[46] *engl.* label

	dh	***dv***	***s***	***h***	***mdc***
Fußboden		×	×	×	
Decke		×	×	×	
Wand	×		×		×
Tisch		×	×	×	

Tabelle 3.2: Merkmale von Raumbegrenzungen und Tisch

Abweichung der Normalen von der Vertikalen

- ***s****ize*
  Fläche

- ***h****eight*
  Höhe

- ***m****inimal_****d****istance_****c****eiling*
  Kleinster Abstand zur Decke

Die jeweils für die hier betrachteten Objekte verwendeten Merkmale sind in Tabelle 3.2 aufgeführt. Wie im Abschnitt 3.3.1 beschrieben, geht das Modell von normalverteilten Merkmalsparametern aus. Für einige Parameter, wie beispielsweise das Merkmal Höhe bei einem Tisch, ist dies hinreichend zutreffend und der berechnete Mittelwert und der Idealwert koinzidieren. Anders verhält es sich jedoch beispielsweise bei dem Merkmal *Abstand zur Decke*, das beim Wandmodell verwendet wird. Der ideale Abstand einer Wand zur Decke ist Null, das heißt beide Elemente berühren sich. Bei einer statistischen Auswertung ergibt sich allerdings eine davon mehr oder weniger starke Abweichung für den Mittelwert. Die wesentlichen Ursachen dafür sind, dass Objekte häufig unvollständig segmentiert sind oder im Übergangsbereich bauliche Elemente den Übergang verblenden. In diesen Fällen wird deshalb bei der späteren Anwendung, ausgehend von diesem Wissen über die tatsächliche geometrische Situation, ein Wert, der zwischen dem statistisch wahrscheinlichsten und dem theoretisch idealen Wert liegt, auf den wahrscheinlichsten Wert abgebildet. Somit wird einerseits vermieden, dass beispielsweise eine Wand, die ideal bis an die Decke heranreicht, einen schlechteren Wert erreicht als eine, die weiter entfernt ist, und andererseits wird dennoch die statistische Häufung berücksichtigt. Für diese abgeleiteten Mindest- oder Maximalgrößen gilt also für den Erkennungsvorgang,

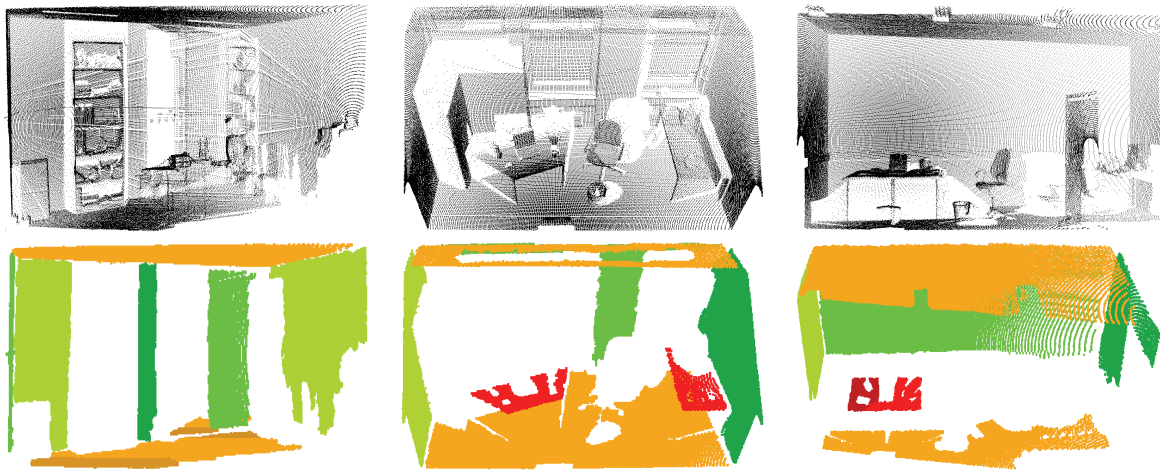
dass ab oder bis zu einem bestimmten Wert keine Veränderung der Klassifizierungsgüte mehr stattfindet. Dies gilt, wie bereits gesagt, beispielsweise für das Merkmal Abstand zur Decke und in umgekehrter Weise beispielsweise für das Merkmal Flächeninhalt einer Wand. Letzterer Typ von Merkmalen hat die Eigenschaft, dass er sehr unspezifisch ist. Eine Mindestgröße ist aber notwendig, um kleinere Artefaktsegmente ausschließen zu können. Da die statische Verteilung dieser Werte sehr unspezifisch ist, wird an dieser Stelle Expertenwissen eingebracht, um den entsprechenden Parameter geschickt zu wählen.

3.3.3 Anwendung: Raumbegrenzungen

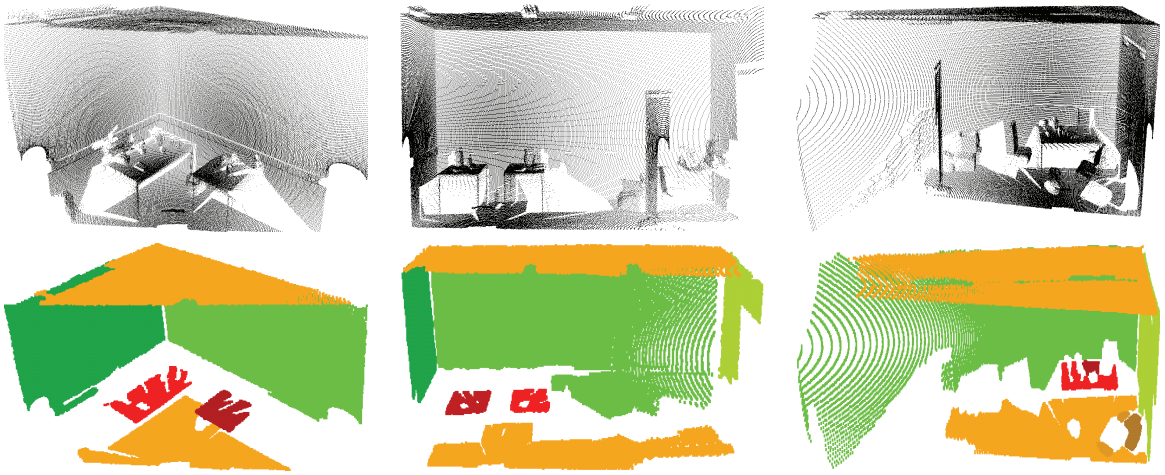
Das Lernen der Merkmale geschieht in zwei Schritten. Zunächst wird mit dem vorgestellten Lernmodul ein Segment einer bestimmten Objektklasse zugeordnet, und die benötigten Eigenschaften werden ausgewählt (vgl. Abb. 3.8(b)). Dann werden diese geometrischen Eigenschaften, wie beispielsweise die Normalenabweichung, bestimmt und alle in einem XML-File abgelegt[47]. Dies wird für eine Reihe von Aufnahmen der gleichen Objektklasse wiederholt. Zum Abschluss werden die erhaltenen Werte statistisch analysiert, d.h. Mittelwert und Abweichung berechnet. In der Abb. 3.9 sind die 3-D Aufnahmen und die extrahierten Flächensegmente dieser Anwendung aufgezeigt. Es handelt sich um eine Reihe von unterschiedlichen Innenraumszenen, die aus verschiedenen Perspektiven aufgenommen wurden. Die Decken- und Bodenelemente wurden in einem Ockerton, die Wände in Grüntönen und die Tische in Rottönen eingefärbt. Um die einzelnen Segmente besser hervorzuheben, wurden benachbarten Segmenten unterschiedliche Farbtöne zugeordnet. Es ergaben sich daraus 13 Parametertupel für die Klasse Tisch und 28 Parametertupel für die Klasse Wand. Für die Klasse Tisch ist in dem rechten Diagramm von Abb. 3.10 der Parameter Höhe gegen den Parameter Vertikale Normalenabweichung aufgetragen. Es zeigt sich, dass die Werte für die Tischhöhe, obgleich es sich um verschiedene Tischmodelle handelt, innerhalb eines 25 mm großen Intervalls liegen[48]. Die Abweichung des durch die segmentierte Tischfläche definierten Normalenvektors beträgt $\sigma = 0.4°$. Was die Fläche betrifft, können, wie im vorherigen Abschnitt bereits erwähnt, die gemessenen Daten nur indirekt zur Parameterbestimmung herangezogen werden. Verursacht durch

[47] Ein für ein mehrfachstrukturiertes Objekt erstelltes XML-File findet sich im Anhang A.5 auf Seite 131.

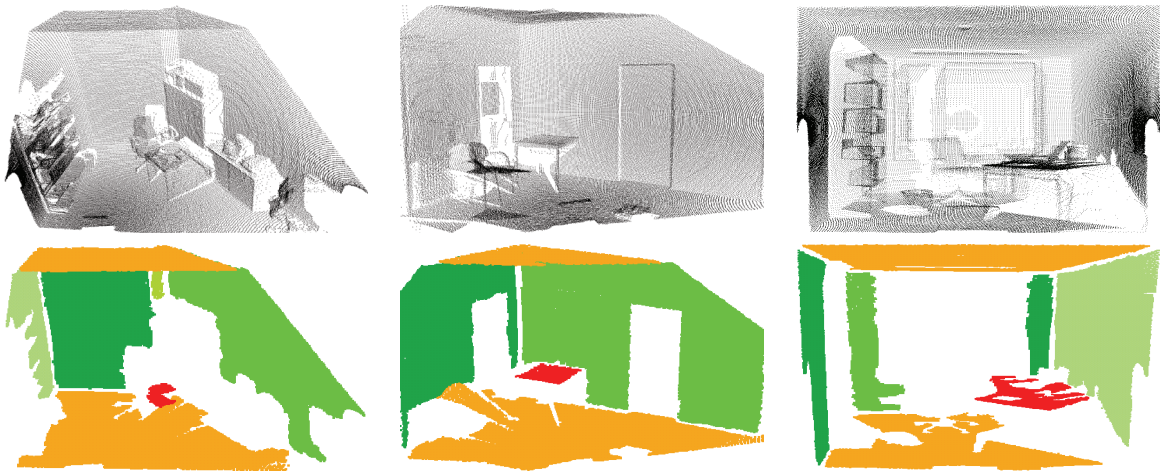
[48] Die negativen Höhenwerte ergeben sich aus den im Scanner liegenden Ursprung.



(a) Datensätze #0272 #0259 #0232



(b) Datensätze #0220 #0223 #0308



(c) Datensätze #0092 #0095 #0273

Abbildung 3.9: Segmentierte Innenraumszenen

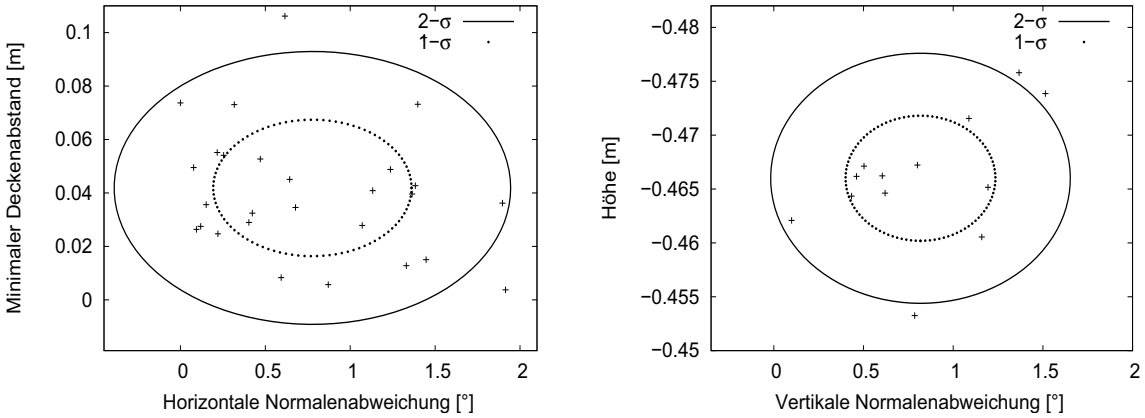


Abbildung 3.10: Merkmalsverteilung bei Wänden (links) und Tischen (rechts)

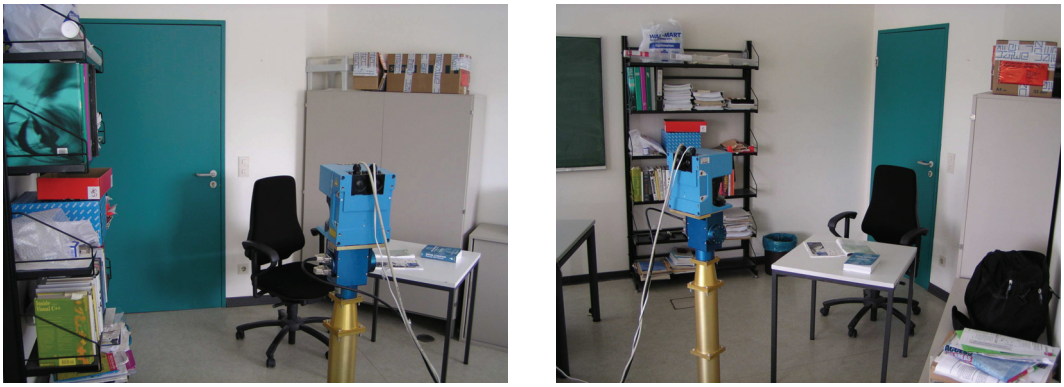
die bei Realaufnahmen vorkommenden Verdeckungen bzw. Verschattungen, schwanken die Messungen stark. Entsprechend unspezifisch müssen die Werte für diesen Parameter gesetzt werden. Ein ähnliches Bild ergibt sich bei der Interpretation der Werte von Wänden. Die Abweichung des Normalenvektors von der Horizontalen beträgt hier $\sigma = 0.6°$ und ist damit auch noch sehr spezifisch. Der mittlere minimale Abstand zur Decke beträgt etwas über 42 mm. Auch hier ist der bei der Erkennung verwendete Wert für das Merkmal Fläche sehr unspezifisch, was auch bei den Aufnahmen in Abb. 3.9 zu sehen ist. Dennoch wird die Verwendung einer Mindestgröße aus den bereits erwähnten Gründen notwendig. Die Sicherstellung einer hinreichenden Güte der extrahierten Raumgrenzen wird insbesondere deshalb wichtig, weil deren Position in den folgenden Arbeitsschritten zur weiteren Datenreduktion und Segmentierung verwendet wird.

3.4 Registrierung

Die Erkennung von Objekten, wie den hier gezeigten Raumbegrenzungen, wird nicht nur zu deren Identifizierung, sondern auch darüberhinausgehend genutzt werden. Im Abschnitt 3.6, *Segmentierung auf funktionalen Ebenen*, werden diese Objekte als Grundlage einer ebenen- und dichtebasierten Segmentierung verwendet. Sie lassen sich darüber hinaus aber auch dazu nutzen, eine Registrierung zweier oder mehrerer Raumansichten zu vollziehen. Hierzu eignen sich die Raumbegrenzungen besonders gut, da sie nicht bewegliche Teile einer Szene darstellen und deshalb räumlich und zeitlich fix sind. Ziel

eines solchen Vorhabens ist es, Objekte von mehreren Seiten aufzunehmen und diese Aufnahmen zu einem möglichst kompletten Ganzen zusammenzufassen. Ein häufig verwendetes Verfahren zur Registrierung von Punktwolken ist der sog. ICP-Algorithmus (Iterative Closest Point). Bei diesem Verfahren wird schrittweise die Orientierung zwischen beiden Punktwolken so verändert, dass eine Gütefunktion, die aus der Summe der Abstände der Testpunkte zum jeweils nächstgelegenen Punkt im Referenzbild gebildet wird, optimiert wird. Dabei können verschiedene Variationen vorgenommen werden, um den Erfolg des Vorgangs zu verbessern [DF01], [RL01]. Bei dem hier gewählten Ansatz wird die Registrierung nicht auf Punktebene, sondern auf Merkmalsebene durchgeführt. Dies hat den Vorzug, dass nicht jeder beliebige Punkt Einfluss auf das Registierungsergebnis haben kann. Andererseits ist dieser Vorgang mit einem Mehraufwand, was die Merkmalsextraktion betrifft, verbunden.

Der Algorithmus wird am Beispiel einer Innenraumszene demonstriert, bei der die Wände nicht orthogonal angeordnet sind. In Abbildung 3.11(a) ist auf der linken Seite die linke Aufnahmeposition und auf der rechten Seite die rechte Aufnahmeposition dieser Szene dargestellt. Im ersten Arbeitsschritt werden alle vertikalen Ebenen in beiden Aufnahmen berechnet. Je zwei dieser Ebenen werden für die Registrierung verwendet. Zunächst werden die Ebenen bezüglich ihrer horizontalen Position geordnet und geordnete zweielementige Teilmengen gebildet. Im ersten Schritt der Registrierung wird der erste Datensatz so verschoben, dass die erste Wand aus dem ersten Datensatz mit der ersten Wand aus dem zweiten Datensatz zur Überdeckung kommt. Das Ergebnis dieses Zwischenschritts ist in Abb. 3.11(b) im ersten Bild zu sehen. Die übereinander liegenden Wände sind diejenigen ganz rechts, wobei aufgrund der Betrachterposition nur ein schmales Band zu sehen ist. Die grünen Punkte gehören hierbei zur linken Aufnahme, die roten entsprechend zu rechten. Die für die Registrierung verwendeten Wandelemente sind im Gegensatz zu den anderen etwas heller dargestellt. Die schwarzen Striche bilden jeweils ein Koordinatenkreuz, dessen Ursprung im Scanner liegt. Der zweite Datensatz wird nun soweit parallel zu den im ersten Schritt verwendeten Wänden verschoben, bis auch die anderen beiden Wände zur Überdeckung kommen. Das Ergebnis ist in Abbildung 3.11(b) im mittleren Bild zu sehen. Das gleiche Registrierungsergebnis, allerdings mit allen Punkten, ist im rechten Bild zu sehen. Dieser Vorgang wird für alle gebildeten Kombinationen der aus jeweils zwei Wänden bestehenden Teilmengen durchgeführt. Gewinner ist jene Kombination, bei der die meisten Wände nach der Registrierung zur Überdeckung kommen. Überdeckung bedeutet, dass der Schwerpunkt der einen Wand



(a) Linke und rechte Aufnahmeposition

(b) Segmentierungszwischenschritte und Segmentierungsergebnis

Abbildung 3.11: Registrierung mit Wandsegmenten

einen maximal zulässigen Abstand zur andern Wand nicht überschreitet. Bei gleicher Überdeckungszahl kommt der minimale Gesamtabstand zum Tragen.

3.5 Strukturbasierte Objekterkennung

Im Abschnitt 3.3, *Erkennung einfach strukturierter Objekte*, wurden Objekte untersucht, die aus nur einem Segment bestehen. Zur Erkennung wurden geometrische Eigenschaften verwendet, die mit der Funktionalität des Objekts korrespondieren. Eine wesentliche Eigenschaft der Oberfläche wurde durch den vorgeschalteten Segmentierungsschritt sichergestellt: Nur hinreichend homogene Oberflächen, die bei all den untersuchten Klassen Voraussetzung sind, werden betrachtet. In diesem Abschnitt wird mit dem Stuhl nun ein Objektbeispiel betrachtet, das sowohl eine kompliziertere Struktur als auch ei-

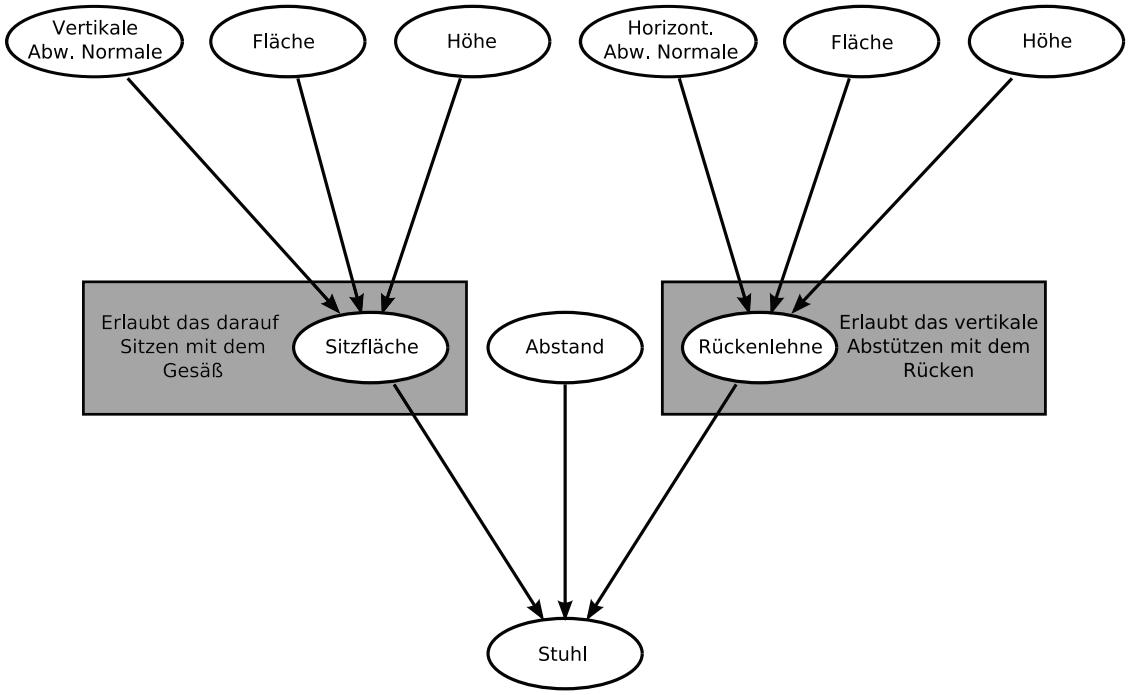


Abbildung 3.12: Strukturelles und funktionales Stuhlmodell

ne vielfältigere Funktionalität (vgl. Minskys Stuhlmodell in Abschnitt 2.7 auf Seite 25) besitzt. Das hier verwendete Modell eines Stuhls (vgl. Abb. 3.12) besitzt, was die Funktionalität betrifft, die Affordanz der Sitzmöglichkeit und die Affordanz der Rückenabstützung. Geometrisch korreliert dies mit hinreichend ebenen und großen Flächen mit entsprechender Ausrichtung in bestimmter Höhe, die hinreichend nahe zueinander positioniert sind. Die Basis des Stuhls spielt bei diesem Modell keine Rolle, zumal sie von der verwendeten Aufnahmeeinheit gewöhnlich nicht eingesehen werden kann. In der Abb. 3.13 wird das konkrete Vorgehen bei der Merkmalsextraktion am Beispiel dreier Stühle graphisch demonstriert. In der obersten Reihe befinden sich Bilder, wie sie von einer externen Kamera, aber aus ähnlicher Perspektive wie die Tiefenaufnahme, gemacht wurden. Die Darstellungen der Reihen zwei und drei vereinen die Resultate unterschiedlicher Verarbeitungsschritte auf den 3-D Punktewolken. Die zweite Reihe nimmt dabei eine Perspektive von links oben ein, die dritte Reihe von der Seite. Die Änderungen der Perspektive beruhen hierbei ausschließlich auf der Darstellungsänderung, der 3-D Datensatz bleibt der gleiche. Die grauen Punkte stellen nicht-segmentierte Punkte dar, die von den folgenden Segmentierungsergebnissen überdeckt werden. Die roten Punk-

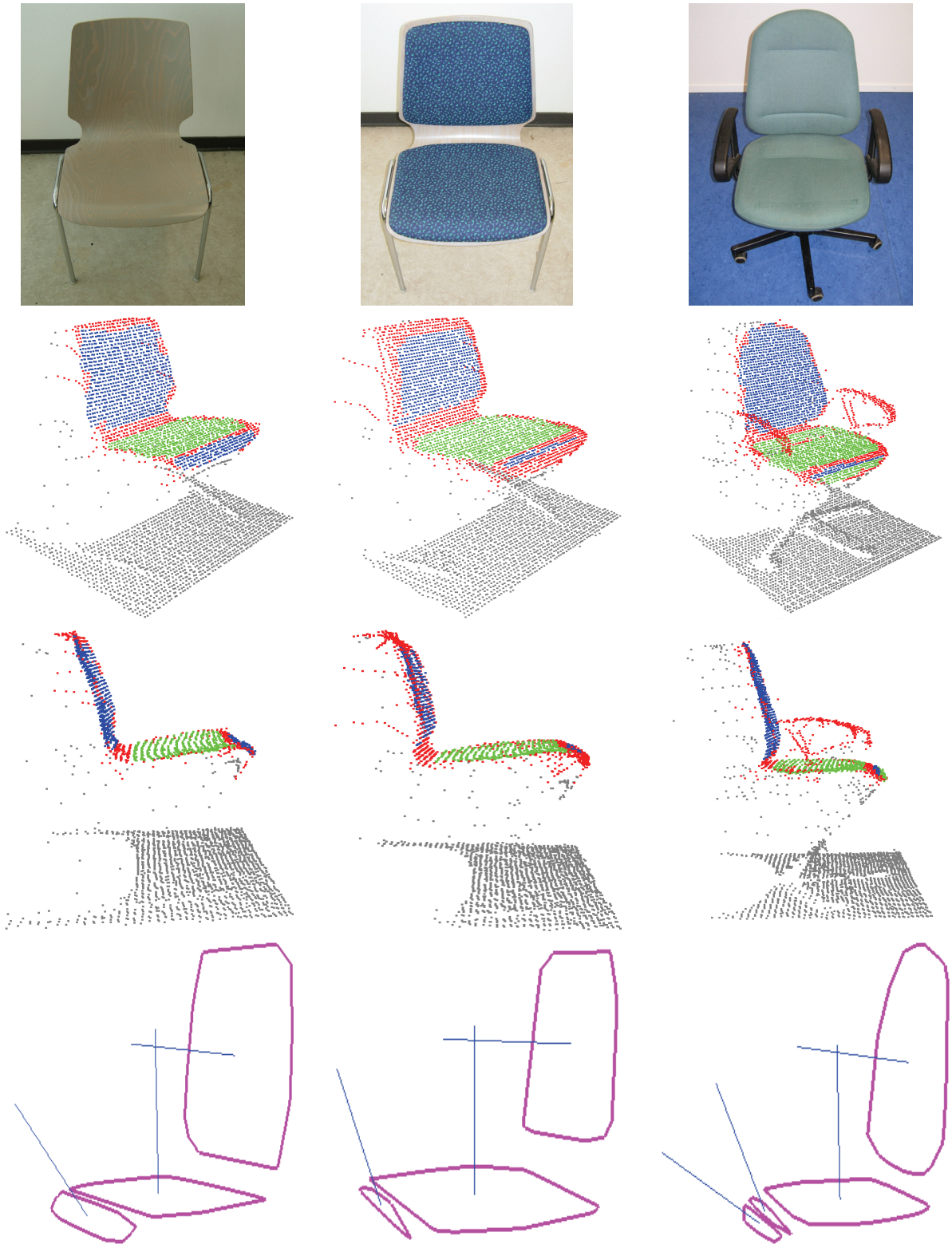
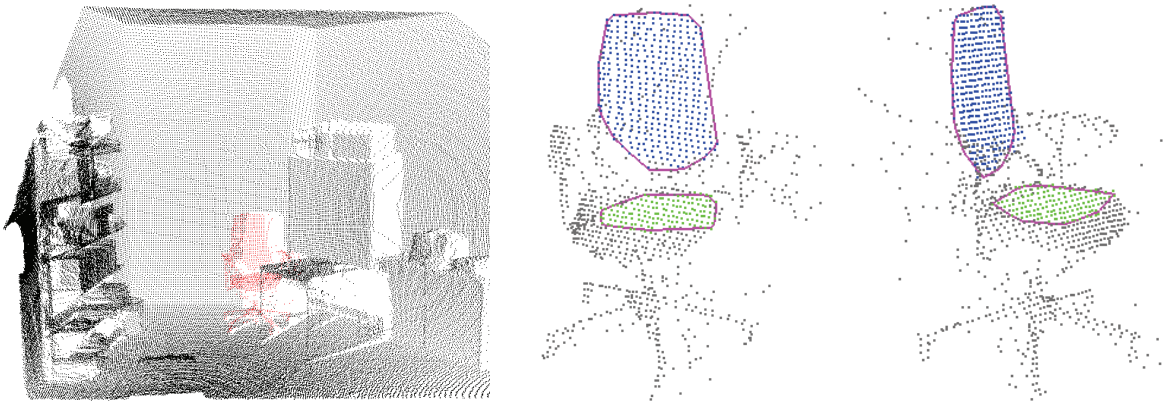


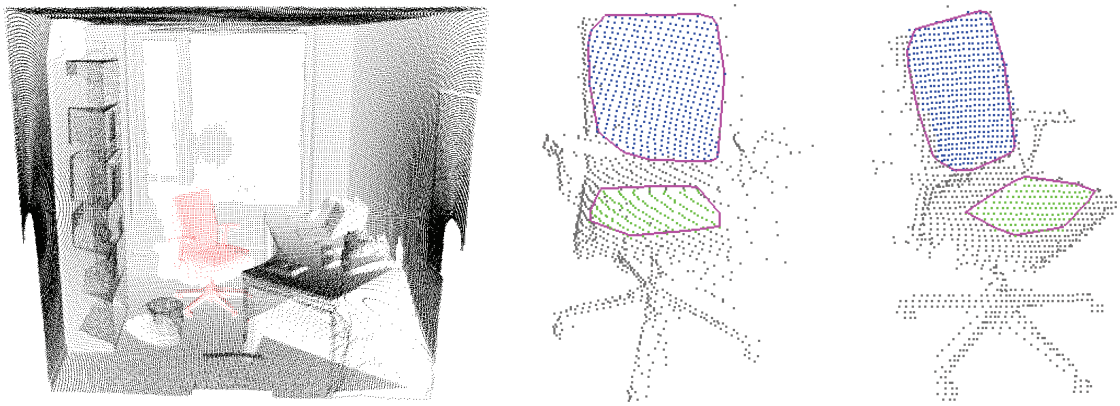
Abbildung 3.13: Segmentierte Stühle mit Zuordnungsgüten von 99.7%, 99.8% und 98.9%

te entsprechen dem Ergebnis eines Segmentierungsschritts, der erst im Abschnitt 3.6, *Segmentierung auf funktionalen Ebenen*, beschrieben wird und bei dem isoliert stehende Objekte aus ihrer Umgebung herausgelöst werden. Dieses Segmentierungsergebnis wird durch die Ebenensegmente überdeckt, die in den Farben Grün und Blau dargestellt sind. Es ist zu sehen, dass bei allen Stühlen die Sitzfläche und Rückenlehne ordentlich erfasst werden. Im vorderen Sitzbereich kommt es zur Übersegmentierung, die ihren Ursprung in den dort stark gewölbten Flächen und der speziellen Ausrichtung von Scanner zu Stuhlobjekt hat. Bei allen Objekten wird das Gestell nur unzureichend erfasst, was wie bereits erwähnt, für das verwendete Modell aber nicht relevant ist. In der letzten Reihe sind die aus den Ebenen berechneten Merkmale, wie sie zur Erkennung verwendet werden, dargestellt. Die rosafarbenen Polygone umschreiben die segmentierten Flächen, die blauen Strecken stellen die Flächennormalen dar, mit denen die Abweichungen zur vertikalen oder zur horizontalen Ebene berechnet werden. Konkret werden dem Modul eine Menge von Ebenensegmenten übergeben, die ihren gemeinsamen Ursprung in dem vorgeschalteten Segmentierungsschritt haben. Daraus werden, entsprechend dem Stuhlmodell, das aus den Elementen Rückenlehne und Sitzfläche besteht, die möglichen Kombinationen gebildet und deren jeweilige Güte berechnet. Es werden dabei also auch die Permutationen mit den kleinen Segmenten im vorderen Sitzbereich untersucht. Dies ist vom Rechenaufwand jedoch vernachlässigbar und macht dafür eine Vorklassifikation, die überprüft, welche der aktuellen Teilelemente welche Funktion übernehmen könnte, überflüssig. Für die in Abb. 3.13 gezeigten Stühle ergeben sich Zuordnungsgüten zur Klasse Stuhl von 99.7%, 99.8% und 98.9%. Bei den Kombinationen mit den im vorderen Sitzbereich auftretenden kleineren Segmenten ergeben sich verschwindende Wahrscheinlichkeiten. Für die ersten Stuhlexemplare wird eine eindeutige Zuordnung in die Kategorie Stuhl erreicht. Beim Bürostuhl kommt es durch die recht große Rückenlehne zu einer stärkeren Abweichung vom Modell, die für größere Kosten und damit eine Abwertung sorgt. Da die anderen Parameter aber gut zum Modell passen, wird dennoch eine eindeutige Zuordnung erreicht.

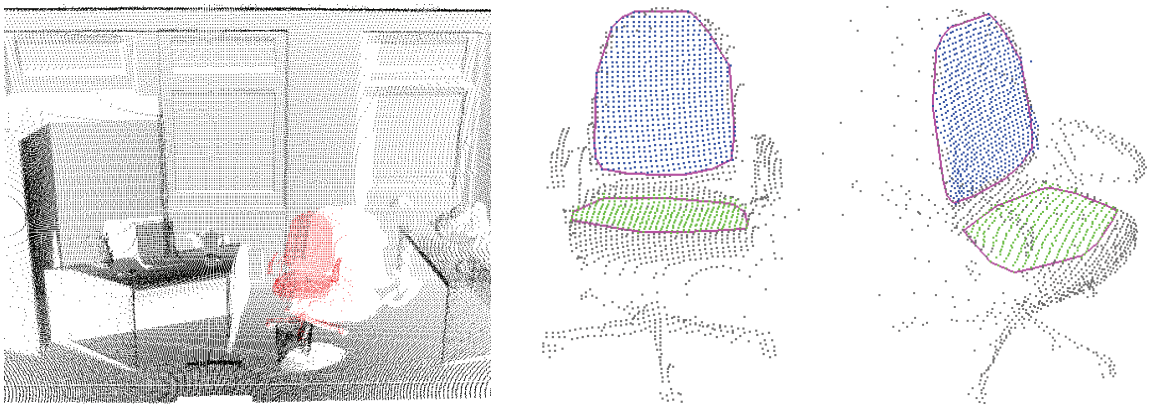
Waren die zuvor gezeigten Stühle aus einem jeweils frontalen Aufnahmewinkel aufgenommen, folgen nun Beispiele bei denen die Stühle beliebig in der Szene positioniert sind. In der linken Spalte von Abb. 3.14 ist der Stuhl jeweils rot markiert, in der mittleren und rechten Spalte sind die Segmentierungsergebnisse aus zwei verschiedenen Perspektiven dargestellt. Der Stuhl in Abb. 3.14(a) erreicht eine Erkennungsgüte von 70.6%. Gründe für diese schwächere Zuordnung sind die senkrecht gestellte Rückenlehne und



(a) Datensatz #092, Erkennungsgüte 70.6%



(b) Datensatz #273, Erkennungsgüte 76.5%



(c) Datensatz #259, Erkennungsgüte 99.9%

Abbildung 3.14: Stühle innerhalb einer Szene

die reduziert erkannte Sitzfläche bei großer Rückenlehnenfläche. Letztere Punkte sind es auch, die bei dem Stuhl in Abb. 3.14(b) zu einer erhöhten Kostensumme und einem damit eher schlechteren Klassifizierungsergebnis führen. Die Verschattung der Sitzfläche und die starke Wölbung im vorderen Bereich erwirken eine nur klein detektierte Sitzfläche. Tendenziell gestärkt wird dieses abschlägige Verhalten durch die relativ große Entfernung zwischen Scanner und Objekt, die dazu führt, dass Fehlerpunkte bzw. nichtmitsegmentierte Punkte durch die dünnere Punktdichte stärker einfließen. Der ideale Wert der Zuordnung des Stuhls in Abb. 3.14(c) ist darauf zurückzuführen, dass alle Werte mit Ausnahme der Fläche der Rückenlehne, sehr gut passen. Deren negativer Einfluss wird aber durch die besondere Güte der anderen Parameter gedämpft.

Wie bei vielen anderen Methoden auch, hängt der Erfolg vom erfolgreichen Abschluss der Vorverarbeitungsschritte ab. Konkret bedeutet dies, dass Störungen infolge einer Über- oder Untersegmentierung oder Verdeckungen die Erkennungsfähigkeit stark einschränken können.

Anstelle der Untersuchung einer größeren Messreihe mit verschiedenen Stuhlobjekten wird das Verhalten der Erkennungsleistung auf Parameteränderungen hin gezeigt. Dazu sind in Tabelle 3.3 die erhaltenen Erkennungsgüten für die Variation der Parameter Sitzfläche und Normalenabweichung bei Beibehaltung der restlichen Parameter aufgezeigt. Es ist zu sehen, dass sich die Werte wie erwartet kontinuierlich ändern. Es deutet sich auch an, dass diese Art der Modellierung mit negativ logarithmischen Wahrscheinlichkeitswerten bei wachsender Parameterzahl einzelne Ausreißer in ihrer potentiell ausschließenden Wirkung dämpft.

Neben der Verwendung von Ebenensegmenten und ihren Relationen können auch weitere Merkmale in die Objektklassifikation einfließen. Eine Möglichkeit ist beispielsweise, die einfache, aber unter bestimmten Randbedingungen robust detektierbare Information über die Ausdehnung eines Objekts zu dessen Klassifikation heranzuziehen, wie sie in [WR06] eingesetzt wurde. Somit kann man auf einfache Weise gewisse Objektkategorien ausschließen oder validieren. Dies kann dann dazu genutzt werden, von der sensorischen Erfassung her gesehen, nur sehr unspezifisch charakterisierbare Objekte zu modellieren. Die Objektklasse Regal ließe sich beispielsweise, wenn auch nicht direkt funktional motiviert, anhand ihrer Ausdehnung und Orientierung, die über eine objektorientierte Bounding-Box und dem Nichtvorhandensein größerer ebener Flächen an der Front erhalten werden können, umsetzen.

	0.12	0.14	0.16	0.18	0.20	0.22	0.24	0.26
12.0	100.0	100.0	99.9	99.4	87.1	9.2	0.1	0.0
14.0	100.0	100.0	99.9	99.3	84.2	7.4	0.0	0.0
16.0	100.0	100.0	99.9	98.6	72.6	3.8	0.0	0.0
18.0	99.9	99.9	99.6	95.6	13.9	1.2	0.0	0.0
20.0	99.6	99.4	97.8	80.1	1.9	0.2	0.0	0.0
22.0	97.2	95.6	84.2	34.1	0.2	0.0	0.0	0.0
24.0	72.6	62.5	29.1	3.8	0.0	0.0	0.0	0.0
26.0	11.3	7.4	1.9	0.2	0.0	0.0	0.0	0.0
28.0	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0
30.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabelle 3.3: Erkennungsverhalten bei Variation der Parameter: Neigungswinkel gegen Flächeninhalt

3.6 Segmentierung auf funktionalen Ebenen

Bei dem zuvor vorgestellten Segmentierungsverfahren werden ebene Flächen detektiert (vgl. Abschnitt 3.2, *Oberflächensegmentierung*). Kriterium dort ist die relative Positionierung der Punkte zu ihrer Umgebung. Die derart detektierten Flächen werden dazu genutzt, (funktional) einfache Objekte wie Tische oder Böden (vgl. Abschnitt 3.3, *Erkennung einfach strukturierter Objekte*) und komplexere Objekte wie Stühle (vgl. Abschnitt 3.5, *Strukturbasierte Objekterkennung*) zu erkennen.

Bei dem in diesem Abschnitt vorgestellten Segmentierungsverfahren wird nun eine dichtebasierte Untersuchung vorgenommen. Dazu wird die funktionale Bezugsebene, also beispielsweise die einem Tisch zugehörige detektierte Tischoberfläche, auf der die Gegenstände positioniert sind, verwendet. Diese legt zum einen die Höhe fest, oberhalb derer die Punkte untersucht werden, zum anderen wird auch der Suchbereich innerhalb der Ebene dadurch beschränkt. Dieser Bereich ist entweder durch die achsenparallelen Ausdehnungsgrenzen des Ebenensegments oder durch die polygonal umschriebene Fläche gegeben.

Der entwickelte Segmentierungsalgorithmus besteht aus folgenden drei Hauptschritten:

- Wahl der Bezugsebene bzw. des Bezugsobjekts, auf dem die Objekte positioniert sind: Diese ist das Ergebnis eines vorgeschalteten Klassifizierungsschritts (vgl. Kap.

3.3, *Erkennung einfach strukturierter Objekte*).

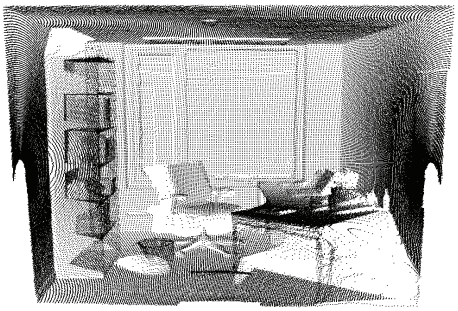
- Trennung der Ausgangspunktwolke von den Raumbegrenzungen, falls die Bezugsebene ein Boden ist, damit auch dicht an den Wänden stehende Objekte extrahiert werden können.

- Abstandsabhängige dichtebasierte Segmentierung der Punkte oberhalb der Bezugsebene.

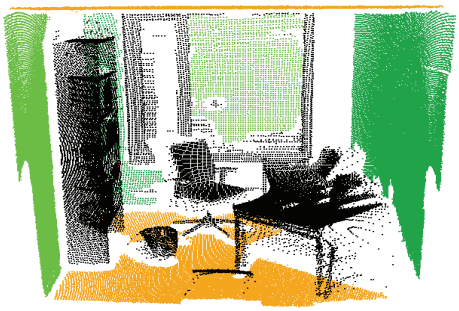
Diese Einzelschritte werden am Beispiel einer Innenraumszene, die aus zwei unterschiedlichen Perspektiven aufgenommen wurde, in der Abb. 3.15 vorgestellt. Die Bilder (a) bis (e) zeigen Aufnahmen, die den Raum als Ganzes zeigen und behandeln. Die Aufnahmen (f) bis (h) beschränken sich auf einen Teil der Gesamtszene, den Tisch, und wurden aus einer zweiten Position heraus aufgenommen.

Die Abb. 3.15(a) zeigt die Szene zusammen mit der Aufnahmeeinrichtung aus der ersten Aufnahmeposition. Daneben zeigt die Abb. 3.15(b) die entsprechende Tiefenaufnahme. Im ersten Verarbeitungsschritt werden die Ebenen im Bild segmentiert und diese entsprechend der Methode aus Kap. 3.3, *Erkennung einfach strukturierter Objekte*, klassifiziert. Resultierend werden die Raumgrenzen, also Wände, Decke und Boden, bestimmt. Für diese erste Untersuchungsreihe sollen die auf dem Boden befindlichen Objekte extrahiert werden. Zunächst werden die Informationen über die Decke und Wände, sowie der Bezugsebene, dem Boden, dazu genutzt die Ausgangspunktwolke von diesen zu trennen. Nur so ist es möglich, beispielsweise dicht oder direkt an einer Wand stehende Objekte von dieser zu lösen und zu segmentieren. Die Abb. 3.15(c) zeigt das erhaltene Ergebnis. Die Decke ist zur besseren Visualisierung dabei in Ocker gehalten, die Wände in Grüntönen. Die schwarzen Punkte stellen die von den Raumbegrenzungen gelösten Punkte dar. Im nächsten Schritt folgt dann die eigentliche Segmentierung auf dieser reduzierten Punktmenge. Dazu werden die Punktabstände von geometrisch benachbarten Punkten dazu genutzt, zusammengehörige Gruppen zu bilden. Unterschreitet die Punktdichte einen bestimmten Grenzwert, werden die beteiligten Punkte der gleichen Gruppe zugeordnet. Da das 3-D Aufnahmesystem fächerförmige Abtastbereiche hat, haben weiter entfernte Punkte automatisch einen größeren Punktabstand, alleine basierend auf der Aufnahmegeometrie. Um dies auszugleichen, wird der Punktabstand mit dem mittleren Punktabstand der beteiligten Punkte zum Ursprung, der im Scanner liegt, ins Verhältnis gesetzt. Somit wird einer Untersegmentierung im Nahbereich (Objekte

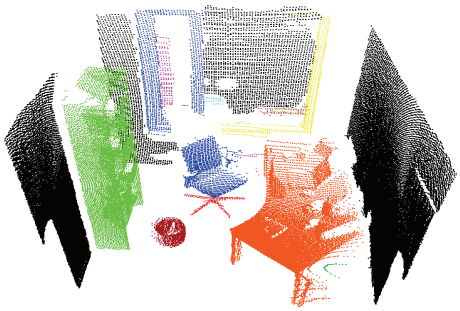
(a) Position A



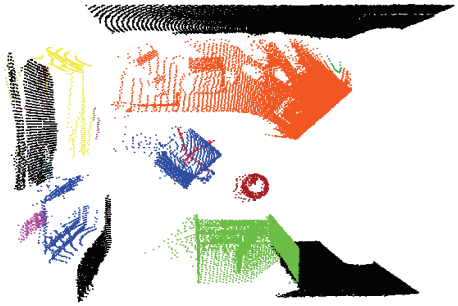
(b) Tiefenbild



(c) Von der Umgebung extrahierte Datenpunkte

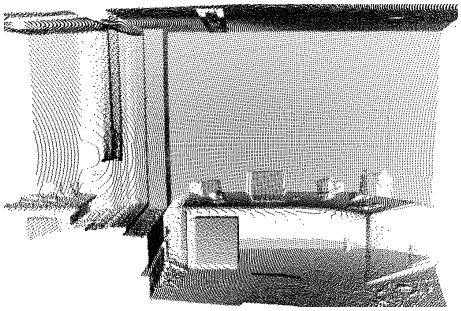


(d) Extrahierte Segmente, Blickwinkel 1

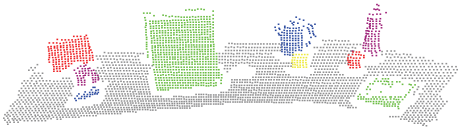


(e) Extrahierte Segmente, Blickwinkel 2

(f) Position B



(g) Tiefenbild



(h) Extrahierte Segmente

Abbildung 3.15: Dichtebasierte Segmentierung auf Basis ausgezeichneter Ebenen

verschmelzen miteinander) und eine Übersegmentierung im Fernbereich (Segmente bestehen aus ganz wenigen, im Extremfall nur einem Punkt) entgegengewirkt. Die beiden Abbildungen 3.15(d) und 3.15(e) zeigen die auf diese Weise erhaltenen Segmente. Hier sind die Wände zur besseren Orientierung in schwarz mit eingezeichnet. Auf der linken Seite wurde das Regal (grün), auf der rechten Seite der Tisch mitsamt Gegenständen (orange) komplett segmentiert. In der Mitte des Raums wurde der Mülleimer (dunkelrot) extrahiert. Beim Bürostuhl kommt es aufgrund der Perspektive und der Bauart zu einer Übersegmentierung, er zerfällt in zwei Teile, Basis und Sitzteil. Da bei den Klassifizierungsalgorithmen entweder der funktionale Anteil untersucht wird oder ein schablonenhafter Vergleich stattfindet, ist das Fehlen der Basis von keinem bzw. nur kleinem Einfluss. Auffällig ist das Segmentierungsergebnis an der Kopfseite. Hier kam es aufgrund des heruntergelassenen Rollos zu einer Fehlklassifikation. Neben dem linken, realen Wandsegment wurde auch das Rollo als Wandsegment segmentiert. Die angelegten Lamellen bilden eine hinreichend ebene Fläche, die zudem ganz nah bis zur Decke reicht. Es ergeben sich daraus mehrere Segmente, die Teile der Fensternische bzw. des Fensterrahmens darstellen.

Die Abb. 3.15(f) zeigt die Teilszene mit dem Tisch aus der zweiten Aufnahmeposition heraus. Das zugehörige Tiefenbild zeigt die Abb. 3.15(g). Hier ist interessant, dass es aufgrund des Einfallswinkels zu einer Spiegelung (Totalreflexion) an der Fensterscheibe kommt, die das Zimmer scheinbar nach links fortsetzt. Die Objekte auf dem Tisch werden dennoch mit Ausnahme der Computermaus alle vollständig segmentiert. Bei dieser existieren einfach zu wenige Reflexionspunkte, als dass sie zu einem Segment kondensierten. Beim Buch zeigt sich ein weiterer Spiegeleffekt, der durch eine transparente Buchschutzfolie verursacht wird.

3.7 Umgebungsboxberechnung

Die Umgebungsboxberechnung setzt auf bereits segmentierten Objekten an. Im *ORCC*-System wird sie aus drei Gründen berechnet:

- Die Umgebungsbox liefert die Information über die Ausdehnung des Objekts. Diese Information kann bei entsprechendem Objektklassenumfang bereits zur Klassifikation von Objekten eingesetzt werden. Dies wurde in [WR06] ebenfalls an Objekten

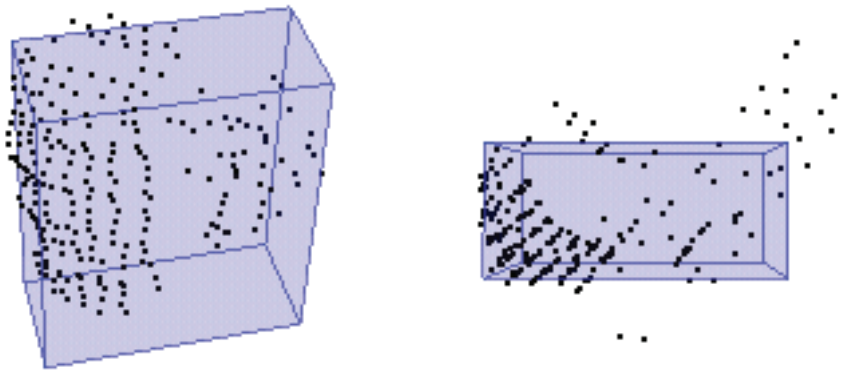


Abbildung 3.16: Umgebungsbox eines Kaffeebechers

einer Büroszene mit dem *ORCC*-System durchgeführt.

- Als weiteres Ergebnis liefert die Umgebungsbox eine Möglichkeit, den Ort eines Objekts festzulegen. Der Schwerpunkt der Punktwolke beispielsweise wäre wegen der unterschiedlichen Punktedichteverteilung ein ungünstiges Maß zur Festlegung der Position.

- Bei der Erkennung teilverdeckter Objekte im folgenden Abschnitt 3.8, *3-D ansichtenspezifische formbasierte Objekterkennung*, ist zur Berechnung der Kollision zwischen Sehstrahl und verdeckendem Objekt eine flächenhafte Erfassung des verdeckenden Objekts nötig. Diese wird durch den Quader der Umgebungsbox der Punktwolke angenähert.

Aufgrund der Scannercharakteristik haben die erhaltenen Daten häufig Scheinpunkte, die sich an Objektkanten bilden (vgl. Abschnitt 3.1, *Sensordaten*). Es wurde deshalb ein Verfahren entwickelt, das diese spezielle Situation beachtet [Mey05]. Die Gütebeschreibung der Umgebungsbox setzt sich aus zwei Termen zusammen, einem punktbasierten Anteil und einem seitenbasierten Anteil. Der erste besteht aus der Summe der einzelnen Punkte zu ihren nächstgelegenen Seitenprojektionen. Der zweite berechnet sich ebenso, bezieht sich allerdings jeweils nur auf den zehnten Teil einer jeweiligen Seite nächstgelegenen Punkte. Ausgehend von einer achsenorientierten Umgebungsbox, wird solange iteriert, bis ein Optimum erreicht ist. Die Abbildung 3.16 zeigt das Beispiel eines Kaffeebechers, der recht viele Scheinpunkte produziert. Auf Seite 101 ist der braune Becher

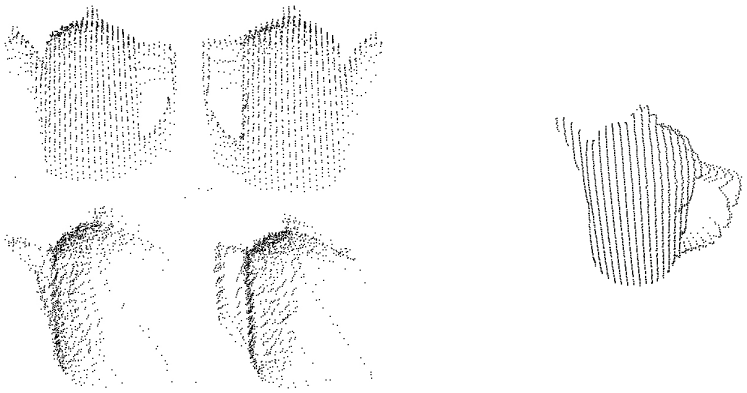
mit ausgeprägtem Henkel in der Mitte der Abbildung 4.2(b) zu sehen.

3.8 3-D ansichtenspezifische formbasierte Objekterkennung

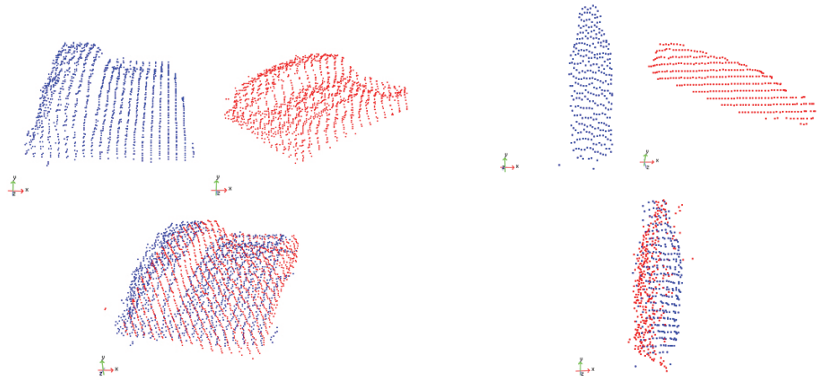
Im Kapitel 2, *Kognitive Aspekte der Objekterkennung*, wurden verschiedene Objekterkennungsstrategien nach kognitionsspezifischen Aspekten hin untersucht. Eine Möglichkeit, die funktionalen Aspekte eines Objekts für dessen Kategorisierung zu verwenden, wurde im Abschnitt 3.5, *Strukturbasierte Objekterkennung*, umgesetzt. Eine technische Schwierigkeit dieser Kategorie von Erkennungssystemen ist, dass diese als Grundlage des Entscheidungsprozesses spezifische Merkmale verwenden, die zwar durch ihre funktionale Orientierung in dieser Anwendung relativ robust sind, dennoch aber zunächst berechnet werden müssen. Eine alternative Möglichkeit besteht darin, die segmentierten Objekte auf unterster Informationsebene zu vergleichen. Zur Klassifizierung wird dabei ein Referenzobjekt verwendet und ein Vergleich zwischen Referenz- und Testobjekt auf der Ebene der Punktpositionen durchgeführt. Dieses pragmatische Vorgehen hat mehrere Vorteile: die Vorverarbeitung entspricht der Segmentierung der Objekte, es ist keine Lernphase nötig, bzw. mit der Aufnahme eines Referenzobjekts identisch, und neue Objektklassen können einfach hinzugenommen werden, ohne dass sich auf unterer Erkennungsebene für die bisher verwendeten Objektkategorien etwas ändert. Entsprechend hat dieses Vorgehen eine lange Tradition und wurde prinzipiell bereits Anfang der achtziger Jahre im Zusammenhang mit Arbeiten zur Oberflächenanpassung als Erkennungsmethode vorgeschlagen [Pot82], [BJ85]. Eine Umsetzung aus jüngerer Zeit findet sich in [Sch98]. Bei der Anwendung des hier entwickelten Erkennungsystems [Bre07] innerhalb *ORCC* wurden Gegenstände aus einer Büroumgebung für die Erkennung herangezogen, vgl. Abb. 3.17(a). Es sind, bezogen auf die Scannerauflösung, viele kleine Objekte darunter. Wie die Bilder in Abb. 3.17(b) zeigen, sind zuweilen deutliche Scheinpunkte[49] im Verschattungsbereich vorhanden. Als Referenzmodelle kommen zwei Verfahren zum Einsatz. Beim multiplen Referenzdatensatz werden mehrere Aufnahmen aus unterschiedlichen Perspektiven erstellt. In der Abb. 3.17(b) ist dies beispielhaft für das Objekt Kaffeekanne dargestellt. In der oberen Reihe sind die Aufnahmen von der Seite aufgenommen, so dass jeweils Schnabel und Henkel sichtbar sind. In der unteren Rei-

[49] vgl. Abschnitt 3.1, *Sensordaten*, auf Seite 42

(a) Auswahl der verwendeten Objekte



(b) Multiple Referenzobjekte bzw. einfaches Referenzobjekt



(c) Registriertes Telefon und registrierte Flasche als Grundlage für den Klassifizierungsprozess

Abbildung 3.17: 3-D formbasierte Objekterkennung [Bre07]

he wurden die Aufnahmen so gemacht, dass entweder nur die Schnabelpartie oder nur die Henkelpartie sichtbar ist. Beim Erkennungsprozess wird das Testobjekt gegen jedes dieser Untermodelle einzeln getestet. Im Fall des einfachen Referenzdatensatzes wird nur eine Aufnahme als Referenzmodell herangezogen, wobei darauf geachtet wird, dass möglichst viele Details in dieser Ansicht erfasst sind.

Die Güte der Registrierung, d.h. der möglichst optimalen Ausrichtung zwischen Test- und Referenzobjekt, dient bei dieser Klassifizierungsmethode als Grundlage zur Bewertung der Klassenzugehörigkeit. Zur iterativen Berechnung dieser Transformation wird der sog. ICP-Algorithmus (vgl. auch Abschnitt 3.4, *Registrierung*) eingängig mit dem Levenberg-Marquardt Algorithmus als Optimierungsmethode verwendet. Dieser berechnet in jedem Iterationsschritt, ausgehend von dem aktuellen Parametersatz, einen verbesserten Parametersatz. Als Maß der Registrierungsgüte dient grundsätzlich die Summe der Punktabstände vom Referenzmodell zum Testmodell. Um Ausreißer oder nicht in beiden Modellen vorhandene Bereiche nicht zu gewichten, werden Distanzen oberhalb eines gewissen Grenzwertes nicht berücksichtigt. Als Abbruchkriterium dient die Größe der Verbesserung pro Iterationsschritt, bezogen auf die Gütefunktion. Wird nun das Gütemaß für die Klassenzugehörigkeit bestimmt, werden die durchschnittlichen Punktabstände zwischen Referenz und Testmodell herangezogen. Dabei ergibt sich das Problem, dass dieser Wert davon abhängt, über welche der beiden Punktmengen er ermittelt wurde. So kann es passieren, dass sich ein kleineres Objekt optimal in einem größeren einpasst, beispielsweise die Kontur einer Tasse in der einer Flasche. Um solche Fehlinterpretationen an dieser Stelle auszuschließen, wird einmal der Abstand ausgehend vom Referenzmodell und einmal ausgehend vom Testmodell berechnet und miteinander verrechnet. Aufbauend auf diesen Werten wird ein Gütemaß für die jeweilige Klassenzugehörigkeit entwickelt.

In der Abbildung 3.17(c) sind die Ergebnisse des Registrierungsvorgangs für einen Telefonapparat und ein Flasche gezeigt. Wie an den Koordinatenkreuzen zu erkennen ist, wurde das rote Testobjekt aus einer jeweils stark geänderten Orientierung gegenüber dem blauen Referenzmodell aufgenommen. Im Fall der Flasche wurde diese sogar liegend positioniert. In beiden Fällen schafft es der Registrierungsalgorithmus, die Objekte nahezu deckungsgleich zu transformieren. Für einen in [Bre07] durchgeführten Test lieferte das System eine entsprechend positive Erkennungsleistung. Bei einem Versuch

wurden von 15 Objekten[50] 11 Stück korrekt klassifiziert, d.h. mit Rangposition 1 erkannt. Bei drei Flaschensorten wurde eine Zuordnung zu der richtigen Art Flasche auf dem zweiten Platz erkannt und stattdessen einer anderen Flaschensorte den Vorzug gegeben. Algorithmische Probleme gab es beim Tacker. Hier lieferte die Registrierung ein falsches Ergebnis, und führte infolge zu einer schlechten Einschätzung. Dies lässt sich bei dem Verfahren leider nicht grundsätzlich ausschließen und hängt natürlich von sehr vielen Faktoren ab. In [Sch98] wird dieses Problem dadurch gedämpft, dass man einfach eine Reihe von Anfangsausrichtungen und entsprechenden Registrierungen anfertigt und dann die beste auswählt. Allerdings steigt dadurch auch der Gesamtaufwand linear an.

Einem grundlegenden Problem wurde sich ebenfalls in [Bre07] zugewandt, nämlich dem der Teilverdeckung. Wird ein Objekt nicht vollständig erfasst, so kann dies bei der automatischen Registrierung schnell zum Scheitern des Vorgangs führen. In dem entwickelten System wurde deshalb die Möglichkeit entwickelt, durch eine automatische Adaption des Referenzmodells im Falle von Verdeckungen ein stabileres Ergebnis zu erreichen. Potentiell ließe sich dem Problem der Verdeckung auch im Sinne einer Active Vision[51] (*dt.* Aktives Sehen) Strategie begegnen, bei der das System auf diese erkannte Verdeckung durch eine Änderung der Position auf eine verdeckungsfreie Blickrichtung ausweicht. In dem hier entwickelten Verfahren, das in der Abb. 3.18 beispielhaft bei einer durch einen Kaffeebecher verdeckten Bierflasche dargestellt ist, ist eine Aufnahmepositionsänderung jedoch nicht nötig. Das System reagiert auch hier aktiv auf die Verdeckungssituation, aber nicht durch eine Anpassung der Blickposition, sondern der Vergleichsschablone, basierend auf dem Wissen über die Ausdehnung des verdeckenden Objekts. Betrachtet man das mittlere Bild in Abb. 3.18, wird schnell klar, dass bei den wenigen Punkten, die der Flaschenhals der Bierflasche liefert, ein Vergleich mit dem ganzen Modell nicht sinnvoll wäre. Der Ansatz wird deshalb von der Idee getragen, dass die verdeckte Modellschablone in ihrer jeweiligen imaginären Position im Registrierungsprozess so beschnitten wird, als läge sie ebenfalls im Schatten des verdeckenden Objekts, hier des blauen Kaffeebechers. Nicht beachtet werden diejenigen Punkte, deren Verbindungslinie zum Scanner mit dem verdeckenden Objekt, vereinfacht repräsentiert durch dessen

[50] Liste der verwendeten Objekte: Bierflasche, Wasserflasche, Weinflasche, Saftflasche, Kola Flasche, große Dose, kleine Dose, Tupperdose, Kaffeebecher, Kaffeekanne, Kaffeemaschine, Kaffeetasse, Locher, Tacker und Telefon.

[51] „Der Begriff Aktives Sehen beschreibt eine Methodik in der Bildverarbeitung, in der die Kamera in einem Bildverarbeitungssystem eine aktive Rolle einnimmt. Gemäß einer Definition von Swain und Stricker [Str93], bezeichnet Aktives Sehen Sehsysteme mit der Eigenschaft, entsprechend der behandelten Aufgabe aktiv Änderungen von Kameraparametern vorzunehmen.“ [BSB02]

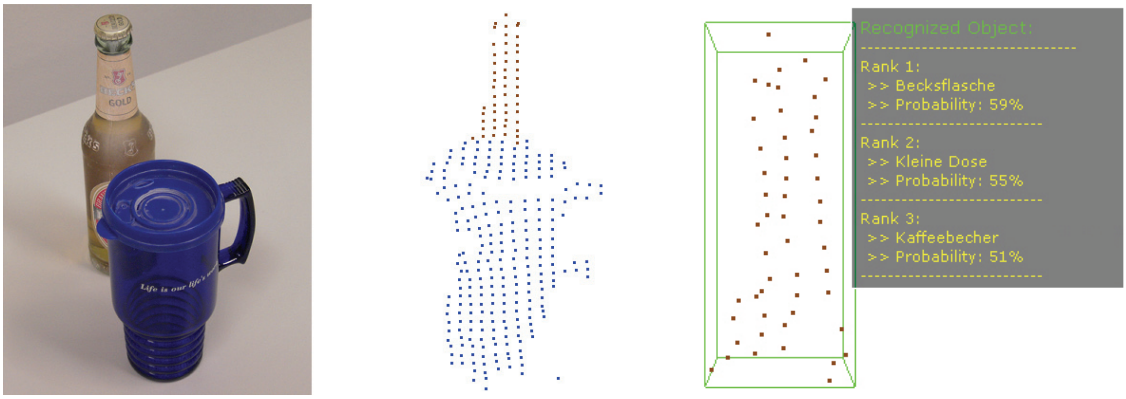


Abbildung 3.18: Erkennung teilverdeckter Objekte [Bre07]

Bounding-Box, kollidieren. Die Schablone wird dabei so in der Höhe positioniert, dass sie in ihrer Gesamtheit direkt auf dem Tisch steht, bzw. der unterst gelegene Punkt die gleiche Höhe hat wie das verdeckende Modell. Bei der untersuchten Szene kann auf diese Weise, obgleich die Flasche nur zu einem Viertel erfasst ist, die Klassenzugehörigkeit korrekt bestimmt werden.

3.9 2-D/3-D Systemkalibrierung

Zielanwendung des *ORCC*-Systems ist die sprachliche Kommunikation eines Menschen mit dem System über die aufgenommene und analysierte Szene. Die Datenbasis besteht aus dreidimensionalen Tiefendaten und zweidimensionalen Farbdaten. Um beide einander zuordnen zu können, muss die Transformation zwischen beiden Aufnahmen berechnet werden, man spricht dabei auch von Registrierung. Die Registrierung von Aufnahmen verschiedener Dimension ist vor allem in der Medizintechnik von Interesse [KPT+05]. Dort geht es beispielsweise darum, konventionelle Röntgenaufnahmen mit Computertomographieaufnahmen zu verknüpfen, um somit die verschiedenen Aufnahmesysteme geometrisch miteinander zu verbinden. Somit können die Vorteile verschiedener Aufnahmesysteme kombiniert und Orte von Interesse in der einen Aufnahme in der anderen lokalisiert werden. In der Abb. 3.19 sind auf der linken Seite die im *ORCC* System beteiligten Komponenten dargestellt: der in blau und schwarz gehaltene Laserscanner und die in Silber aufgesetzte Videokamera. Um die Abbildungsvorschrift zwischen beiden Aufnahmesystemen zu erhalten, müssen zwei Kalibrierungen vorgenommen werden. Zum

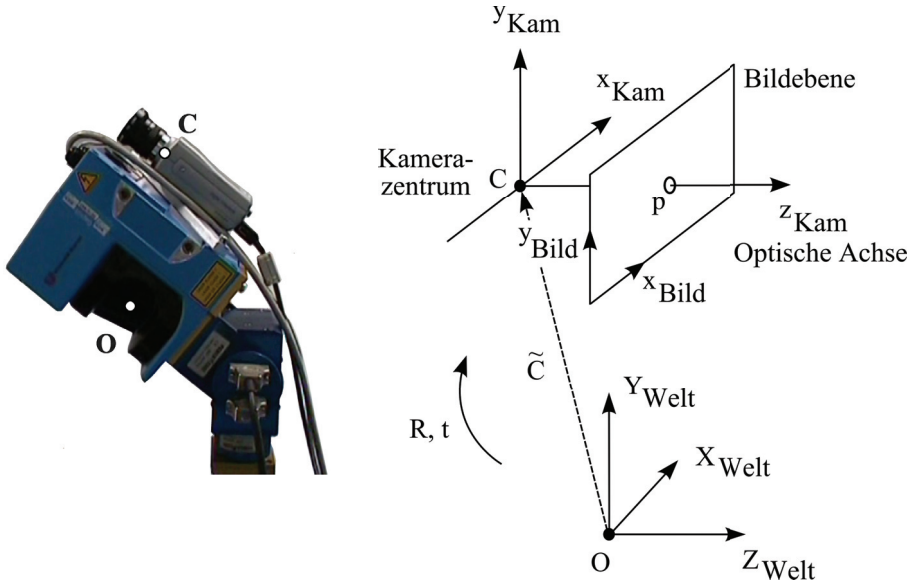


Abbildung 3.19: Transformation zwischen Kamera- und Weltkoordinaten [HZ04]

einen müssen die internen Parameter der Farbbildkamera, wie die Brennweite, bestimmt werden, zum anderen deren sog. externe Kameraparameter, d.h. die Positionierung gegenüber einem äußeren Weltzentrum, das in unserem Falle dem Scannermittelpunkt entspricht. Bekannt ist die Auflösung des Fotosensors der Kamera[52], die 1280×960 Pixel beträgt. Die Abbildung der Farbbildkamera ergibt sich unter Verwendung homogener Koordinaten für ein Lochkameramodell als folgende lineare Abbildung [HZ04]:

$$\underbrace{\begin{pmatrix} fX \\ fY \\ Z \end{pmatrix}}_{\text{Bild-koordinaten}} = \underbrace{\begin{bmatrix} f_x & 0 & p_x & 0 \\ 0 & f_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{=[K \mid \mathbf{0}],\ \text{Kameramatrix } K} \underbrace{\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}}_{\text{Kamera-koordinaten}}$$

bzw. in kompakter Schreibweise:

$$\mathbf{x}_{Bild} = K\,[I \mid \mathbf{0}]\,\mathbf{X}_{Kam}$$

Die Überführung von kartesischen Koordinaten in homogene geschieht durch eine erweiterte Komponente mit Eintrag 1. Die Rückführung auf kartesische Koordinaten erfolgt durch Division der einzelnen Komponenten durch die letzte und Streichung dieser. Diese

[52] Bei der Farbbildkamera handelt es sich um das Modell DFW-SX900 der Firma Sony.

Abbildungsvorschrift beschränkt sich in dieser Form ausschließlich auf das von der Kamera vorgegebene Koordinatensystem. Um dieses mit dem äußeren sog. Weltkoordinatensystem zu verknüpfen, muss die Abbildungsvorschrift entsprechend erweitert werden. Zunächst wird ein Punkt aus dem Weltkoordinatensystem in das Kamerakoordinatensystem transformiert. Für einen nicht homogenen[53] Punkt $\tilde{\mathbf{X}}_{Welt}$ im Weltkoordinatensystem bedeutet dies:

$$\tilde{\mathbf{X}}_{Kam} = R\left(\tilde{\mathbf{X}}_{Welt} - \tilde{\mathbf{C}}\right) \ ,$$

wobei $\tilde{\mathbf{C}}$ dem Kamerazentrum in Weltkoordinaten und R der die Orientierung des Kamerakoordinatensystems im Weltkoordinatensystem beschreibenden 3×3 Matrix entspricht. In homogenen Koordinaten kann diese Abbildung geschrieben werden als:

$$\mathbf{X}_{Kam} \ = \ \begin{bmatrix} R & -R\tilde{\mathbf{C}} \\ \mathbf{0}^T & 1 \end{bmatrix} \mathbf{X}_{Welt}$$

Zusammen mit der Projektionsabbildung ergibt sich folgende Gesamtabbildung:

$$\mathbf{x}_{Bild} \ = \ K\left[I \mid \mathbf{0}\right] \begin{bmatrix} R & -R\tilde{\mathbf{C}} \\ \mathbf{0}^T & 1 \end{bmatrix} \mathbf{X}_{Welt} = KR\left[I \mid -\tilde{\mathbf{C}}\right] \mathbf{X}_{Welt}$$

Indem man $-R\tilde{\mathbf{C}}$ durch t ersetzt, unterdrückt man die Sichtbarkeit der Abhängigkeit vom Kamerazentrum und erhält eine vereinfachte Abbildungsvorschrift:

$$\begin{aligned} \mathbf{x}_{Bild} \ &= \ \underbrace{K\left[R \mid t\right]}_{= P} \mathbf{X}_{Welt} \\ &\qquad \text{Kameraprojektionsmatrix} \end{aligned}$$

Das hier verwendete Kameramodell hat insgesamt also 10 Freiheitsgrade, 4 für K, 3 für R und 3 für $\tilde{\mathbf{C}}$.

Bei der Kalibrierung geht es im ersten Schritt darum, die sog. inneren Kameraparameter, d.h. die Werte der Parameter der Kameramatrix K, zu bestimmen. Im zweiten Schritt werden die externen Parameter, d.h. der Versatz zwischen Farbbildkamera und Laserscanner, bestimmt. Es gibt grundsätzlich zwei mögliche Kalibrierungsarten [Zha98]: photogrammetrische Kalibrierung oder Selbstkalibrierung. Bei der photogrammetrischen Kalibrierung wird ein Kalibrierobjekt bekannter Geometrie, meistens aus senkrecht zueinander stehenden Ebenen mit Schachbrettmuster genutzt, um aus den 2-D/3-D Punktkorrespondenzen auf die Abbildungsparameter zu schließen. Bei der Selbstkalibrierung

[53] Nicht homogene, also kartesische Punktbeschreibungen, werden in diesem Abschnitt mit einer Tilde gekennzeichnet.

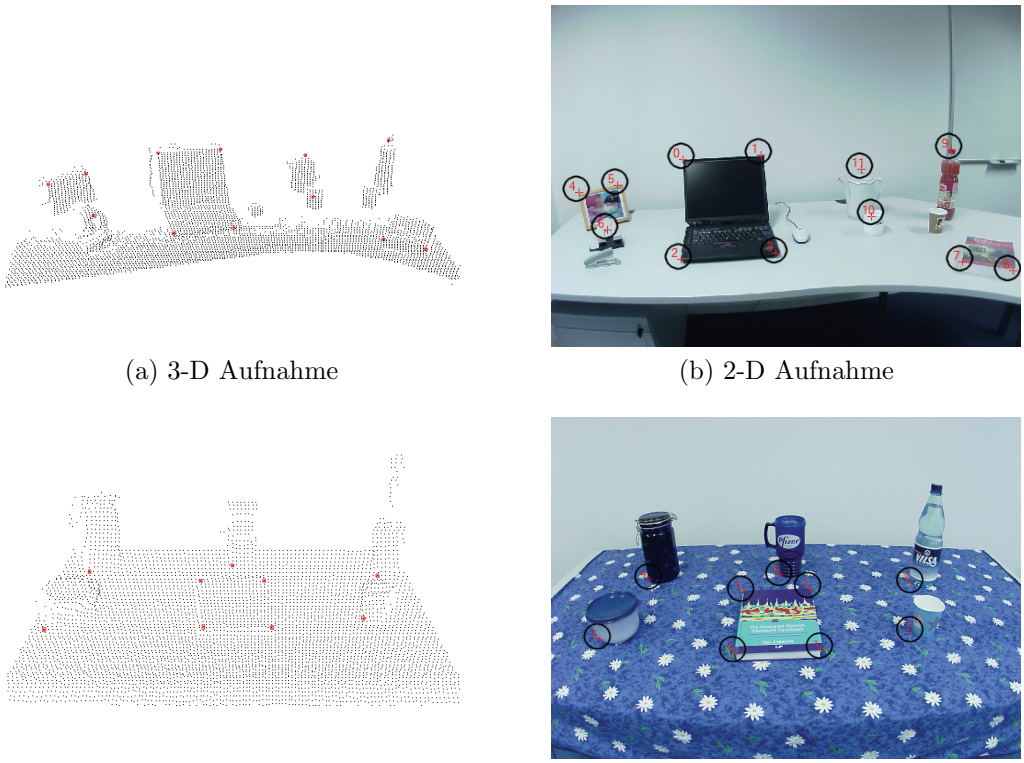
hingegen werden ausschließlich die zugeordneten Punktkorrespondenzen aus unterschiedlichen Aufnahmen für die Kalibrierung genutzt. Entsprechend relativ instabiler sind die damit erreichten Ergebnisse. Die hier verwendeten Verfahren liegen zwischen beiden Verfahren in ihrer Reinform. Bei der Kalibrierung der Farbbildkamera wird zwar ein Muster verwendet, das allerdings nur zweidimensionale Informationen enthält. Bei der Bestimmung der externen Parameter wird kein spezielles Muster benötigt. Es werden natürlich selektierbare Punkte in den beiden Aufnahmen, also sowohl dem 2-D Farbbild als auch dem 3-D Tiefenbild, verwendet. Allerdings ist eine Benutzerinteraktion nötig, um die Punkte aus den verschiedenen Aufnahmesystemen zu selektieren und einander zuzuordnen.

Für die Berechnung der inneren Parameter wird das Verfahren von [Zha98] in der Anwendung von [VA05] basierend auf Methoden aus OpenCV [Sou07] verwendet. Bei diesem Verfahren wird zunächst analytisch eine initiale Lösung für die internen Kameraparameter bestimmt. Diese basiert auf Verwendung der durch Punktkorrespondenzen bestimmten Homographien[54]. Dazu werden von einem ebenen Schachbrettmuster aus verschiedenen Perspektiven Aufnahmen gemacht. Für jede Aufnahme wird eine Homographieabbildung bestimmt, die für ein Gleichungssystem jeweils zwei weitere Beiträge liefert. Ab einer Anzahl von 3 Homographien kann das Gleichungssystem bis auf einen linearen Faktor eindeutig bestimmt werden, woraus sich dann die initialen Werte der internen Kameraparameter, und infolge auch der externen Parameter, berechnen lassen. Ausgehend von der berechneten initialen Schätzung, werden die Parameter mittels weiterer iterativer Optimierungsverfahren weiter verbessert. In diesem Optimierungsschritt wird dann auch die Kameraverzerrung beachtet. Wie man in der Abbildung 3.20(d) an der eigentlich geraden Tischkante sehen kann, verzeichnet das Objektiv, entsprechend einem starken Weitwinkelobjektiv, stark kissenförmig. Dies ist neben der Art auch auf die geringe Güte der verwendeten Komponente zurückzuführen. Die Linsenverzerrung ist durch 4 Parameter gegeben:

$$\begin{aligned} \tilde{x} &= x + x\left(k_1 r^2 + k_2 r^4\right) + \left(2p_1 xy + p_2\left(r^2 + 2x^2\right)\right) \\ \tilde{y} &= y + y\left(k_1 r^2 + k_2 r^4\right) + \left(2p_2 xy + p_1\left(r^2 + 2y^2\right)\right) \end{aligned}$$

Die Parameter k_1 und k_2 sind die radialen Verzerrungskoeffizienten, p_1 und p_2 die tangentialen Verzerrungskoeffizienten. Die Werte von $\tilde{x}$ und $\tilde{y}$ für die realen Bildkoordinaten,

[54] Unter einer Homographie versteht man die Abbildung zwischen zwei Ebenen, hier der zwischen dem Schachbrettmuster und der Kamerabildebene.



(a) 3-D Aufnahme

(b) 2-D Aufnahme

(c) 3-D Aufnahme

(d) 2-D Aufnahme

Abbildung 3.20: Kalibrierung (Datensätze #277 und #281) [Kan07]

die von x und y für die idealisiert verzerrungsfreien Bildkoordinaten. Nach einer initialen Bestimmung der Koeffizienten werden alle Parameter zusammen, beschrieben durch ein nichtlineares Optimierungsproblem, weiter verbessert.

Um danach die externen Parameter, die hier gleichzusetzen sind mit dem Versatz zwischen Farbbildkamera und Laserscanner, zu bestimmen, werden Punktkorrespondenzen aus der jeweiligen Szene verwendet (vgl. Abb. 3.20). Im ersten Schritt werden initiale Werte mittels DLT (Direkte Lineare Transformation) berechnet (vgl. [Sou07], [HZ04]), gefolgt von einem auf dem Downhill-Simplex Algorithmus basierenden Verfahren, das in der Diplomarbeit von Christian Kanthak entwickelt wurde [Kan07]. Die Güte der erhaltenen Parameter wird dabei durch den Abstand der rückprojizierten Punkte im 2-D Bild bestimmt (vgl. Abb. 3.21). Für jeden der gewählten 2-D Kalibrierpunkte (durch rote Fadenkreuze gekennzeichnet) werden die entsprechenden 3-D Kalibrierpunkte mit dem aktuellen Parametersatz in das 2-D Bild zurückprojiziert (durch blaue Fadenkreuze gekennzeichnet). Im Idealfall einer perfekten Kalibrierung wären beide deckungsgleich.

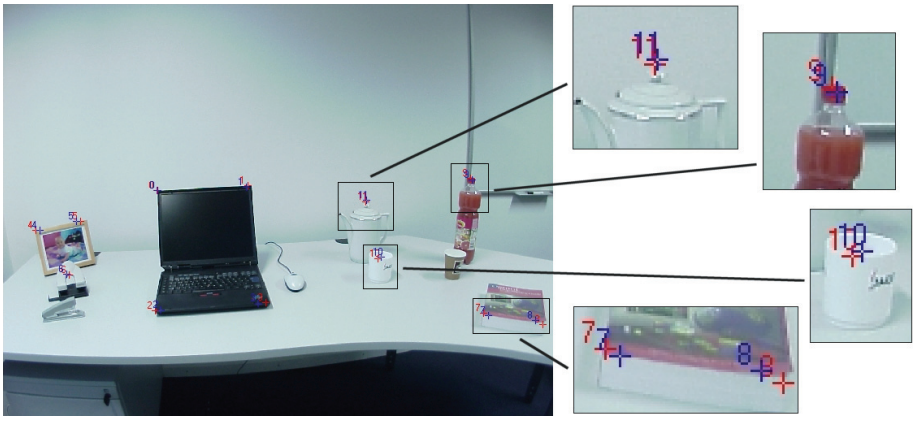


Abbildung 3.21: Kalibrierung: Rückprojektionsfehler [Kan07]

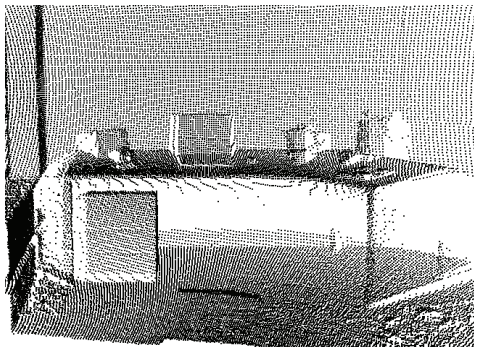
Ansonsten gibt deren euklidischer Gesamtabstand die Güte der Projektionsmatrix an:

$$\min \sum_i d\left(\mathbf{x_i}_{Kam}, P\mathbf{X_i}_{Welt}\right)^2$$

Dass die Kalibrierung an einigen Stellen nicht optimal ist, liegt unter anderem an der beschränkten Auflösung des Scanners, die die notwendige Punktauswahl zuweilen schwierig gestaltet, und der extrem starken Verzeichnung des Objektivs.

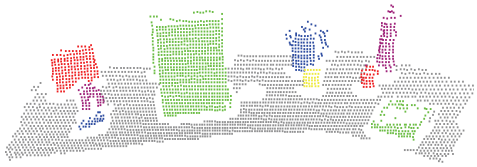
3.10 Texturausschnittsberechnung

Die farbbasierte Segmentierung beliebiger Objekte in unbestimmter Umgebung ist unmöglich. Deshalb werden i.d.R. die Informationen über die farbliche oder geometrische Beschaffenheit für die Segmentierung der Objekte verwendet. Selbst dann wird es aber sehr schwierig, wenn, wie beispielsweise in Abb. 3.20(d) zu sehen, der Hintergrund sehr unregelmäßig und farbähnlich zu den zu segmentierenden Objekten ist. Im *OR-CC*-System wird deshalb umgekehrt die geometrische Situation des Hintergrunds dazu genutzt die Objekte dreidimensional zu segmentieren. Basierend auf der Detektion bestimmter funktionaler Ebenen können die Objekte dreidimensional, unabhängig von der farblichen Charakteristik, segmentiert werden (vgl. Abschnitt 3.6, *Segmentierung auf funktionalen Ebenen*). Andererseits ist die Textur eines Objekts aber maßgeblich für dessen Wahrnehmung und damit auch für deren verbale Referenzierung. Um die dazu notwendigen Texturausschnitte der Objekte zu erhalten, werden insgesamt folgende



(a) 3-D Aufnahme

(b) 2-D Aufnahme



(c) 3-D Segmente

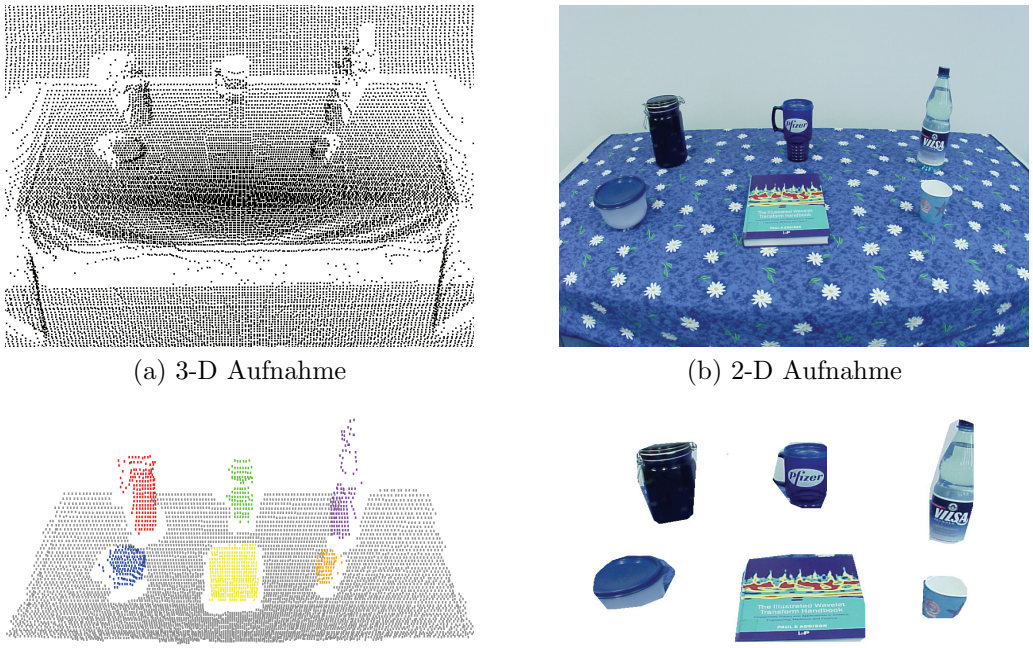
(d) 2-D Segmente

Abbildung 3.22: 2-D/3-D Registrierung (Schreibtischszene)

Schritte durchlaufen:

- 3-D Segmentierung (vgl. Kap. 3.6, *Segmentierung auf funktionalen Ebenen*)
- 2-D/3-D Registrierung (vgl. Kap. 3.9, *2-D/3-D Systemkalibrierung*)
- Berechnung der 2-D Korrespondenzpunkte
- Bildung eines konvexen Polygons aus den 2-D Punkten
- Berechnung des Bildausschnitts

In den Abb. 3.22 und 3.23 wurde die Ausschnittsberechnung für zwei Szenen durchgeführt, wobei einige der Zwischenschritte nicht dargestellt sind. Die Ausgangssituation ist in beiden Szenen eine 3-D Tiefenaufnahme (Abb. 3.22(a) und 3.23(a)) und eine dazugehörige 2-D Aufnahme (Abb. 3.22(b) und 3.23(b)). In beiden Szenen wurden zunächst die Tischoberflächen detektiert (vgl. Kap. 3.3, *Erkennung einfach strukturierter Objekte*), und dann die darauf befindlichen Objekte segmentiert (vgl. Kap. 3.6, *Segmentierung auf funktionalen Ebenen*). Die Transformation zwischen 3-D und 2-D Aufnahme wur-



(a) 3-D Aufnahme (b) 2-D Aufnahme

(c) 3-D Segmente (d) 2-D Segmente

Abbildung 3.23: 2-D/3-D Registrierung (Tischszene)

de, wie in Abschnitt 3.9, *2-D/3-D Systemkalibrierung*, vorgestellt, für beide Aufnahmen gesondert gerechnet. Damit sind alle Ausgangswerte und Parameter, die zur Texturausschnittsberechnung nötig sind, vorhanden. Zunächst werden für jeden 3-D Punkt eines 3-D Segments die entsprechenden 2-D Punkte in der Farbaufnahme bestimmt. Daraus resultiert eine Menge von Punktkorrespondenzen, auf denen das konvexe Polygon gebildet wird. Da die 3-D Punktwolke sehr 'löcherig' ist, ist dies auch die Korrespondenzpunktwolke in der Farbaufnahme. Deshalb kann nicht entschieden werden, an welchen Stellen ein Einschnitt des Objekts und an welchen Stellen die Auflösungsdichte des Systems zu gering ist. Durch die Verwendung des konvexen Polygons gehen keine Anteile des Texturausschnitts verloren. Solange der Anteil von Hintergrunds- bzw. Umgebungspunkten nicht zu groß ist, können diese Ausschnitte dennoch hinreichend für eine farbliche Beschreibung genutzt werden. In der Abb. 3.22(d) sind die Texturausschnitte zu den Segmenten in Abb. 3.22(c) zu sehen. Bis auf die Maus wurden alle auf dem Tisch befindlichen Objekte erfolgreich dreidimensional segmentiert. Für die Maus existieren zu wenig Scanpunkte, als dass sie als eigenes Objekt erkannt würde. Insgesamt ist für kleine Objekte das Auflösungsvermögen des Scanners in einem ungünstigen Verhältnis. So existieren für die Porzellantasse nur 36 Scanpunkte und für den Tacker gerade mal

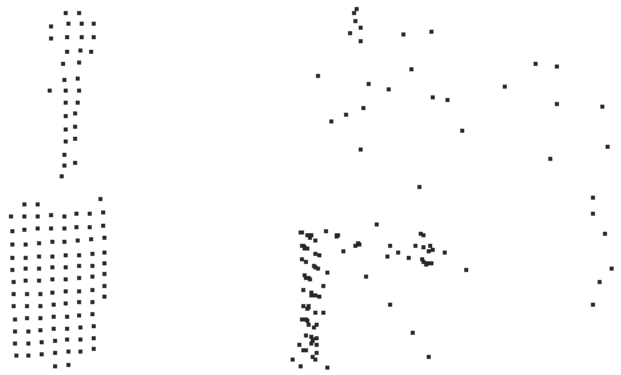


Abbildung 3.24: Grenzen des Aufnahmesystems: Wasserflasche aus Plastik in Frontal- und Seitenansicht

27 Scanpunkte. Betrachtet man die mit Hilfe der Kameraparameter berechneten zugehörigen polygonalen Segmente im 2-D Bild, fällt auf, dass diese tendenziell etwas zu klein ausgeschnitten werden. Dies hängt unter anderem damit zusammen, dass an den Objektkanten sog. Scheinpunkte entstehen, welche die Objekte zum Hintergrund hin abflachen (vgl. Abschnitt 3.1, *Sensordaten*). Dies führt letztlich dazu, dass Randpunkte teilweise nicht mehr zum Objekt zugeordnet werden. Da es an dieser Stelle nur auf die Ableitung einer farblichen Beschreibung ankommt, kann diese Verkleinerung der erhaltenen Ausschnitte jedoch akzeptiert werden. Welche zusätzlichen Seiteneffekte bei der Aufnahme auftreten können, zeigt die Abb. 3.24. Bei dieser Flasche aus Abb. 3.23(a) kommt es am Etikett zu einer Kantenbildung und darüber hinaus zu Transmissionen. Dennoch bleibt der resultierende 2-D Ausschnitt aussagekräftig genug, da für die Ausschnittsberechnung nur der projizierte Rand von Belang ist. Bei dieser Szene zeigt sich der grundsätzlich große Vorteil dieser Art von Segmentierung. Obgleich die verwendete Tischdecke ein sehr unregelmäßiges Muster hat und die Objekte in ähnlichen Farbtönen gehalten ist, was eine rein farbbasierte Segmentierung sehr schwierig machte, können die Objekte eindeutig segmentiert werden.

Die Übertragung der Segmentierungsergebnisse aus der der 3-D Aufnahme in das 2-D Bild funktioniert hinreichend, hat aber ihre Grenzen bedingt durch die Scannerauflösung und die Kalibriergüte. Dies könnte durch Optimierung der technischen Ausstattung verbessert werden. Eine andere Möglichkeit wäre, die auf diesem Weg erhaltenen Segmentierungen als initialen Wert für einen im 2-D Bild fortgesetzten Segmentierungsschritt zu verwenden, beispielsweise mittels aktiver Konturen [KWT87], [ABFJB03].

3.11 Farbnormierung und Farbbeschreibung

Bisher war das Hauptaugenmerk dieser Arbeit auf die Verwendung von 3-D Daten gerichtet. Mit den letzten beiden Abschnitten wurde der Weg zu einer farbintensitätsbasierten Untersuchung geöffnet, indem die Transformation zwischen der 3-D und der 2-D Aufnahme berechnet wurde. Farbe ist für uns Menschen bei der visuellen Wahrnehmung offensichtlich von großer Bedeutung. Wir Menschen sind dabei in der Lage, Farbe unter stark variierenden Bedingungen zu erkennen [Pau01]. Diese Fähigkeit der an sich robusten Wahrnehmung führt zu gewissen Seiteneffekten, die im Anhang A.2, *Optische Wahrnehmung*, an einigen Beispielen gezeigt werden. Die Schwierigkeit bei der Erkennung von Farben durch ein künstliches System hat ihre Ursache in verschiedenen physikalischen Gesetzmäßigkeiten. Lokal können Reflexionen dafür sorgen, dass gewisse Stellen überblendet werden, global können Änderungen des einfallenden Lichts die Reflexionswerte aller Gegenstände stark verändern. Dies gilt natürlich auch für den menschlichen Beobachter; er ist allerdings in der Lage, sich der Situation anzupassen, indem er beispielsweise Glanzpunkte ignoriert. Bei einem Kamerasystem kommen zudem noch Unzulänglichkeiten der Optik hinzu, die zu Abberation[55] und damit auffallenden Bildfehlern führen können. Bevor man deshalb versucht, Farbe zur Beschreibung oder Klassifikation von Objekten heranzuziehen, müssen Möglichkeiten entwickelt werden, diese negativen Effekte zu dämpfen. Grundsätzlich gibt es dafür zwei unterschiedliche Herangehensweisen: Farbnormierung und Farbkalibrierung. In der Arbeit von [Fin07] wurden insgesamt vier verschiedene Verfahren innerhalb des *ORCC*-Systems, dort im Zusammenhang für die farbbasierte Objekterkennung, entwickelt, umgesetzt und getestet. Zwei der Verfahren arbeiten direkt auf den aufgenommen Bildern und gehören zur Gruppe der Farbnormierungsverfahren, bei den beiden anderen wird ein in der Szene positioniertes Farbkalibriermuster (vgl. Abb. 3.25) dazu genutzt, die farbliche Situation an dieser Stelle für das Gesamtbild zu erfassen. Zum Einsatz kommt das Munsell ColorChecker[56] Kalibriermuster. Es vereint typisch in der Natur vorkommende Farben mit

[55] Unter Abberation versteht man Abbildungsfehler bei einer optischen Anordnung. Ihr Ursprung liegt in physikalischen Gesetzmäßigkeiten, beispielsweise der Tatsache, dass Licht verschiedener Wellenlänge beim Übergang in ein anderes Medium unterschiedlich stark gebrochen wird.

[56] Das Gitter besitzt insgesamt 24 verschiedene Farbkästchen. Deren englische Bezeichnung lautet: dark skin (0), light skin(1), blue sky (2), foliage (3), blue flower (4), bluish green (5), orange (6), purplish blue (7), moderate red (8), purple (9), yellow green (10), orange yellow (11), blue (12), green (13), red (14), yellow (15), magenta (16), cyan (17), white (18), neutral 8 (19), neutral 6.5 (20), neutral 5 (21), neutral 3.5 (22), black (23)

einer Grauwertskala.

Bei der Gruppe der Farbnormierungsverfahren werden gewisse Voraussetzungen angenommen, aus denen dann Normierungsvorschriften abgeleitet werden. Bei dem sog. CCN-Verfahren (Comprehensive Color Normalization) [FSC98] wird unter anderem davon ausgegangen, dass das reflektierte bzw. von der Kamera aufgenommene Licht nur von der Lichtquelle und der direkten Oberfläche abhängt, nicht jedoch von Streulicht innerhalb der Szene. „Die Grundidee dieses iterativen, zweistufigen Verfahrens ist es, jeden Farbpixel isoliert zu normieren, wobei Intensitätsschwankungen eliminiert werden; jeder Pixel wird in der Summe der Komponenten zu eins normiert. Dann wird jeder Farbkanal unabhängig global normiert, was Farbschwankungen unterdrückt; jeder Farbkanal wird so normiert, dass die Summe der Farbkomponenten im Kanal sich zu einem Drittel der Pixelanzahl addiert. Diese beiden Schritte werden iteriert, bis sich keine Änderung ergibt“ [Pau01]. Somit soll der Einfluss durch Änderung von Beleuchtungsrichtung und Farbtemperatur[57] neutralisiert werden. Bei dem ebenfalls zu den Normierungsverfahren zählenden Farbrotationsverfahren [PCN98] wird zugrunde gelegt, dass sich die Farbwerte in einem Bild durchschnittlich zu grau summieren. Infolge wird dies dazu genutzt, den Farbraum so zu transformieren, dass er diese Bedingung erfüllt. Dazu wird zunächst die Hauptachse der Punkte im Farbraum, beispielsweise durch Hauptkomponentenanalyse (*engl.* Principal Component Analysis), bestimmt. Die Hauptachse wird dann auf die sog. Unbuntachse (Schwarz-Weiss Achse) transformiert und im Anschluss skaliert.

Für die kalibrierbasierten Verfahren ist ein normiertes Kalibriermuster notwendig, mit dem in der Testszene die farbliche Situation erfasst wird. Die Abb. 3.25 zeigt ein solches Muster, das 18 Farbwerte und 6 Grauwerte enthält. Der Ablauf dieses Verfahrens ist in dieser Abbildung schematisch gezeigt. „Ausgangspunkt ist zum einen eine Szenenaufnahme unter Referenzbedingungen R und zum anderen die zu kalibrierende Szenenaufnahme mit unbekannten Bedingungen U gegeben. Beide Aufnahmen enthalten dasselbe Farbkalibriermuster. Bei der in der Referenzszene vorherrschenden Beleuchtung kann es sich um eine Normlichtart wie CIE D65, D50, A, B, C oder eine beliebige andere handeln, die approximiert werden soll. Zunächst wird zwischen einer zu kalibrierenden Szene und

[57] Die Farbtemperatur ist „eine zur Kennzeichnung der Strahlung einer Lichtquelle dienende Größe. Sie ist definiert als diejenige Größe eines schwarzen Strahlers, bei der dieser eine Strahlung der gleichen Farbart wie die zu charakterisierende Lichtquelle aussendet“ [Pau99]. Bei höherer Temperatur verschiebt sich das Maximum des Spektrums zur kürzeren Wellenlängen hin (Wiensches Verschiebungsgesetz). Eine Glühlampe hat eine Farbtemperatur von $2500 - 3000$ K, die Sonne bei wolkenlosem Himmel von ca. 6000 K.

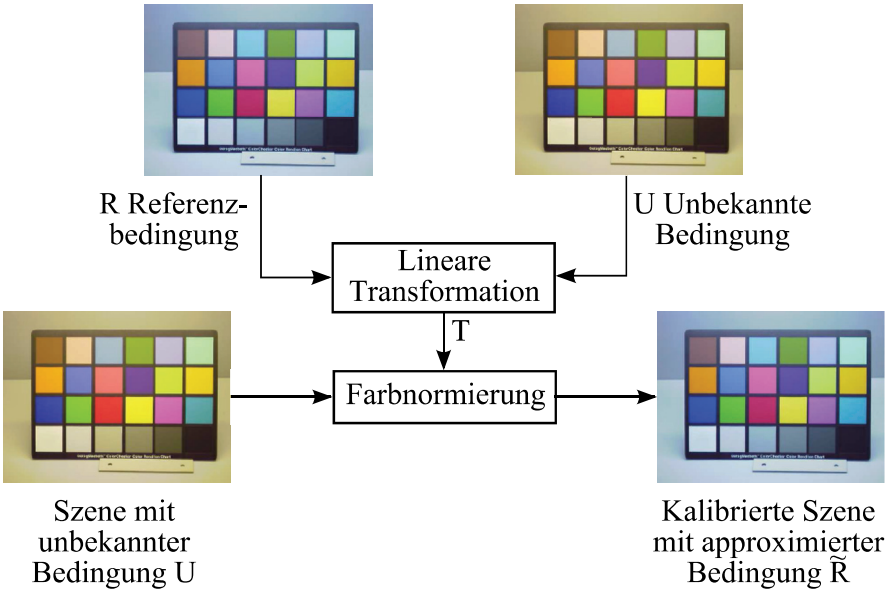


Abbildung 3.25: Farbnormierung durch Approximierung der Referenzbedingung nach [MHPM04]

der Referenzszene ein linearer Zusammenhang in Form einer linearen Transformationsmatrix hergestellt ... . Im Anschluss daran können Szenen mit denselben unbekannten Bedingungen U in Szenen unter ähnlichen Bedingungen R wie die der Referenzbedingungen mittels dieser Transformationsmatrix überführt werden .... “ [MHPM04]. Um die Transformation zu gewinnen, werden zunächst die Farbwerte der einzelnen Farbtafeln sowohl in der unter normierten Bedingungen als auch in der Realszene aufgenommen. Für die Berechnung der Transformation, welche einen Punkt aus der realen Aufnahme einen Farbpunkt der Referenzaufnahme zuordnet, wird ein überbestimmtes Gleichungssystem aufgestellt. Die Extraktion der einzelnen Farbwerte erfolgt in einem halbautomatisierten Prozess. Mittels SVD[58] wird die Lösung gefunden, die die Summe der quadratischen Fehler, d.h. der Abstände zwischen Referenzfarbwerten und projizierten Werten, minimiert. Mit dieser erhaltenen Transformation ist es dann möglich, zu beliebigen Punkten in der Szene, die unter unbekannten Bedingungen aufgenommen wurde, einen entsprechenden angenäherten Farbwert, welcher der Referenzumgebung entspricht, zuzuordnen. Ein weiteres normierungsbasiertes Verfahren wurde in [Fin07] entwickelt. Hier ist die Idee, die globale Transformationsvorschrift durch eine durch lokale Referenzpunkte definierte zu

[58] Singular Value Decomposition, *dt.* Singulärwertzerlegung

Abbildung 3.26: Farbreduktion auf Standardfarbwerte [Fin07], links: Originalbild, rechts: farbreduziertes Bild

ersetzen. Dazu werden zunächst für jeden Punkt im anzupassenden Bild die nächstgelegenen Farbwerte des Kalibriergitters bestimmt. Diese spannen ein lokales Koordinatensystem im Farbraum auf. Der zu transformierende Farbwert wird hiernach in Koordinaten dieses lokalen Koordinatensystems zum Ausdruck gebracht. Dieser Punkt wird danach vom lokalen Koordinatensystem des Testbilds in das korrespondierende lokale Koordinatensystem des Referenzbildes übertragen und die Farbwerte dort festgestellt. In [Fin07] wurde weiterführend ein histogrammbasiertes Objekterkennungsverfahren entwickelt.

Basierend auf den erhaltenen Farbwerten kann darauf aufbauend eine farbliche Beschreibung der Objekte abgeleitet werden. Dazu werden im ersten Schritt die kalibrierten Bildpunkte auf die Farben des GretagMacbeth ColorChecker[59] Kalibriergitters abgebildet, indem der jeweils nächste Farbwert bestimmt wird. Zugrunde gelegt für die Abstandsberechnung ist der euklidische Abstand. In der Abb. 3.26 ist dies für eine komplette Szene gezeigt. In großen Teilen des Bildes stimmen die Werte mit den erwarteten überein.

Für eine verbale Referenzierung der Farben innerhalb eines Dialogs ist eine große Zahl an Farbbeschreibungen aus mehrerlei Gründen nicht sinnvoll. Zum einen wäre es nicht möglich, eine hinreichend kohärente Farbbeschreibung eines Objekts zu erhalten, zum anderen wären die Differenzierungen für einen Dialog ungeeignet. Deshalb werden diese Farbwerte auf gewählte 7 Farb- oder Helligkeitsbeschreibungen zusammengefasst: Rot, Grün, Blau, Orange, Gelb, Hell und Dunkel. Rot ergibt sich aus den Farben des

[59] vgl. Fußnote 56 auf Seite 86

Kalibriermusters gemäßigtes Rot, Rot und Magenta, Grün aus den Farben Blattwerk, bläuliches Grün, gelb Grün und Grün, Blau aus den Farben Himmelsblau, blaue Blume, leicht violettes Blau, Blau und Zyanblau, Orange aus den Farben Orange und oranges Gelb, Gelb aus der Farbe Gelb, Hell aus den Farben helle Haut, weiß, verschiedenen Grauwerten und Dunkel aus dunkler Haut, Lila, einem Grauton und Schwarz.

Die Fehlzuordnungen haben ihre Ursachen in verschiedenen Faktoren. Zum einen wird zur Kalibrierung ein in einem halbautomatisierten Verfahren ausgewählter Bereich der einzelnen Farbfelder genutzt. Bedingt durch diese Auswahl, können die Farbwerte, die der Transformationsberechnung zugeführt werden, streuen. Desweiteren wird bei der Berechnung bei dem einen Verfahren eine globale Transformation des Farbraums durchgeführt. Dies bedeutet aber, dass die Transformation optimal im Sinne der Mittel der Farbwerte ist und einzelne Farbbereiche nur suboptimal transformiert werden können. Ein anderes Problem ist, dass die Referenzaufnahme nur unter schwachen Referenzbedingungen durchgeführt werden konnte, da keine Normlichtkabine zur Verfügung stand. An dieser Stelle wäre deshalb eine Erweiterung des Verfahrens im Hinblick auf eine Zuordnung zu den im kalibrierten Bild im Kalibriermuster vorhandenen Farbwerten zu untersuchen.

3.12 Geometrische Szenenbeschreibung

Ziel dieses Moduls ist es, eine umfassende, geometriebasierte Szenenbeschreibung zu erstellen, die danach in eine für den Sprachdialog aufgearbeitete Szenenbeschreibung überführt wird (vgl. Abschnitt 3.13, *Räumliche Relationen und Benutzerdialog*). Hier fließen die Ergebnisse ein, die bei der 3-D basierten Objekterkennung erhalten wurden, wie auch die zugehörigen Farbbeschreibungen, die über die 3-D/2-D Zuordnung berechnet wurden. Viele Objekte wie Decken und Wände, die vorher in Zwischenschritten berechnet wurden und für die weitere Auswertung nötig waren, sind an dieser Stelle nicht mehr von Bedeutung. Die Ausgabe erfolgt als xml-Datei und enthält folgende Elemente:

- **Funktionalitätsbasiert erkannte einfach strukturierte Objekte als Referenzobjekte**
  Aufgelistet werden hier die Objekte der Klasse Tisch und Boden. Da auf ihnen andere Objekte positioniert werden können, dienen sie als Referenzobjekte und können entsprechend angesprochen werden „*Zeige alle Objekte auf dem Tisch*“

- **3-D ansichtenspezifisch erkannte Objekte**
  Für Segmente, die auf den Referenzobjekten positioniert sind, werden die Ergebnisse des 3-D ansichtenbasierten Erkennungsmoduls aufgeführt. Es werden drei alternative Ergebnisse zusammen mit dem berechneten Grad der Zuordnung angegeben. Dies wird in dem später geführten Dialog dazu genutzt werden können, die Klassifizierung zu überprüfen.

- **Farbbeschreibung**
  Die Farbbeschreibung wird dazu genutzt, Objekte über ihre Farbe anzusprechen. Dies setzt hier voraus, dass sie eine markante Farbe besitzen. Die möglichen Eigenschaften sind: rot, grün, blau, orange, gelb, hell und dunkel. Es wird jeweils der Anteil am Gesamtbild des berechneten 2-D Ausschnitts angegeben.

- **Zentrierter Abstand und Richtungswinkel**
  Bei dieser Form der geometrischen Beschreibung werden die Mittelpunkte der objektorientierten Bounding-Boxen dazu genutzt, den horizontalen Abstand und Richtungswinkel zu berechnen.

- **Minimaler und maximaler Abstand und Richtungswinkel**
  Im Gegensatz zu den vorher aufgeführten Angaben werden hier die direkt gemessenen Punkte für die geometrische Beschreibung genutzt. Es werden minimaler und maximaler Punktabstand sowie minimaler und maximaler Richtungswinkel berechnet.

Diese xml-Datei enthält die Daten in einer Form, die einer linguistischen Anwendung noch nicht dienlich ist. Sie werden deshalb in einem von Joana Hois [HWBR06] entwickelten Verfahren in eine ontologisch basierte und für die Sprachverarbeitung genehme Form transformiert. Dies wird im nächsten Abschnitt weiter erläutert.

3.13 Räumliche Relationen und Benutzerdialog

Ziel dieser Komponente ist die sprachliche Kommunikation mit dem System über die räumliche Anordnung der Objekte in der Szene. Die sprachliche Interaktionsfähigkeit, bezogen auf das Umfeld, bildet allgemein eine Kernkomponente für zukünftige autonome Systeme, beispielsweise zur Routenbestimmung, wie bei dem autonomen Rollstuhlsystem

Rolli [MF07]. Bei dem hier vorgestellten Anwendungsszenario wird die räumliche Referenzierung von Objekten innerhalb einer Innenraumszene demonstriert. Dies könnte den Vorlauf zu einer beispielsweise haptischen Interaktion darstellen. Um mit dem System in einen verbalen Dialog über die in der Szene befindlichen Objekte treten zu können, müssen die Bereiche Szeneninterpretation und Sprachverarbeitung miteinander verknüpft werden. Dazu werden Ergebnisse der linguistischen und psychologischen Forschung verwendet, um die metrischen Daten der visuellen Szenenanalyse in verbale Beschreibungen zu konvertieren und umgekehrt [HWBR06].

Im Folgenden wird die Modellierung der Szene, wie sie für die räumliche und sprachliche Interpretation durchgeführt wird, vorgestellt. Dann wird auf die verwendete räumliche Relation eingegangen und zum Schluss der Ablauf der Dialoganwendung vorgestellt. Die Abbildung 3.27 zeigt die Anforderungen und den Ablauf dieses Moduls. Zunächst wird die von der visuellen Komponente des *ORCC*-Systems erstellte geometrische Szenenbeschreibung dazu genutzt, eine für den Dialog brauchbare Szenenbeschreibung zu erstellen. Diese besteht aus einer sogenannten Ontologie, die es erlaubt, neben der Modellierung auch Schlussfolgerungen auf der erstellten Wissensbasis zu berechnen. Eine Ontologie besteht konkret aus Begriffen (*engl.* concepts) bzw. Instanzen davon und Relationen zwischen diesen. Darüber hinaus existieren Regeln, mit denen man Schlussfolgerungen, basierend auf den Relationen zwischen den Instanzen, berechnen kann. Ein weiterer Vorteil von Ontologien ist deren Verknüpfbarkeit. Sie können in andere Ontologien eingebettet sein und können damit neben den speziell benötigten Konzepten, hier für die sprachliche Referenzierung von Objekten, auch andere, allgemeine Wissensbasen enthalten [HSB06]. Die Basis für die Beschreibung der zugrunde liegenden räumlichen Relationen bildet das Ontologie-Rahmenwerk Dolce (Descriptive Ontology for Linguistic and Cognitive Engineering), das ursprünglich für das Semantik Web Projekt [BL88] entwickelt wurde. Diese Ontologie wurde mit Schwerpunkt auf linguistische und kognitive Aspekte entwickelt und passt entsprechend zu der hier angestrebten Aufgabe. Als Upper-Ontologie ist sie, wenn gleich sie für besagte Zwecke entwickelt wurde, noch nicht auf eine bestimmte Domäne festgelegt, sondern liefert ein Gerüst von Begriffen, die es erlauben, darauf aufbauend eine spezifische Domäne in einer Domänen-Ontologie zu modellieren. Die Formulierung selbst erfolgt in der sog. Web Ontology Language (OWL). Bei Dolce werden die Hauptbegriffe in vier Kategorien unterschieden: *Endurants*, *Perdurants*, *Qualities* und *Abstracts*. Ein entsprechender Kategorienbaum zu Dolce findet sich in [CMO04]. Das Concept Endurants beschreibt Entitäten bzw. Instanzen, die unab-

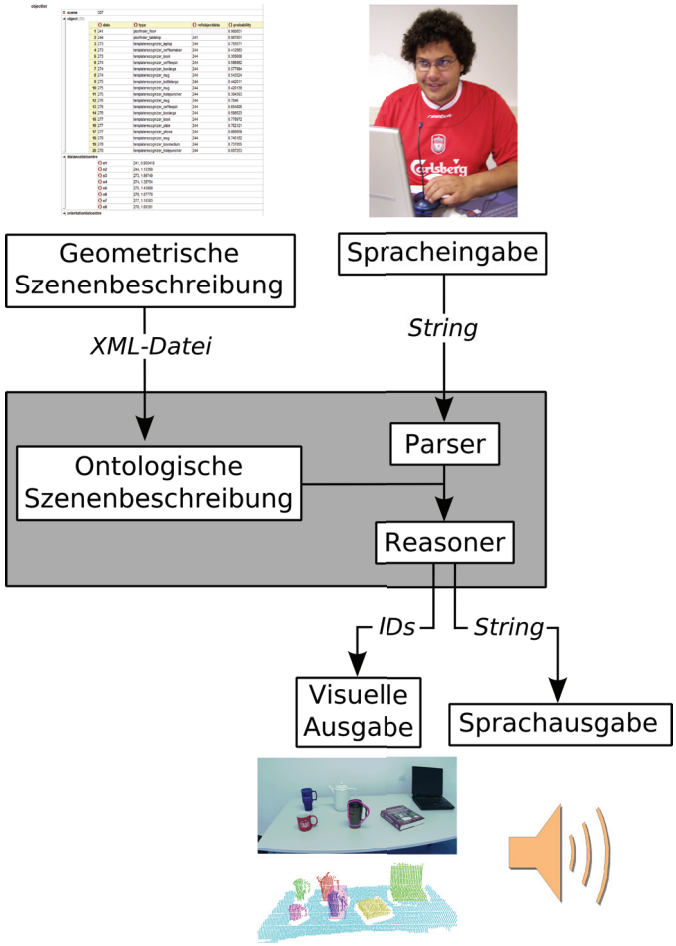


Abbildung 3.27: Dialogsystem

hängig von der Zeit vorhanden sind (enduring entities). Konkrete Gegenstände sind von deren Unterkonzept *Physical Object*. Perdurants beschreibt Entitäten, die in der Zeit anfangen (perduring entities), wie beispielsweise Ereignisse[60]. Das Concept *Qualities* einspricht Basisentitäten, die gemessen oder wahrgenommen werden können, wie Größe,

[60] „Classically, the difference between enduring and perduring entities (which we shall also call endurants and perdurants) is related to their behavior in time. Endurants are wholly present (i.e., all their proper parts are present) at any time they are present. Perdurants, on the other hand, just extend in time by accumulating different temporal parts, so that, at any time they are present, they are only partially present, in the sense that some of their proper temporal parts (e.g., their previous or future phases) may be not present. E.g., the piece of paper you are reading now is wholly present, while some temporal parts of your reading are not present any more. Philosophers say that endurants are entities that are in time, while lacking however temporal parts (so to speak, all their parts flow with them in time). Perdurants, on the other hand, are entities that happen in time, and can have temporal parts (all their parts are fixed in time).“ [CMO04]

Länge oder Farbe[61]. Die konkreten Werte von Instanzen des Concepts *Quality* entsprechen einer Entität des Concepts *Quale*, die einen Wert im aufgespannten Qualityraum darstellt. *Quale* ist ein Unter-Concept des Concepts *Abstract*, das sich dadurch auszeichnet, dass es einerseits nicht vom Type *Quality* ist und andererseits keine zeitlichen oder räumlichen Instanzen vom Type *Quality* zugeordnet haben. Die in der verwendeten Domänen-Ontologie[62] definierten Objekte sind funktional hierarchisch organisiert. Dies soll es zukünftig ermöglichen, bei einer Anfrage nach einem Trinkgefäß flexibel zu reagieren. Objekttypen, die bei dem Dialog als nicht definiert erkannt werden, werden als physikalische Objekte behandelt. Bevor auf die für diese Anwendung in [HWBR06] entworfene ontologische Darstellung der räumlichen Relation eingegangen wird, werden kurz die möglichen Arten von linguistischen räumlichen Relationen aufgeführt.

Wie bereits erwähnt ist die Absicht des Dialogs, Objekte in Bezug auf andere Objekte zu referenzieren. Die Ergebnisse linguistischer Forschung zeigen, dass es grundsätzlich drei verschiedene Referenzierungsarten gibt: Intrinsische, relative und absolute [Lev03], [The05]. Im Fall der intrinsischen Relationenbeschreibung wird das zu referenzierende Objekt, der Referent, in Bezug auf die intrinsische Orientierung des Bezugsobjekts (Relatum) beschrieben. Die intrinsische Orientierung eines Stuhls beispielsweise wäre typischerweise durch die Ausrichtung, in die ein potentieller Benutzer gerichtet wäre, gegeben. Bei der relativen Relationenbeschreibung wird der Referent aus dem Blickwinkel eines dritten Objekts in Bezug auf das Relatum beschrieben. Im Falle der absoluten Relationenbeschreibung entspricht der Blickwinkel einer festen externen Orientierung. In der Anwendung hier wird von einer relativen Orientierung ausgegangen, wobei linguistisch hauptsächlich projektive Ausdrücke von Relationen genutzt werden. Der Blickwinkel entspricht dabei der des Aufnahmesystems und damit beispielsweise der Situation einer auf einem Rollstuhl montierten Aufnahmeeinheit. In einer weiterentwickelten Version könnte dies auch auf eine im Raum zu verfolgende Person ausgerichtet werden.

Für die Zuordnung der sprachlichen Relation zur geometrischen Relation werden sog. nicht überlappende Akzeptanzbereiche (*engl.* acceptance areas) verwendet [Her94], [MT06]. In unserem Fall entspricht dies in der Horizontalebene einer einheitlichen Sek-

[61] „'Quality' is often used as a synonymous of 'property', but this is not the case in DOLCE: qualities are particulars, properties are universals." [CMO04]

[62] Eine Domänen-Ontologie bezeichnet die für eine konkrete Domäne definierte Ontologie im Gegensatz zu einer generell gültigen Upper-Ontologie.

toraufteilung von 90°. Die Sektoren für die projektiven Ausdrücke „behind“ und „in front of“ liegen symmetrisch auf der Verbindungslinie, die durch die Mittelpunkte von Scanner (Ursprung) und Relatum gegeben ist. Entsprechend nehmen die Ausdrücke „to the left“ und „to the right“ die beiden anderen Sektoren auf. Als geometrische Orte von Referenz und Relatum werden die aus den Bounding-Boxen abgeleiteten zentrierten Entfernungswerte und Richtungswinkel verwendet. Diese scharfe Aufteilung der Regionen kann in den Grenzregionen natürlich zu einem volatilen Verhalten führen, da bereits kleine Positionsänderungen eine Zuordnung in einen anderen Sektor bewirken können. Da es an dieser Stelle vorrangig um die prinzipielle Umsetzung der Verknüpfung von Erkennung und Sprache geht, kann dieser Umstand akzeptiert werden.

Beim Aufbau der Domänenontologie werden für alle möglichen und sinnvollen Objektpaare die sprachlichen Relationen bestimmt und in der Domänenontologie abgespeichert. Diese stehen dann bei dem folgenden Sprachdialog direkt zur Verfügung und können durch den Reasoner interpretiert werden. Dem eigentlichen Benutzerdialog ist eine sog. Trainingsphase optional vorgeschaltet. Diese soll es ermöglichen, per Benutzerdialog unsichere Klassifizierungsergebnisse zu revidieren [HWBR06]. Beim Sprachdialog werden die von der Spracheingabe erhaltenen Strings auf die räumlichen Akzeptanzbereiche geparst. Eine semantische Analyse ist nicht nötig, da für die auf der Spracheingabeseite eingesetzte Spracherkennung eine auf die Aufgabe abgestimmte Grammatik definiert ist. Der Reasoner durchsucht dann die aufgebaute Domänenontologie nach den passenden räumlichen Relationen, bei denen die Objekte in der entsprechenden Art zueinander positioniert sind. Geht es speziell darum, den nächsten Nachbarn innerhalb des entsprechenden Akzeptanzbereichs zu bestimmen, wird eine spezielle geometrische Betrachtung angehängt. Als Ausgabe werden zum einen die Identitäten der entsprechenden Objekte berechnet und zum anderen ein Sprachstring erzeugt, der an die Sprachsynthese weitergeführt wird.

Kapitelzusammenfassung

Dieses Kapitel betrachtet die einzelnen Module der entwickelten Umgebung *ORCC*. Ziel des Systems ist es eine komplette Szenenanalyse einer Innenraumszene zu erreichen. Die einzelnen Module werden in der Reihenfolge ihrer Anwendung demonstriert, sind jedoch nicht auf ein einzelnes Beispiel beschränkt, sondern erklären die Aspekte an einer Vielzahl von verschiedenen Aufnahmen in unterschiedlichen Szenen.

Im kommenden Kapitel wird nun der Ablauf des Systems mithilfe einer gleichbleibenden Szene betrachtet. Dazu werden die in diesem Kapitel beschriebenen Methoden verwendet und ihre jeweiligen Ergebnisse analysiert.

4 Anwendung

Die Vision, Maschinen durch einen möglichst offenen Dialog in einer möglichst offenen Szene zu instruieren, setzt viele Fähigkeiten voraus. Ein notwendiger Beitrag ist es, die Maschine mit einer Beschreibung der in der Szene vorhandenen Objekte zu versorgen. In Erreichung dieser Aufgabe haben sich große Teile dieser Arbeit damit beschäftigt, Objekte zu segmentieren und zu kategorisieren. Es entsteht hier eine Szenenbeschreibung, die Informationen über die Position, Kategorisierungenalternativen und Farbbeschreibungen enthält. Die Entstehung dieser Szenenbeschreibung wird in diesem Kapitel komplett und beispielhaft an einer Innenraumszene dargestellt. Der sich anschließende Dialog nutzt diese Information, um Gegenstände unter Verwendung räumlicher Relationen anzusprechen. Somit wird demonstriert, dass das entwickelte System nicht nur in der Lage ist, einzelne Objekte zu identifizieren, sondern auch die räumliche Anordnung der Teile zu erfassen und in einem sprachgesteuerten Dialog einzusetzen.

Bei der in diesem Kapitel untersuchten Beispielszene handelt es sich wieder um eine reale Innenraumszene, bestehend aus einem mit Objekten belegten Tisch mit zwei Stühlen und einem Abfalleimer davor. In der Abb. 4.1 sind die einzelnen Verarbeitungsschritte und ihre Abhängigkeiten aufgezeigt. Ausgehend von der 3-D und 2-D Aufnahme der Szene werden folgende Zwischenschritte abgearbeitet:

- Oberflächensegmentierung: Zusammenhängende Oberflächensegmente werden detektiert
- Funktionalbasierte Klassifizierung: Raumbegrenzungen und Tischoberfläche werden detektiert
- Punktextraktion: Raumbegrenzungen werden entfernt
- Dichtebasierte Segmentierung: Auf dem Tisch oder Boden stehende Objekte werden detektiert
- Formbasierte Klassifizierung: Zuvor detektierte Objekte werden klassifiziert
- Umgebungsboxberechnung: Ausdehnungen der zuvor detektierten Objekte werden bestimmt
- 3-D / 2-D Kalibrierung: Kameramatrix und Orientierung zwischen Scanner und Kamera werden berechnet
- Ausschnittsberechnung: Zuvor berechnete dreidimensionale Segmente werden ins 2-D Bild übertragen
- Farbnormierung: Farbwerte werden mittels Farbmuster normiert
- Farbbeschreibung: Farbverteilung der Segmente wird berechnet
- Szenenbeschreibung und Dialog: Einzelergebnisse werden zu einer Gesamtbeschreibung zusammengefasst. Basierend auf der Szenenbeschreibung, kann mit dem System über die räumliche Situation der Objekte gesprochen werden

Der linke Teilast beinhaltet im wesentlichen die Schritte, die auf den 3-D Ausgangsdaten beruhen, der linke die Schritte, die hauptsächlich mit den Farbbildern zusam-

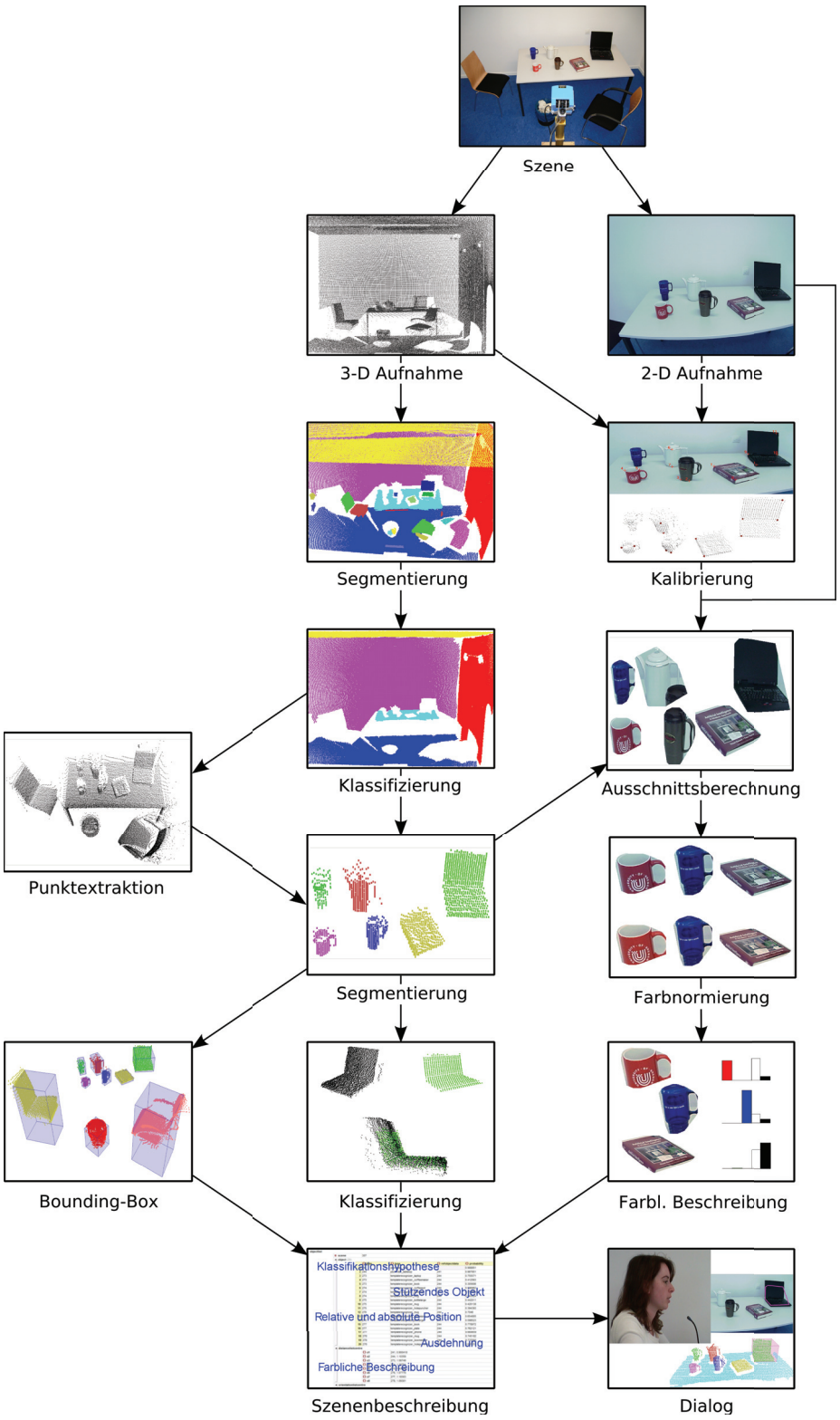


Abbildung 4.1: Einzelschritte hin zur Szenenbeschreibung und einem Mensch-Maschine Dialog

menhängen. Am Ende münden alle Berechnungen in der Szenenbeschreibung und dem Benutzerdialog.

Oberflächensegmentierung

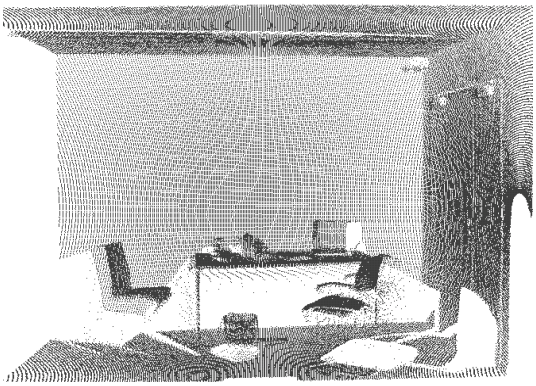
Die Bilder in Abb. 4.2 zeigen in Teilbild (a) die Tiefenbildaufnahme des Scanners (vgl. Abschnitt 3.1, *Sensordaten*) und in Teilbild (b) die Farbbildaufnahme der am Scanner montierten Farbbildkamera. Das Bild 4.3 gibt einen Gesamtüberblick über die Szene, die aus einem Tisch mit den darauf stehenden Gegenständen Kaffeekanne, Kaffeebecher, Buch und einem Laptop besteht. Auf dem Boden stehen zudem zwei Stühle und ein Abfalleimer. Im ersten Verarbeitungsschritt werden im 3-D Bild zusammenhängende und hinreichend eben geformte Flächenstücke segmentiert (vgl. Abschnitt 3.2, *Oberflächensegmentierung*). Das Ergebnis ist in Abb. 4.4 zu sehen. Wie in dieser Abbildung zu erkennen ist, werden neben den Raumbegrenzungen auch der Tisch und einzelne Segmente der Tischobjekte erkannt. Beim linken Stuhl werden die Rückenlehne und die Sitzfläche, beim rechten Stuhl wird aufgrund eines schleifenden Schnitts zwischen Scanner und Stuhl nur dessen Sitzfläche erkannt.

Funktionalbasierte Klassifizierung

Bei der funktionalbasierten Klassifizierung (vgl. Abschnitt 3.3, *Erkennung einfach strukturierter Objekte*) geht es zum einen darum, im Sinne der Eigenstruktur einfache Objekte wie Decken, Wände, Böden und Tische zu erkennen. Das Ergebnis für diese Szene ist in Abb. 4.5 gezeigt. Deren robuste Detektion ist wichtig, da sie für die Trennung der Umgebungsflächen bzw. der Aufstellflächen von den weiteren Objekten notwendig sind. Zusätzlich gibt es auch die Möglichkeit, strukturell komplexere Objekte wie Stühle zu erkennen (vgl. Abschnitt 3.5, *Strukturbasierte Objekterkennung*). Dort werden dann mehrere Segmente des zu klassifizierenden Objekts und ihre Relationen zueinander, als auch zur Umgebung, betrachtet.

Punktextraktion

In diesem Schritt werden alle zuvor erkannten Umgebungsflächen dazu genutzt, die Gegenstandspunkte zu extrahieren. Dazu werden der Punktwolke alle raumelementnahen Punkte entfernt.



(a) 3-D Aufnahme

(b) 2-D Aufnahme

Abbildung 4.2: Szenenaufnahme und Segmentierungsergebnis

Abbildung 4.3: Externe Szenenaufnahme

Abbildung 4.4: Oberflächensegmentierung

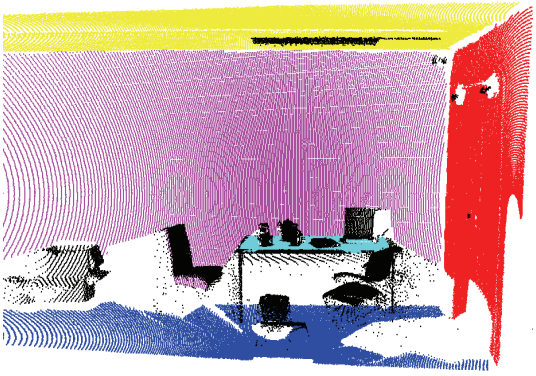


Abbildung 4.5: Erkannte Wände, Decke, Fußboden, Tischoberfläche und extrahierte Punkte

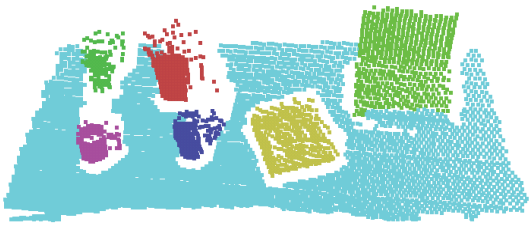


Abbildung 4.6: Dichtebasierte 3-D Segmentierung

Dichtebasierte Segmentierung

Während bei der zuvor durchgeführten Segmentierung die lokale Ausrichtung der Punkte zur Segmentierung genutzt wurde, wird hier eine dichtebasierte Segmentierung durchgeführt. Dabei spielen zwei Faktoren eine Rolle: Das Vorhandensein einer Bezugsfäche und die lokale Punktdichte. Die Abb. 4.6 zeigt das Ergebnis für die Tischelemente. Bezogen auf die Tischebene werden Punkte, die einen maximalen Abstand zu ihrem nächsten Nachbarn nicht überschreiten, zu einem Segment zusammengefasst (vgl. Abschnitt 3.6, *Segmentierung auf funktionalen Ebenen*).

Formbasierte Klassifizierung

Die erhaltenen Segmente werden nun durch ein formbasiertes dreidimensionales Verfahren klassifiziert (vgl. Abschnitt 3.8, *3-D ansichtenspezifische formbasierte Objekterkennung*). Dazu wird das zu klassifizierende Objekt gegen Referenzobjekte ausgerichtet und deren Passgüte zur Beurteilung des Klassifikationsergebnisses herangezogen. In Abb. 4.7 ist das Registrierungsergebnis für einen Laptop gezeigt, wobei das Referenzobjekt in schwarz gehalten ist.

Umgebungsboxberechnung

Ein sich der Segmentierung anschließender Schritt ist die Berechnung der Umgebungsbox (vgl. Abschnitt 3.7, *Umgebungsboxberechnung*). Das von Meyer [Mey05] entwickelte Verfahren ist in der Lage diese Umgebungsbox objektorientiert und auch unter gewisser Nichtberücksichtung von Schattenpunkten zu berechnen. Innerhalb dieser Szenenanalyse wird die Umgebungsbox zur Lokalisierung des Objektmittelpunkts benutzt. Man kann sie auch zur Objektklassifikation verwenden [WR06]. Für ausgewählte Objekte ist die erhaltene Umgebungsbox in der Abb. 4.8 dargestellt.

3-D / 2-D Kalibrierung

Mit diesem Schritt wird die rein 3-D basierte Berechnung verlassen und der Übergang zu 2-D Daten geschaffen. Mit Hilfe von Benutzervorgaben werden korrespondierende 3-D/2-D Punktepaare (vgl. Abb. 4.9) bestimmt, aus denen dann die 3-D/2-D Transformation berechnet wird (vgl. Abschnitt 3.9, *2-D/3-D Systemkalibrierung*).

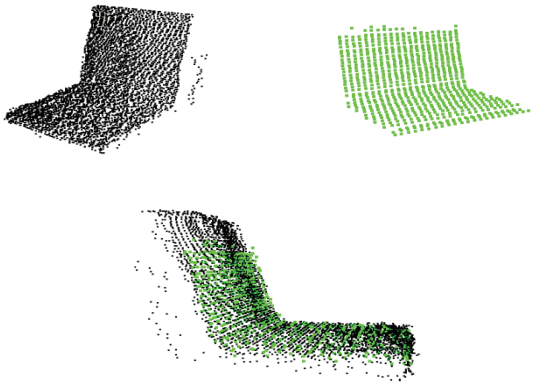


Abbildung 4.7: 3-D Registrierung eines Laptops zur formbasierten Klassifizierung

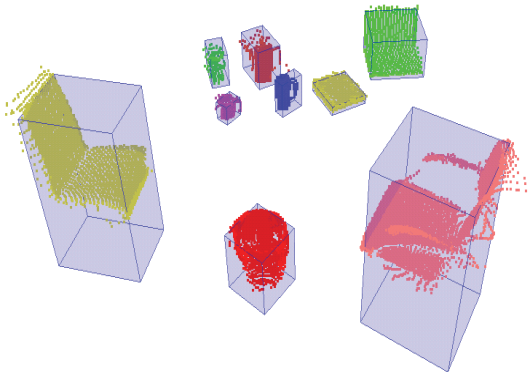


Abbildung 4.8: Umgebungsboxberechnung zur Positionsbestimmung

Abbildung 4.9: Punktkorrespondenzen zur Berechnung der 3-D/2-D Transformation

Abbildung 4.10: 2-D Ausschnittsberechnung

Abbildung 4.11: Ergebnis der Farbkalibrierung

Abbildung 4.12: Farbbeschreibungsvektor

Ausschnittsberechnung

Basierend auf den zuvor erhaltenen dreidimensionalen Segmenten, kann nun unter Verwendung der berechneten Transformation zwischen Scanner und Kamera der korrespondierende Bildausschnitt berechnet werden. In der Abb. 4.10 ist das Ergebnis für die Tischobjekte gezeigt (Abschnitt 3.10, *Texturausschnittsberechnung*).

Farbnormierung und Farbbeschreibung

Die zuvor erhaltenen Texturausschnitte werden in diesem Schritt in eine Farbbeschreibung überführt. Im ersten Schritt wird eine farbliche Normierung der Texturausschnitte durchgeführt, deren Ergebnisse in Abb. 4.11 zu sehen sind. Damit sollen Variationen in den Beleuchtungsverhältnissen eliminiert werden (vgl. Abschnitt 3.11, *Farbnormierung und Farbbeschreibung*). Auf diesen Texturausschnitten wird dann eine farbliche Beschreibung erstellt, deren Farbverteilung in Abb. 4.12 zu sehen ist.

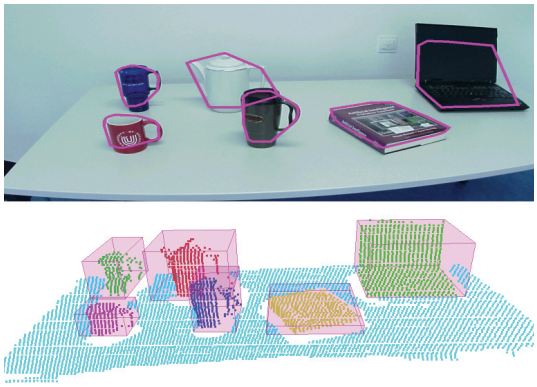
Szenenbeschreibung und Dialog

Die vorhergehenden Verarbeitungsschritte sorgen in ihrer Gänze für eine vielfältige, geometriebasierte Beschreibung der aufgenommenen Szene. Für die sprachbasierte Interaktion mit dem System werden diese in eine passende Szenenbeschreibung überführt (vgl. Abschnitt 3.13, *Räumliche Relationen und Benutzerdialog*). Es besteht die Möglichkeit, falsche Klassifizierungen zu korrigieren oder unbekannte Objekte dem System bekannt zu machen.

In der Abb. 4.13 sind vier Dialoge, die mit dem System geführt wurden, im Ergebnis dargestellt. Die Dialoge werden in englischer Sprache geführt. Beim Dialog A fragt der Benutzer das System, *What is the object to the left of the dark mug.* Das System erwidert mit der Sprachausgabe *There is a mug.* Die Tasse wird zudem in der dreidimensionalen Darstellung durch eine achsenorientierte Umgebungsbox, in der zweidimensionalen durch ein Polygon hervorgehoben. Dies ist in der Abb. 4.13(a) dargestellt. Auf die Frage *What is the object in front of the laptop*, antwortet das System mit *There is a book* (vgl. Abb. 4.13(b)). Die Frage *What are the objects on the table* wird beantwortet mit *There are six objects, a laptop, a book, mugs, a coffeepot* (vgl. Abb. 4.13(c)). Die Aufforderung *Show me the objects to the right of the blue mug* erhält als Reaktion *There is a coffee pot* (vgl. Abb. 4.13(d)).

(a) Dialog A, What is the object to the left of the dark mug

(b) Dialog B, What is the object in front of the laptop



(c) Dialog C, What are the objects on the table

(d) Dialog D, Show me the objects to the right of the blue mug

Abbildung 4.13: Sprachliche Referenzierungen der Tischobjekte

5 Zusammenfassung und Ausblick

Die Schaffung eines Systems, das autonom die Umwelt exploriert, seine Interaktion mit Menschen und der Umwelt situativ anpasst und selbständig aus der Erfahrung Verhaltensänderungen ableitet, ist ein visionäres Fernziel der Robotik und der Kognitionswissenschaften.

Das in diesem Zusammenhang entwickelte und hier vorgestellte System *ORCC* untersucht den Aspekt der visuellen Wahrnehmung und Szenenanalyse und bildet als solches die Grundlage für viele mögliche darauf aufbauende Anwendungsszenarien, seien sie autonomer oder interaktiver Art, mit dem Ziel einer weiteren Analyse der Szene oder einer haptischen Aktion mit den Gegenständen der Szene. Das hier umgesetzte Anwendungsbeispiel ist das einer interaktiven, sprachbasierten Mensch-Maschine Interaktion. Diese Anwendung ist beispielhaft für die Konstruktion eines künstlichen Systems, da sie die Verknüpfung einer durch den Roboter erzeugten, technisch basierten Wahrnehmung der Szene mit dem durch den Menschen abgebildeten kognitiven Modell dieser Szene, erfordert. Um dies zu erreichen, müssen psychologische und linguistische Erkenntnisse, hier speziell im Bereich der Raumkognition, genutzt werden.

Dieses System zeichnet sich durch folgende Eigenschaften aus:

- Die zu untersuchenden Szenen sind Innenräume, die Erkennung ist aber nicht auf eine spezielle Klasse von Objekten beschränkt, sondern es können beliebige im Alltag vorkommende Objekte genutzt werden.

- Neue Gegenstände können dem Erkennungsmodul hinzugefügt werden, ohne dass das System neu trainiert werden muss. Dabei wird berücksichtigt, dass die Objekte teilweise verdeckt sein können.

- Das System ist in der Lage mehrere Datenquellen zu fusionieren. Dadurch können die technischen Vorzüge verschiedener Datenquellen in einem multimodalen Erkennungsprozess genutzt werden. Somit können sowohl die formbeschreibenden

dreidimensionalen Daten als auch die farbbeschreibenden zweidimensionalen Daten für die Erkennung und Referenzierung eines Objekts verwendet werden.

- Das System erstellt eine geometriebasierte Szenenbeschreibung, die in eine für die Sprachkommunikation dienliche Form umgewandelt wird. Die Szenenbeschreibung enthält neben alternativen Klassifizierungsergebnissen auch eine Lage- und Farbbeschreibung der Objekte.

- Es ermöglicht, sprachlich mit dem System über die erkannten Objekt und deren räumliche Anordnung zu kommunizieren. Das System gibt seinerseits visuelle und sprachliche Rückmeldung über die erkannten Objekte.

Eine Szenenanalyse stellt verschiedene Anforderungen: einerseits muss das System eine möglichst gute Objekterkennung leisten, andererseits muss es diese Informationen über die Objekte möglichst offen für sich anschließende Fragestellungen aufarbeiten. In dieser Arbeit wird sich auf Innenraumszenen beschränkt, wobei die Wahl der Objekte offen gehalten wird und einfach erweiterbar ist. Die Besonderheit dabei ist, dass das System verschiedene Datenquellen, distanz- und farbbasierte, derart fusioniert, dass eine initiale dreidimensionale Segmentierung von in der speziellen Formausprägung unbekannten Objekten auf den zweidimensionalen Bereich übertragen wird. Dadurch wird es ermöglicht, texturunbekannte Objekte zu segmentieren und farblich zu beschreiben. Das System ist in der Lage, dieses Wissen für eine bidirektionle Mensch-Maschine Interaktion aufzubereiten. Dabei zeigt sich, wie wichtig bereits für einfache Kommunikationsstränge das Wissen über die kognitiv räumliche Modellierung der Welt durch den Menschen ist, hier genutzt zur räumlichen Referenzierung von Gegenständen.

Verbesserungen oder Weiterentwicklungen könnten in verschiedener Richtung erreicht werden, zum einen durch die Erweiterung der multimodalen Nutzung und durch die Verbesserung der Methoden, und zum anderen durch Anpassung der globalen Strategie:

- Multimodale Verarbeitung
  Neben der Verbesserung der einzelnen Methoden, kann auch die Erweiterung der Nutzung verschiedener Datenquellen zur Leistungserweiterung beitragen:

  Im Sinne der Wiedererkennung von Objekten wäre es sinnvoll, zuvor auf 3-D Daten funktional erkannte Objekte über ihre 2-D Textur wiederzuerkennen. Der Vorteil

ist, dass farbbasierte Erkennungsmethoden relativ robust und schnell und in gewissem Maß unabhängig von Verdeckungen sind. Zudem liefern Farbbildkameras ihre Daten in höherer Frequenz als dies ein Laser-Scanner erlaubt. Für die funktionale Erkennung hingegen haben die 3-D Daten tendenziell Vorteile, da sie direkt die dreidimensionale Form des Objekts repräsentieren.

Die Segmentierung von unbekannten oder formunspezifischen Objekten ist eine schwierige Aufgabe. Hier könnte die Erweiterung der hier rein auf 3-D Daten basierenden Segmentierung Vorteile bringen, beispielsweise bei der Texturausschnittsberechnung. Der verwendete 3-D Sensor liefert im Vergleich zur Farbbildkamera eine viel geringere Winkelauflösung. Ausgehend von der hier berechneten dreidimensionalen Lösung, könnte man diese als Startlösung einer im zweidimensionalen Bild weitergeführten Berechnung verwenden.

- Methoden
  Die Verbesserung der einzelnen Methoden verbessert direkt auch die Gesamtperformanz des Systems. Ein Ansatzpunkt wäre beispielsweise die Verbesserung des der 3-D basierten Erkennung zugrunde liegenden Registierungsvorgangs.

- Szenenanpassung
  Schaut man sich existierende Erkennungssysteme an, erkennt man, dass diese gut innerhalb einer bestimmten Domäne und Aufgabenstellung funktionieren. Wollte man dieses System also dahingehend erweitern, dass es domänenunspezifischer arbeitet, müsste es selbständig in der Lage sein zu erkennen, in welcher Umgebung und in welcher Handlungssituation es ist. Abhängig davon würden dann die passenden Erkennungsstrategien gewählt werden. Dieser Wahrnehmungszustand bzw. die Erwartung der Wahrnehmung basiert auf der Historie des Erlebten und der aktuellen Wahrnehmung und Handlungsabsicht. Ein solches System ginge in Komplexität weit über das hier angestrebte Anwendungsszenario hinaus. Es wäre sehr viel Allgemeinwissen, aber auch das Wissen über menschliche Aktionsmuster und Denkmodelle nötig, um die Situation richtig zu interpretieren. Es müsste eine Modellierung entwickelt werden, die dieses mächtige Wissen integriert, die vor allem aber parametrisierbar, das heißt situativ adaptierbar, bleibt. Ein Ansatzpunkt hierfür sind Ontologien, die innerhalb dieser Arbeit im Zusammenhang mit räumlichen Kalkülen verwendet wurden.

Insgesamt zeigt sich, dass durch die Kombination verschiedener Aufnahmesysteme eine hinreichend vollständige Szenenbeschreibung einer mit beliebigen Objekten ausgestatteten Büroumgebung erreicht werden kann. Diese kann für eine Interaktion zwischen Mensch und Maschine eingesetzt werden. Eine beispielhafte Anwendung des Systems findet sich als Video `http://www.informatik.uni-bremen.de/~wuenstel/ki-video/`.

A Anhang

A.1 ORCC-System: Technische Beschreibung

Das *ORCC*-System wurde mit der Absicht konstruiert, als Entwicklungsplattform für die in dieser Arbeit beschriebene Szeneninterpretation mit anschließender Mensch-Maschine-Interaktion zu dienen. Eine technische Hauptanforderung an das System ist, dass es einfach um weitere Module erweiterbar ist und die Ergebnisse wiederum einfach von anderen Modulen oder auch externen Gruppen ausgelesen werden können. Zudem sollte das System weitestgehend plattformunabhängig entwickelt werden.

Als Ergebnis entstand ein Client-Server basierter Entwurf, bei dem Clients verschiedene Dienste anbieten und der Server diese Clients verwaltet (vgl. Abb. A.1). Diese Dienste können hardwarenah sein, wie bei der Bildaufnahme oder der Sprachkommunikation, es kann sich aber auch um reine softwarebasierte Dienste handeln, die beispielsweise die Punkte einer Aufnahme segmentieren. Die im Verbund aktiven Clients melden ihre Dienste beim Server an, der die aktiven Dienste und die dazugehörigen Rechner auflistet und diese Liste an alle Clients weitergibt. Somit kann jeder Client einen Dienst, den er als Vorverarbeitungsschritt benötigt, aktivieren. Diese Programme können, müssen aber nicht, auf dem eigenen Rechner installiert sein. Neben den Diensten kann der Benutzer auch gezielt Objekte, beispielsweise segmentierte Punktmengen, auswählen. Damit können zielgerichtet die Auswirkungen von Operation im Versuchsstadium getestet werden. Die Ergebnisse werden jeweils in einer SQL-Datenbank abgelegt. Dies erleichtert einerseits den gleichzeitigen Zugriff von mehreren Clients, andererseits hat es in der Entwicklungsphase den Vorteil, dass sehr einfach, auch plattformübergreifend, die

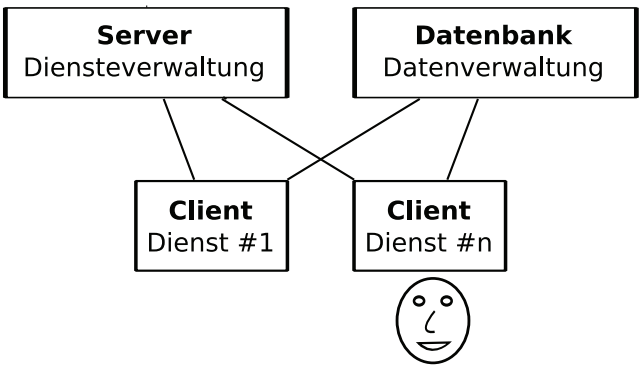


Abbildung A.1: Client-Server Modell des *ORCC*-Systems

Daten ausgetauscht werden können. Die Implementierung selbst erfolgte in C++ unter Verwendung der allgemein numerischen Bibliotheken Numerical Recipes [PTVF92], der Geometrie bezogenen Bibliothek CGAL[63], der Vision bezogenen Bibliothek OpenCV[64], der GUI-Bibliothek Qt[65] und der Szenengraph bezogenen Bibliothek OpenSG[66].

Durch diese Struktur konnte sichergestellt werden, dass zu jedem Zeitpunkt neue Projektmitglieder sich mit ihren speziellen Aufgaben in die Weiterentwicklung des Gesamtsystems einfügen konnten. Da die Grundstruktur eines Dienstes unverbindlich vorgegeben ist, mussten keine neuen Module zur Darstellung oder Datenbankkommunikation entwickelt werden, sie konnte vielmehr übernommen oder angepasst werden.

[63] Computational Geometry Algorithms Library, `http://www.cgal.org`
[64] Intel®Open Source Computer Vision Library, `http://www.sourceforge.net/projects/opencvlibrary`
[65] Qt, `http://www.trolltech.com`
[66] OpenSceneGraph, `http://www.openscenegraph.org`

A.2 Optische Wahrnehmung

Es gibt einige eindrucksvolle Beispiele, die zeigen, wie unser Gehirn die optischen Sinneseindrücke um eigene Interpretationen erweitert und damit eine biologisch gefilterte oder erweiterte Form der Wahrnehmung konstruiert. Diese Gesetzmäßigkeiten finden in den, Anfang des 20sten Jahrhunderts entdeckten, Gestaltfaktoren bzw. Gestaltgesetzen[67] ihren Ausdruck [Dah06]. Die Gestaltpsychologen „untersuchten systematisch die Regeln, nach denen unser Sehsystem Objektmerkmale zu kohärenten Einheiten - zu Gestalten - zusammenfasst“ [Eng06]. So sehen wir beispielsweise bei der Abb. A.2 links ein inliegendes Dreieck, obgleich es faktisch nicht existiert. Das entsprechende Gestaltgesetz, das diesen Vorgang beschreibt, heißt *Gesetz der guten Fortführung*. Entsprechende neurologische Untersuchungen haben gezeigt [MK05]: „Die Großhirnrinde betätigt sich als ‚kreativer Lückenfüller‘, der aktiv fehlende Sinnesdaten zu ‚plausiblen‘ Gesamteindrücken ergänzt.“ Es ermöglicht uns, unvollständige Bilder im Allgemeinen durch sinnhaftige Eindrücke zu ergänzen[68]. Ein anderes Beispiel für ein Gestaltgesetz ist in Abb. A.2 rechts zu sehen [Dah06]. Wenn man dieses Bild betrachtet, entsteht zunächst der

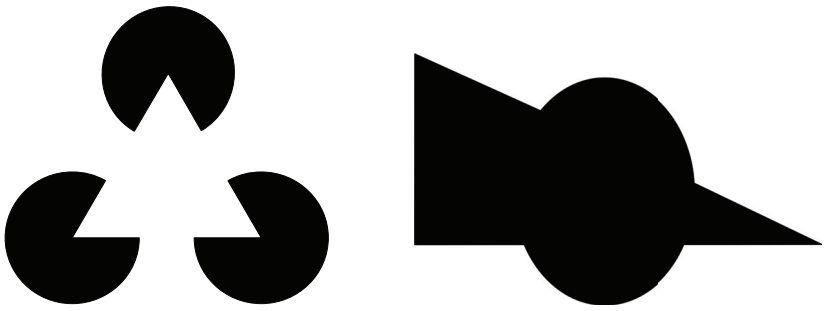


Abbildung A.2: Gestaltgesetze
links: Gesetz der guten Fortführung (Kanizsas Dreieck)
rechts: Gesetz der guten Gestalt oder Gesetz der Vertrautheit

Eindruck eines Dreiecks und einer Ellipse. Das zugehörige Gestaltgesetz heißt *Prägnanz* oder *Gute Gestalt*. Diese Fähigkeit, einfache Objekte innerhalb einer Anordnung von verschiedenen Objekten oder trotz Teilverdeckung zu erkennen, entstand vermutlich als

[67] Bedeutender Mitbegründer der Gestaltpsychologie ist Max Wertheimer, 1880–1943.
[68] Einen ähnlichen Effekt gibt es bei der akustischen Wahrnehmung. Die Pfeife der englischen Polizei beispielsweise produziert zwei Töne in leicht verschobener Tonlage. Die dadurch produzierte Schwebung ergänzt unser Wahrnehmungssystem zu einem entsprechend tiefen Ton.

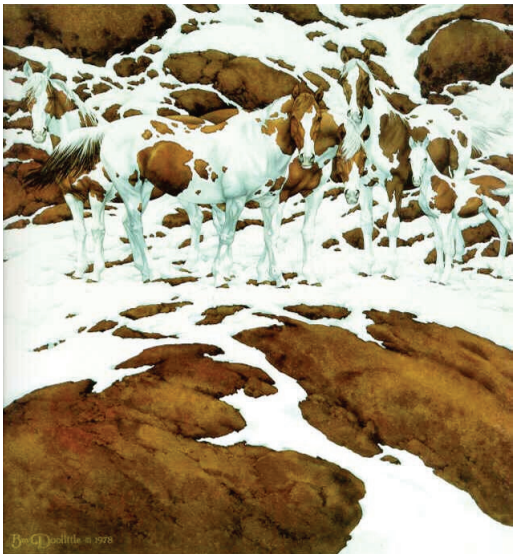


Abbildung A.3: Bev Doolittle: Pintos 1978

Überlebensvorteil des Menschen, um auch in widrigen Umständen entweder Gefahrenobjekte besser zu erkennen oder bei der Jagd oder beim Sammeln Nahrungsmittel schneller zu sehen. Betrachtet man das gleiche Bild, kann bei einigen Menschen jedoch auch der Eindruck der Frontpartie eines Rennwagens entstehen. In diesem Fall lautet das Gestaltgesetz das *Gesetz der Vertrautheit oder Bedeutung* und besagt, dass Objekte eine Gesamtheit bilden, wenn sie dadurch eine Bedeutung erlangen, also dem Betrachter vertraut sind. Ebenfalls das Gesetz von der Vertrautheit spielt neben anderen, wie *gestaltgerechter Linienfortsetzung*, bei der Erkennung der Pferde in Abb. A.3[69] ein Rolle. Darüber hinaus zeigt diese Abb. A.3 auch das den Gestaltpsychologen wichtige Thema der Figur-Grund-Trennung auf [Gol02].

Ein Beispiel, das für eine andere Klasse von Effekten steht, welche direkt mit dem Aufbau des Auges zusammenhängen, zeigt das Bild A.4 [GHKW05]. Schaut man eine gewisse Zeit auf das Fadenkreuz im linken Bild und dann auf das Fadenkreuz im rechten leeren Bild, erscheinen türkisfarbene Quader. Dies rührt daher, dass zunächst nur die roten Zäpfchen gereizt werden, d.h. aktiv sind, und die blauen und grünen während dieser Zeit nicht aktiv sind [Dor04]. Wenn danach auf dem weißen Bild alle Zapfen durch die weiße Fläche gereizt werden, liefern die neu aktivierten blauen und grünen Zapfen stärkere Impulse im Vergleich zu den bereits längere Zeit aktiven und infolge

[69] Quelle: `http://www.bnr-art.com/doolitt/pintos.htm`

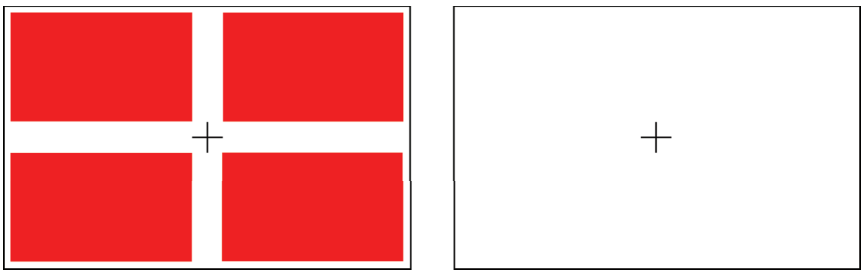


Abbildung A.4: Farbnachbild

gedämpften roten Zapfen.

Dass die Farbwahrnehmung und die physikalischen Werte nicht linear zueinander korrespondieren, zeigt die linke Abbildung A.5[70]. So haben die jeweiligen Farben in jeder Reihe den gleichen Abstand bzgl. des Farbtons (*engl.* hue) in HSV-Farbmodelldarstellung. Dennoch nehmen wir Menschen die Farbtöne in der oberen Reihe wesentlich kontrastreicher wahr als in der unteren Reihe. Ursache hierfür ist die Tatsache, dass im Gelbbereich zwei Farbzapfen, nämlich die roten und die grünen, annähernd ihren maximalen Absorptionsbereich haben, während die roten Farbtöne nur noch durch die roten Zapfen wahrgenommen werden können. Dies führt dazu, dass im Gelbbereich bereits kleinere Änderungen der Wellenlänge eine große Änderung der Aktivierungen nach sich ziehen und die menschliche Wahrnehmung in diesen Bereichen besonders sensibel ist [Gre98]. Die resultierende farbliche Auflösung in Abhängigkeit von der menschlichen Wahrnehmung zeigt sich in der rechten Abbildung in A.5. Im Gelbbereich ist die farbliche Auflösung besonders groß, ein anderes Minimum findet sich im Bereich, wo die grünen und roten Zäpfchen zwar stark reduzierte Sensibilität aufweisen, aber dafür auch noch die blauen Zapfen aktiv sind.

Während das erste Beispiel vorrangig die Form, das zweite und dritte die Farbe als Wahrnehmungsmerkmal inne hatte, hängt das letzte hauptsächlich mit dem Kontrast, also gewissermaßen der Unterschiedlichkeit von vorkommenden Intensitäten, zusammen [Ade00]. Die Abb. A.6[71] zeigt, wie sehr sich die Umgebung auf die Einschätzung bei Helligkeiten auswirken kann. „Sieht man die Felder A und B auf dem linken Bildabschnitt, erscheinen sie eindeutig unterschiedlich hell. Verbindet man sie jedoch mit einer

[70] Quelle: `http://de.wikipedia.org/wiki/Bild:Demo_farbabstand_hsb.gif`
[71] Quelle: `http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html`

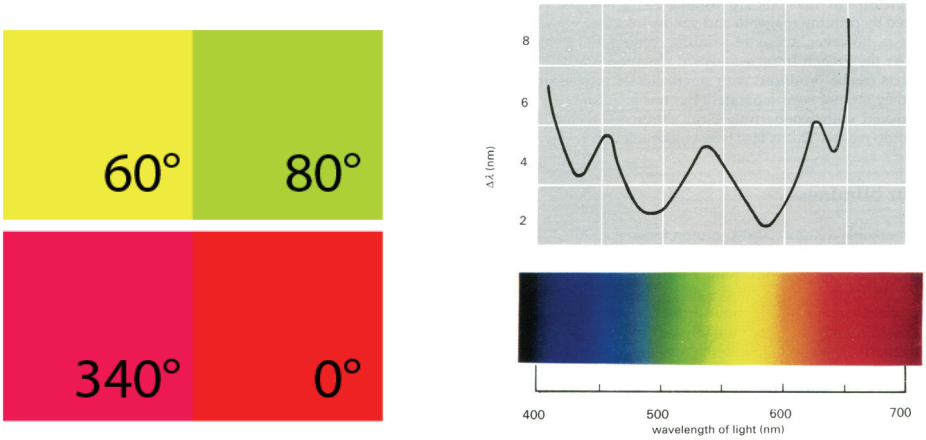


Abbildung A.5: Farbkontrast - Farbwahrnehmung, Farbtonauflösung [Gre98]

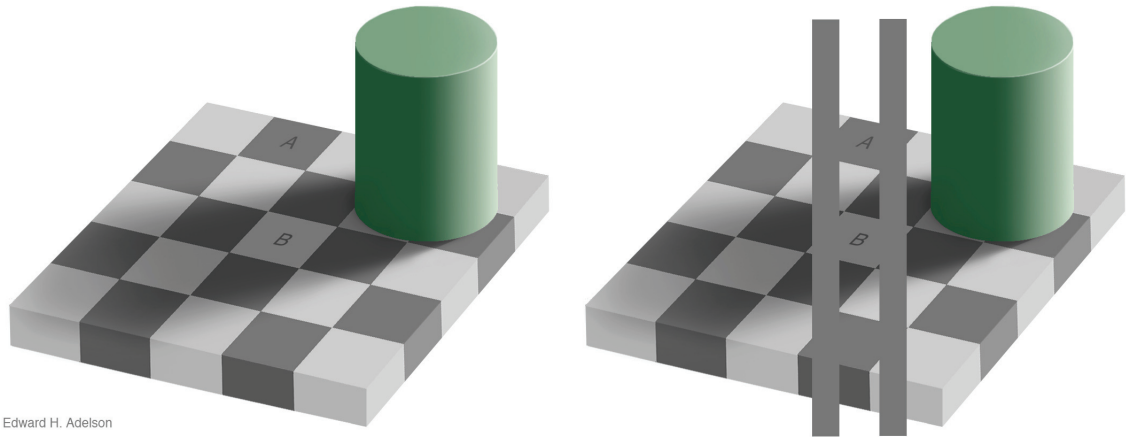


Abbildung A.6: Simultaner Kontrasteffekt

Farbbrücke wird klar, dass sie beide identische Helligkeitswerte besitzen. Neben dem lokalen Kontrasteffekt, der dazu führt, dass bei dunklen direkten Nachbarn das Feld relativ heller scheint, spielt hier noch ein anderer Effekt eine Rolle. Der Kontrasteindruck wird hier dadurch verstärkt, dass kontinuierliche Übergänge (hier im Verschattungsbereich) eher ignoriert werden.

A.3 Begriffe der Wahrscheinlichkeitsrechnung

In diesem Kapitel werden einige Theoreme der Wahrscheinlichkeitstheorie beschrieben, die für den Bereich dieser Arbeit von Interesse sind. Da die Wahrscheinlichkeitstheorie eng mit der Maßtheorie verbunden ist, werden einige Begriffe von dort Verwendung finden, ohne hier explizit definiert zu werden. Die entsprechenden Begriffe können beispielsweise in [WHK04] nachgeschlagen werden. Die verwendeten Definitionen bilden keine umfassende Einführung in das Thema, sondern werden in Bezug auf die Wichtigkeit für diese Arbeit ausgewählt. [WHK04], [Haf98], [Sac04].

Definition A.3.1 (Wahrscheinlichkeitsverteilung, Wahrscheinlichkeitsraum)
Sei Ω eine nichtleere Menge und $\mathcal{E}$ eine σ-Algebra über Ω. Die Abbildung $P : \mathcal{E} \to [0, 1]$ hat die folgenden Eigenschaften:

(1) $P(\Omega) = 1$

(2) $\{A_n : n\}$ eine endliche oder abzählbare Menge von paarweise disjunkten Ereignissen mit $P(\bigcup_n A_n) = \sum_n P(A_n)$

P heißt Wahrscheinlichkeitsverteilung über Ω. $(\Omega, \mathcal{E}, P)$ heißt Wahrscheinlichkeitsraum.

Definition A.3.2 (Zufallsvariable)
Sei $(\Omega, \mathcal{E}, P)$ ein Wahrscheinlichkeitsraum und $(\Omega', \mathcal{E}')$ ein Ereignisraum. Eine Abbildung $X : \Omega \to \Omega'$ heißt Zufallsvariable falls X $(\mathcal{E}, \mathcal{E}')$-meßbar ist: $X^{-1}(A') \in \mathcal{E} \ \forall \ A' \in \mathcal{E}'$

Der folgende Satz ermöglicht es, die Wahrscheinlichkeitsverteilung des Wahrscheinlichkeitsraums und den Ereignisraum, in den die Zufallsvariable abbildet, anzupassen:

Satz A.3.1 (Verteilung)
Sei $(\Omega, \mathcal{E}, P)$ ein Wahrscheinlichkeitsraum, X sei eine Zufallsvariable über Ω mit Werten im Ereignisraum $(\Omega', \mathcal{E}')$. $P_X(F) = P(X^{-1}(F)) \ \forall F \in \mathcal{E}'$ definiert eine Wahrscheinlichkeitsverteilung über Ω', nämlich die Verteilung der Zufallsvariablen X.

Im Falle von reellwertigen Zufallsvariablen kann die Verteilung wie folgt definiert werden:

Definition A.3.3 (Verteilungsfunktion)
Sei $(\Omega, \mathcal{E}, P)$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Die Funktion $F(x) = P(X^{-1}(]-\infty, x])), \ x \in \mathbb{R}$ heißt Verteilungsfunktion der Zufallsvariablen X.

Definition A.3.4 (Dichte einer Wahrscheinlichkeitsverteilung)
Sei F eine kontinuierliche und stückweise stetig differenzierbare Verteilungsfunktion. Die Ableitung $f(x) = \frac{dF(x)}{dx}$ heißt Dichte der Wahrscheinlichkeitsverteilung, die durch die Verteilungsfunktion bestimmt ist.

Definition A.3.5 (Erwartungswert und Varianz)
Habe die reellwertige Zufallsvariable X die Verteilungsfunktion $F(t) = \int_{-\infty}^{t} f(s)ds$, mit $f(s)$ einer kontinuierlichen Dichteverteilung und $\int_{-\infty}^{\infty} |s|f(s)ds < \infty$.

$$E(X) = \int_{-\infty}^{\infty} sf(s)ds$$

heißt Erwartungswert der Zufallsvariablen X.
Sei $\int_{-\infty}^{\infty} s^2 f(s)ds < \infty$

$$V(X) = \int_{-\infty}^{\infty} (s - E(X))^2 f(s)ds = E((X - E(X))^2) = E(X^2) - E(X)^2$$

heißt Varianz der Zufallsvariablen X. $\sigma(X) = \sqrt{V(X)}$ heißt Standardabweichung der Zufallsvariablen X.

Bemerkung A.3.1 (Stochastische Unabhängigkeit)
Zwei Ereignisse A und B heißen stochastisch unabhängig, wenn gilt:

$$P(A \cap B) = P(A) \cdot P(B)$$

Bemerkung A.3.2 (Bedingte Unabhängigkeit - Conditional Independence)
Seien A, B und C drei Ereignisse. [Wil07]
A heißt bedingt unabhängig von B, gegeben C, wenn gilt:

$$P(A|B, C) = P(A|C)$$

Bemerkung A.3.3 (Bayes Regel)
Die Bayessche Regel trifft eine Aussage über den Zusammenhang von Ursache und Effekt bei bedingten Wahrscheinlichkeiten. [Sac04] Seien A, B zwei Ereignisse, dann gilt:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Eine wichtige Verteilung ist die Normalverteilung bzw. Gausskurve. Viele in der Natur vorkommende Verteilungen oder auch Messreihen entsprechen einer Gaussverteilung. So wird beispielsweise die räumliche Verteilung eines eindimensionalen Elementarteilchens häufig als Gaussches Wellenpaket dargestellt. In Fällen, bei denen die Verteilung keiner einfachen Normalverteilung entspricht, kann man diese durch eine Linearkombination von Normalverteilungen annähern [Fin02].

Beispiel A.3.1 (Kontinuierliche Normalverteilung)
Sei $\Omega = \mathbb{R}$ und $\mathcal{E}$ ein Borel-Algebra über Ω. Die Normal-Dichteverteilung ist

$$p(x) = \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

wobei μ der Mittelwert und σ die Standardabweichung der Zufallsvariablen $X = (x)$ ist. Die Abkürzung ist $X \sim \mathcal{N}(\mu, \sigma^2)$. Es gilt

$$Var(X) = \sigma^2$$

$$E(X) = \mu$$

Beispiel A.3.2 (Kontinuierliche n-dim. Normalv. - stochastisch unabhängig)
Sei $\Omega = \mathbb{R}^n$ und $\mathcal{E}$ eine Borel-Algebra auf Ω. Die n-dimensionale Normal-Dichteverteilung ist

$$p(x_1, \cdots, x_n) = \prod_{j=1}^{n} \frac{1}{\sigma_j \sqrt{2\pi}} \exp^{-\frac{(x_j - m_j)^2}{2\sigma_j^2}}$$

wobei m_j die Mittelwerte und σ_j die Standardabweichungen der j-ten Komponente der Zufallsvariablen $X = (x_1, \cdots, x_n)$ ist.
Diese Verteilung entspricht einem Modell von n stochastisch unabhängigen normalverteilten eindimensionalen Zufallsvariablen.

Für die Anwendbarkeit sind einige Eigenschaften der Gaussverteilung von besonderem Interesse [Bos76], denn sie ist reproduktiv, d.h. die Summe von Gaussverteilungen ist wieder eine Gaussverteilung. Dies hat zur Folge, dass „ein Netzwerk, das nur stetige Variablen mit linearen Gaussverteilungen enthält, eine gemeinsame Verteilung hat, bei der es sich um eine multivariate Gaussverteilung über alle Variablen handelt“ [RN03].

Satz A.3.2 (Normalverteilung ist reproduktiv)

1. Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ und $a \neq 0$.
   Dann gilt für die Zufallsvariable $Y = aX + b$: $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

2. Seien X_1 und X_2 zwei normalverteilte, stochastisch unabhängige Zufallsvariablen: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
   Dann gilt für die Zufallsvariable $Y = X_1 + X_2$: $Y \sim \mathcal{N}(\mu_1 + \mu_1, \sigma_1^2 + \sigma_2^2)$

Im Falle von nicht stochastisch unabhängigen Zufallsvariablen bzw. Einträgen in einem multi-dimensionalen Zufallsvektor muss die Kovarianz als Maß für den wechselseitigen Zusammenhang zweier Zufallsvariablen Berücksichtigung finden [Sac04].

Beispiel A.3.3 (Kontinuierliche n-dim. Normalv. - stochastisch abhängig)
Im Falle eines n-dimensionalen Zufallvektors $X = (X_1, \ldots, X_n)$ ist die n-dimensionale Normalverteilung in ihrer allgemeinen Form gegeben durch

$$X \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

mit

$\boldsymbol{\mu}$	Mittelwert-Vektor
$\boldsymbol{\Sigma}$	Kovarianz Matrix, $\boldsymbol{\Sigma} = \begin{pmatrix} Kov(X_1, X_1) & \cdots & Kov(X_1, X_n) \\ \vdots & \ddots & \vdots \\ Kov(X_n, X_1) & \cdots & Kov(X_n, X_n) \end{pmatrix}$
Kov	Kovarianz, $Kov(X_i, X_j) = E\left((X_i - E(X_i))\,(X_j - E(X_j))\right)$
$\mathbf{x}$	Evidenz-Vektor
n	Dimension

Da die Kovarianz zweier Zufallsvariablen symmetrisch ist, ist auch die Kovarianzmatrix symmetrisch. Im Sonderfall, dass die Kovarianzen zwischen den einzelnen Einträgen Null sind, d.h. $Kov(X_i, X_j) = 0$ für $i \neq j$ wird die Matrix zu einer Diagonalmatrix mit den Varianzen auf der Diagonalen:

$$\boldsymbol{\Sigma} = \begin{pmatrix} Var(X_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & Var(X_n) \end{pmatrix}$$

Für die Umkehrmatrix der Kovarianzmatrix ergibt sich mit $Var(X_i) = \sigma^2(X_i)$:

$$\mathbf{\Sigma^{-1}} = \begin{pmatrix} \frac{1}{\sigma^2(X_1)} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma^2(X_n)} \end{pmatrix}$$

Damit vereinfacht sich die Verteilung mit $\sigma_i = \sigma(X_i)$ zu:

$$\begin{aligned} X \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) &= \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\prod_i \sigma_i^2}} \exp^{\sum_i -\frac{(\mathbf{x_i}-\boldsymbol{\mu_i})^2}{2\sigma_i^2}} \\ &= \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \prod_i \exp^{-\frac{(\mathbf{x_i}-\boldsymbol{\mu_i})^2}{2\sigma_i^2}} \\ &= \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp^{-\frac{(\mathbf{x_i}-\boldsymbol{\mu_i})^2}{2\sigma_i^2}} \end{aligned}$$

Setzt man die Unkorreliertheit der Zufallsvariablen voraus, vereinfacht sich die Kovarianzmatrix und damit die Verteilungsfunktion stark. Während die allgemeine Form der Normalverteilung einem beliebig orientierten Ellipsoid im Raum entspricht, ist dies im Spezialfall der Unkorreliertheit ein achsenorientiertes Ellipsoid. „Obwohl diese Vorgehensweise eigentlich inkorrekt ist, ergeben sich durch diese Heuristik häufig brauchbare Ergebnisse“ [Bur05].

Bemerkung A.3.4 (Berechnung Kovarianz)
Die Kovarianz zweier Zufallsvariablen X, Y lässt sich über das mittlere Abweichungsprodukt berechnen [Sac04]. Sei eine hinreichend große Menge von n Beobachtungswerten (x_i, y_i) gegeben, dann ergibt sich die Schätzung der Kovarianz zu

$$\sigma_{XY} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n-1} \quad .$$

A.4 Bayessche Netzwerke

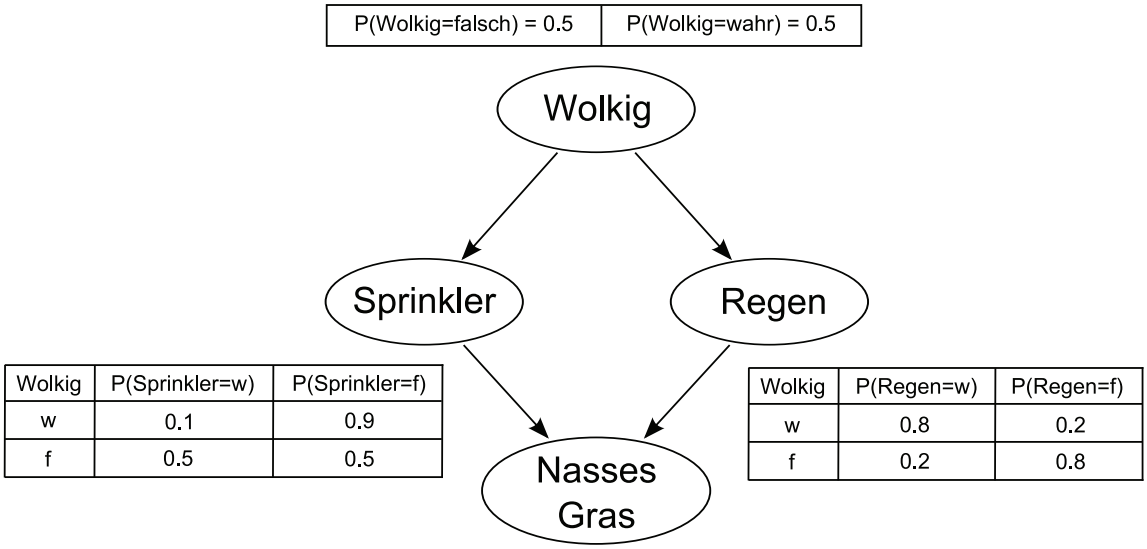
Grafische Modelle (*engl.* graphical models) verbinden die Wahrscheinlichkeitstheorie mit der Graphentheorie und damit Aspekte von Unsicherheit mit Aspekten von Komplexität [Mur01]. Man unterscheidet zwischen ungerichteten Netzwerken, sog. Markovnetzwerken, innerhalb derer auch Zyklen zulässig sind, und gerichteten Netzwerken, wie Bayesschen Netzen. Die Knoten innerhalb eines Graphen repräsentieren Zufallsvariablen, und das Nichtvorhandensein von Kanten repräsentiert bedingte Unabhängigkeiten. Aufgabe des Netzes ist es, das Schließen mit unsicherem Wissen zu ermöglichen und auch nachvollziehbar zu machen. Ein bekanntes System, das Bayessche Netze einsetzt, ist das medizinische Expertensystem „Pathfinder“ zur pathologischen Lymphknoten-Diagnose, das sehr erfolgreich 60 Krankheiten bzw. Diagnosen mit 100 Symptomen probabilistisch verknüpft [Hec91].

Ein Bayessches Netzwerk ist formal wie folgt definiert [CGH97]:

Definition A.4.1 (Bayessches Netzwerk)
Ein Bayessches Netzwerk ist ein Tupel (D, P), bestehend aus einem gerichteten azyklischen Graph D (*engl.* Directed Acyclic Graph, DAG) und einer Menge $P = \{p(x_1|\pi_1), \ldots, p(x_n|\pi_n)\}$, $|P| = n$, von bedingten Wahrscheinlichkeitsverteilungen (CPD, *engl.* Conditional Probability Distribution). Dabei ist Π_i die Menge der Elternknoten des Knotens X_i in D. Mit P ist auch die zugehörige Verbundwahrscheinlichkeitsverteilung JPD (*engl.* Joint Probability Distribution) gegeben mit $p(x) = \prod_{i=1}^{n} p(x_i|\pi_i)$.

Ein einfaches Beispiel, in dem die Zufallsvariablen diskreter, insbesondere binärer Natur sind, soll die Zusammenhänge verdeutlichen. Dafür wird ein in der Literatur bekanntes Beispiel aus [RN03] verwendet. Dieses wird auch in der Bibliothek PNL als Beispiel aufgeführt [Cor03].

Beispiel A.4.1 (Sprinkler/Regen-Modell)
Bei dem in Abb. A.7 dargestellten mehrfach verbundenen Netzwerk (Multiply Connected Network - MCN) gibt es vier diskrete Zufallsvariablen *Wolkig*, *Sprinkler*, *Regen* und *NassesGras* [RN03]. Da die Zufallsvariablen diskret sind, kann ihre bedingte Wahrscheinlichkeitsverteilung als bedingte Wahrscheinlichkeitstafel (CPT, engl. Conditional Probability Table) dargestellt werden. Die CPD der Zufallsvariable *Wolkig* entspricht, da



Wolkig	P(Sprinkler=w)	P(Sprinkler=f)
w	0.1	0.9
f	0.5	0.5

Wolkig	P(Regen=w)	P(Regen=f)
w	0.8	0.2
f	0.2	0.8

Sprinkler	Regen	P(NassesGras=w)	P(NassesGras=f)
w	w	0.99	0.01
w	f	0.9	0.1
f	w	0.9	0.1
f	f	0.0	1.0

Abbildung A.7: Einfaches Bayessches Netzwerk [RN03]

sie ohne Kinderknoten ist, einer Vorabwahrscheinlichkeit [Mur01]. Die Zustände der Zufallsvariablen *Sprinkler* und *Regen* hängen vom Wetter, d.h. der Zufallsvariablen *Wolkig* ab. Die Zufallsvariable *NassesGrass* hat zwei Elternknoten bzw. Zufallsvariablen nämlich *Sprinkler* und *Regen*.

Mit diesem Netzwerk lassen sich nun 3 Arten von Inferenz durch folgende Fragestellungen demonstrieren [Fin02], [Wah06]:

- **Diagnostische Inferenz / Diagnose (*engl.* diagnosis)**
  Hierbei wird von dem Symptom (Effekt, Auswirkung) auf die Diagnose (Ursache) geschlossen. Für das Gras-Beispiel in Abb. A.7 entspricht dies einer Schlussfolgerung der Art: *Das Gras ist nass, wurde seine Nässe durch den Sprinkler oder durch*

Regen verursacht?

$$
\begin{aligned}
\mathbf{0.7079} &= \dots \\
&= \frac{P(R = w, NG = w)}{P(NG = w)} \\
&= \mathbf{P}(\mathbf{Regen} = \mathbf{wahr} | \mathbf{NassesGras} = \mathbf{wahr}) \\
&> \mathbf{P}(\mathbf{Sprinkler} = \mathbf{wahr} | \mathbf{NassesGras} = \mathbf{wahr}) \\
&= \frac{P(S = w, NG = w)}{P(NG = w)} \qquad (\textit{Def. Bedingte Wahrscheinlichkeit}) \\
&= \frac{\sum_{W,R} P(S = w, NG = w, W, R)}{\sum_{S,W,R} P(NG = w, S, W, R)} \qquad \textit{(Marginalisierung)} \\
&= \Big(P(NG = t | S = w, R = w) P(S = w | W = w) P(R = w | W = w) P(W = w) \\
&\quad + P(NG = t | S = w, R = f) P(S = w | W = w) P(R = f | W = w) P(W = w) \\
&\quad + P(NG = t | S = w, R = w) P(S = w | W = f) P(R = w | W = f) P(W = f) \\
&\quad + P(NG = t | S = w, R = f) P(S = w | W = f) P(R = f | W = f) P(W = f) \Big) \\
&\quad \div \sum_{S,W,R} P(NG = w, S, W, R) \\
&= \frac{0.99 \cdot 0.1 \cdot 0.8 \cdot 0.5 + 0.9 \cdot 0.1 \cdot 0.2 \cdot 0.5}{\sum_{S,W,R} P(NG = w, S, W, R)} \\
&\quad + \frac{0.99 \cdot 0.5 \cdot 0.2 \cdot 0.5 + 0.9 \cdot 0.5 \cdot 0.8 \cdot 0.5}{\sum_{S,W,R} P(NG = w, S, W, R)} \\
&= \frac{0.0396 + 0.0090 + 0.0495 + 0.1800}{\sum_{S,W,R} P(NG = w, S, W, R)} \\
&= \frac{0.2781}{0.6471} \\
&= \mathbf{0.4298}
\end{aligned}
$$

Es ist also wahrscheinlicher, dass der Regen zu dem nassen Gras geführt hat.

- **Kausale Inferenz / Voraussage (*engl.* prediction)**
  Hierbei wird von der Diagnose (Ursache) auf das Symptom (Effekt) geschlossen. Für das Gras-Beispiel in Abb. A.7 entspricht dies einer Schlussfolgerung der Art: *Es ist wolkig, wie wahrscheinlich ist das Gras nass?*

$$
\begin{aligned}
& \mathbf{P}(\mathbf{NassesGras} = \mathbf{wahr} | \mathbf{Wolkig} = \mathbf{wahr}) \\
= \; & \frac{P(NassesGras = wahr, Wolkig = wahr)}{P(Wolkig = wahr)} \\
= \; & \frac{\sum_{S,R} P(NG = w, W = w, S, R)}{P(W = w)} \\
= \; & \Big(\; P(NG = t | S = w, R = w) P(S = w | W = w) P(R = w | W = w) P(W = w) \\
& + P(NG = t | S = w, R = f) P(S = w | W = w) P(R = f | W = w) P(W = w) \\
& + P(NG = t | S = f, R = w) P(S = f | W = w) P(R = w | W = w) P(W = w) \\
& + P(NG = t | S = f, R = f) P(S = f | W = w) P(R = f | W = w) P(W = w) \; \Big) \\
& \div P(W = w) \\
= \; & \frac{0.99 \cdot 0.1 \cdot 0.8 \cdot 0.5 + 0.9 \cdot 0.1 \cdot 0.2 \cdot 0.5}{P(W = w)} \\
& + \frac{0.9 \cdot 0.9 \cdot 0.8 \cdot 0.5 + 0.0 \cdot 0.9 \cdot 0.2 \cdot 0.5}{P(W = w)} \\
= \; & \frac{0.0396 + 0.0090 + 0.3240 + 0.0000}{P(W = w)} \\
= \; & \frac{0.3726}{0.5} \\
= \; & \mathbf{0.7452}
\end{aligned}
$$

- **Interkausale Inferenz / Wegerklären (*engl.* explaining away), Berksons Paradoxon**

  Hierbei wird zwischen verschiedenen Diagnosen (Ursachen) des gleichen Symptoms (Effekts) geschlossen. Für das Gras-Beispiel in Abb. A.7 entspricht dies einer Schlussfolgerung der Art: *Das Gras ist nass, und es hat geregnet, wie wahrscheinlich war der Sprinkler an?*

$$
\begin{aligned}
\mathbf{0.7009} \; &= \; \mathbf{P}(\mathbf{Sprinkler} = \mathbf{w} | \mathbf{NassesGras} = \mathbf{w}) \\
&> \; \mathbf{P}(\mathbf{Sprinkler} = \mathbf{w} | \mathbf{NassesGras} = \mathbf{w}, \mathbf{Regen} = \mathbf{w}) \\
&= \; \mathbf{0.1945}
\end{aligned}
$$

  Die Elternknoten, hier Sprinkler und Regen, sind also bedingt abhängig (*engl.* conditionally dependent). Das Eintreten des einen Ereignisses, hier des Regens, macht das andere, hier den Sprinkler, weniger wahrscheinlich.

Bemerkung A.4.1 (Knoten)
Die Knoten in einem Bayesschen Netzwerk sind Zufallsvariablen, deren Zustände wiederum Datenpunkten entsprechen. Entsprechend können die Werte *diskret* oder *kontinuierlich* und *univariat* oder *multivariat* sein. Da innerhalb des Netzes Übergänge zwischen den einzelnen Datentypen vorkommen können, gibt es eine entsprechende Menge von Knotentypen, die die jeweiligen Ein- und Ausgabetypen miteinander verbinden. Die Tabellen A.1 und A.2 zeigen die möglichen Knotentypen und deren Übergangsmöglichkeiten einmal für diskrete Kindknoten und einmal für kontinuierliche Kindknoten.

Dabei gilt [Kev02], [Cor04]:

- $\pi(j)$ ist die initiale Zustandswahrscheinlichkeit oder Vorabwahrscheinlichkeit der Zufallsvariablen X_1 mit $\pi(j) = P(X_1 = j)$.

- $A(i, j)$ beschreibt die Übergangswahrscheinlichkeit der Zustände der Zufallsvariablen X_{t-1} zu X_t mit $A(i, j) = P(X_t = j|X_{t-1})$. A ist eine sog. stochastische Matrix, in der sich die Zeilen zu Eins aufsummieren.

- $\sigma(u, W, j)$ ist die Softmax-Funktion, eine Verallgemeinerung der sigmoidalen Funktion - bei der es nur zwei Zustände gibt - auf m Zustände. Es gilt $P(Y = j|U = \mathbf{u}) = \sigma(\mathbf{u}, \mathbf{W}, j) = \frac{\exp(\mathbf{u}'\mathbf{W}_j)}{\sum_{k\in K}\exp(\sum \mathbf{u}'\mathbf{W}_k)}$, wobei $\mathbf{u}$ ein kontinuierlicher n-dimensionaler Vektor mit den Elternzuständen ist, $\mathbf{W}$ die $n \times m$ Gewichtsmatrix und $\mathbf{W}_k$ ist die k-te Spalte von $\mathbf{W}$. Im Falle von $\dim(\mathbf{u}) = 2$ ergibt sich, vorausgesetzt der diskrete Elternknoten sei im Zustand 1: $P(Y = 1|U = \mathbf{u}) = \frac{\exp(\mathbf{u}'\mathbf{W}_1)}{\exp(\mathbf{u}'\mathbf{W}_0)+\exp(\mathbf{u}'\mathbf{W}_1)}$ $= \frac{1}{\frac{\exp(\mathbf{u}'\mathbf{W}_0)}{\exp(\mathbf{u}'\mathbf{W}_1)}+1} = \frac{1}{1+\exp(\mathbf{u}'(\mathbf{W}_0-\mathbf{W}_1))}$. Dies entspricht einer sigmoidalen Funktion.

Diskrete Kindknoten			
Diskr. Eltern-knoten	*Kont. Eltern-knoten*	*Name der Verteilung*	*CPD*
-	-	Multinomial	$P(Y = j) = \pi(j)$
i	-	Bed. Multinomial	$P(Y = j\|X = i) = A(i, j)$
-	u	Softmax	$P(Y = j\|U = u) = \sigma(u, W, j)$
i	u	Bed. Softmax	$P(Y = j\|U = u, X = i) = \sigma(u, W_i, j)$

Y: diskreter Kindknoten, X: diskreter Elternknoten, U: kontinuierlicher Elternknoten

Tabelle A.1: CPDs für diskreten Kindknoten [Kev02]

Kontinuierliche Kindknoten			
Diskr. Eltern-knoten	*Kont. Eltern-knoten*	*Name der Verteilung*	*CPD*
-	-	Gauss	$P(Y = y) = \mathcal{N}(y; \mu, \Sigma)$
i	-	Bed. Gauss	$P(Y = y\|X = i) = \mathcal{N}(y; \mu_i, \Sigma_i)$
-	u	Linear Gauss	$P(Y = y\|U = u) = \mathcal{N}(y; Wu + \mu, \Sigma)$
i	u	Bed. linear Gauss	$P(Y = y\|U = u, X = i) = \mathcal{N}(y; W_i u + \mu_i, \Sigma_i)$

Y: kontinuierlicher Kindknoten, X: diskreter Elternknoten, U: kontinuierlicher Elternknoten

Tabelle A.2: CPDs für kontinuierlichen Kindknoten [Kev02]

A.5 Merkmalsparameter eines Stuhls

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<pbobject>

   <objecttype>chair</objecttype>

   <centobject>385</centobject>

   <property>
      <id>383</id>
      <subpart>sittingsurface</subpart>
      <proptype>deviation_vertical</proptype>
      <value>2.0991</value>
   </property>

   <property>
      <id>383</id>
      <subpart>sittingsurface</subpart>
      <proptype>size</proptype>
      <value>0.15546</value>
   </property>

   <property>
      <id>383</id>
      <subpart>sittingsurface</subpart>
      <proptype>center_height</proptype>
      <value>-0.670433</value>
   </property>

   <property>
      <id>385</id>
      <subpart>backrest</subpart>
      <proptype>deviation_horizontal</proptype>
      <value>10.6773</value>
   </property>

   <property>
      <id>385</id>
      <subpart>backrest</subpart>
      <proptype>size</proptype>
      <value>0.184231</value>
   </property>

   <property>
      <id>385</id>
      <subpart>backrest</subpart>
      <proptype>center_height</proptype>
      <value>-0.388433</value>
   </property>

   <relation>
      <id1>383</id1>
      <id2>385</id2>
      <reltype>distance_horizontal</reltype>
      <value>0.0293782</value>
   </relation>

</pbobject>
```

Listing A.1: Merkmalsparameter des rechten Stuhls aus Abb. 3.13

Literaturverzeichnis

[ABFJB03] Aubert, Gilles ; Barlaud, Michel ; Faugeras, Olivier D. ; Jehan-Besson, Stéphanie: Image Segmentation Using Active Contours: Calculus of Variations or Shape Gradients? In: *SIAM Journal of Applied Mathematics* 63 (2003), Nr. 6, S. 2128–2154

[Ade00] Adelson, Edward H.: Lightness Perception and Lightness Illusions. In: Gazzaniga, M. S. (Hrsg.): *The New Cognitive Neurosciences*. 2nd edition. MIT Press, 2000

[BB95] Bogoni, L. ; Bajcsy, R.: Interactive Recognition and Representation of Functionality. In: *Computer Vision and Image Understanding* 62 (1995), September, Nr. 2, S. 194–214

[BET95] Buelthoff, Heinrich H. ; Edelman, Shimon Y. ; Tarr, Michael J.: How Are Three-Dimensional Objects Represented in the brain? In: *Cerebral Cortex* 5 (1995), Nr. 3, S. 247–260

[Bie87] Biederman, I.: Recognition-by-Components: A Theory of Human Image Understanding. In: *Psycholog. Rev.* 94 (1987), Nr. 2, S. 115–147

[BJ85] Besl, Paul J. ; Jain, Ramesh C.: Three-Dimensional Object Recognition. In: *ACM Computing Surveys* 17 (1985), März, Nr. 1, S. 75–145

[BK99] Bauer, Eric ; Kohavi, Ron: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. In: *Machine Learning* 36 (1999), Nr. 1/2, S. 105–139

[BL88] Berners-Lee, Tim: *Semantic Web Roadmap*. Version: 1988. `http://www.w3.org/DesignIssues/Semantic.html`, Abruf: 09.12.2007

[Bos76] Bosch, Karl: *Elementare Einführung in die Wahrscheinlichkeitsrechnung*. 1. Reinbeck : Rowohlt, 1976. – ISBN 3–499–27025–0

[BR83] Bickhard, Mark H. ; Richie, D. M.: *On the Nature of Representation: A Case Study of James Gibson's Theory of Perception*. Praeger Publishers, 1983

[Bre07] Breden, Nils: *Formbasierte Objekterkennung im 3-D Raum*, Universität Bremen, Diplomarbeit, 2007

[BS89] Bronstein, I. ; Semendjajew, K.: *Taschenbuch der Mathematik*. Teubner Verlagsgesellschaft, Leipzig, 1989

[BSB02] Bahlmann, C. ; Schael, M. ; Burkhardt, H.: *Aktives Sehen (Skriptum 2002)*. Albert-Ludwigs-Universität Freiburg, Institut für Informatik Lehrstuhl für Mustererkennung und Bildverarbeitung, 2002

[Bur05] Burkhardt, H.: *Mustererkennung (Vorlesungsskriptum WS 2005/2006)*. Lehrstuhl für Mustererkennung und Bildverarbeitung, Albert-Ludwigs-Universität Freiburg, 2005

[CGA07] CGAL Editorial Board: *CGAL User and Reference Manual*. 3.3, 2007

[CGH97] Castillo, Enrique ; Gutierrez, Jose M. ; Hadi, Hadi: *Expert Systems and Probabilistic Network Models*. 1. New York : Springer, 1997. – ISBN 0–387–94858–9

[CMO04] C. Masolo, A. Gangemi N. G. S. Borgo B. S. Borgo ; Oltramari, A.: Ontology Library (final). WonderWeb Deliverable D18 / Laboratory For Applied Ontology - ISTC-CNR, Trento, Italien. 2004. – Forschungsbericht

[CN06] *Kapitel* Introductory Remarks. In: Christensen, H. I. ; Nagel, H.-H.: *Sampling th Spectrum of Approaches*. Springer, 2006, S. 1–4

[Cor03] Corporation, Intel: Probabilistic Network Library / Intel. 2002-2003. – Forschungsbericht

[Cor04] Corporation, Intel: SoftMax distribution / Intel. 2004. – Forschungsbericht

[CTSO03] Chen, Ding-Yun ; Tian, Xiao-Pei ; Shen, Yu-Te ; Ouhyoung, Ming: On Visual Similarity Based 3D Model Retrieval. In: *Comput. Graph. Forum* 22 (2003), Nr. 3, S. 223–232

[Dah06] Dahm, Markus: *Grundlagen der Mensch-Computer-Interaktion*. Pearson Studium, 2006

[DF01] Dalley, G. ; Flynn, P. J.: Range image registration: A software platform and empirical evaluation. In: *3DIM01*, 2001, 246–253

[Dor04] Dorner, Verlag E.: *Linder 2, LP 2004 – Lösungen*. Version: 2004. `www.dorner-verlag.at/dateien/ed-mat/3-7055-0420-7_Loe2.pdf`, Abruf: 22.05.2007

[Eck02] Eckelt, P.: *Quantentheorie (Vorlesungsskriptum 2001/2002)*. Westfälische Wilhelms-Universität Münster, 2002

[ECV03] *ECVision: European Research Network for Cognitive Computer Vision*. Version: 2003. `http://www.ecvision.org`

[Ede92] *Kapitel* Visual perception. In: Edelman, Shimon: *Encyclopedia of Artificial Intelligence*. Bd. 2. S. Shapiro, ed., Wiley, New York, 1992, S. 1655–1663

[Ede99] Edelman, Shimon: *Representation and Recognition in Vision*. Massachu-

setts Institute of Technology, 1999

[Eng06] *Kapitel* Neuronale Grundlagen der Merkmalsintegration. In: Engel, Andreas K.: *Neuropsychologie*. Springer Heidelberg, 2006, S. 55–65

[Fin02] Finck, Alex: Seminar über Bayes'sche Netzwerke, Bayesian Inference / Freie Universität Berlin. 2002 (9.12.2004). – Forschungsbericht

[Fin03] Fink, Gernot A.: *Mustererkennung mit Markov-Modellen - Theorie - Praxis - Anwendungsgebiete*. B.G. Teubner Stuttgart - Leipzig - Wiesbaden, 2003

[Fin07] Finken, Julia: *Farbbasierte Objekterkennung unter variablen Beleuchtungsverhältnissen*, Universität Bremen, Diplomarbeit, 2007

[FKK+05] Fay, Rebecca ; Kaufmann, Ulrich ; Knoblauch, Andreas ; Markert, Heiner ; Palm, Günther: Combining Visual Attention, Object Recognition and Associative Information Processing in a NeuroBotic System. In: Wermter, Stefan (Hrsg.) ; Palm, Günther (Hrsg.) ; Elshaw, Mark (Hrsg.): *Biomimetic Neural Learning for Intelligent Robots. Intelligent Systems, Cognitive Robotics, and Neuroscience.* Bd. 3575. Berlin, Heidelberg : Springer, 2005, S. 118–143

[FMK+03] Funkhouser, Thomas A. ; Min, Patrick ; Kazhdan, Michael M. ; Chen, Joyce ; Halderman, Alex ; Dobkin, David P. ; Jacobs, David P.: A search engine for 3D models. In: *ACM Transactions on Graphics* 22 (2003), Januar, Nr. 1, S. 83–105

[Fre95] Freeman, W.: *Orientation histogram for hand gesture recognition.* `citeseer.ist.psu.edu/freeman94orientation.html`. Version: 1995

[FS99] Freund, Y. ; Schapire, R.: A short introduction to boosting. In: *Journal of Japanese Society for Artificial Intelligence* 14 (1999), Nr. 5, S. 771–780

[FSC98] Finlayson, Graham D. ; Schiele, Bernt ; Crowley, James L.: Comprehensive Colour Image Normalization. In: Burkhardt, Hans (Hrsg.) ; Neumann, Bernd (Hrsg.): *ECCV (1)* Bd. 1406, Springer, 1998 (Lecture Notes in Computer Science). – ISBN 3–540–64569–1, 475–490

[GG00] Gérard, Philippe ; Gagalowicz, André: Three dimensional model-based tracking using texture learning and matching. In: *Pattern Recognition Letters* 21 (2000), Nr. 13-14, S. 1095–1103. – ISSN 0167–8655

[GHKW05] Grünewald, Uli ; Hostert, Alexandra ; Kneser, Jakob ; Wieghaus, Georg: Wunder Wahrnehmung - von Sinnestäuschungen und Hirngespinsten, / WDR, Quarks & Co. 2005. – Forschungsbericht

[Gib50] Gibson, J. J.: *The Perception of the Visual World*. Boston, Massachusetts : Houghton Mifflin, 1950

[Gib82] Gibson, James J.: *Die Sinne und der Prozeß der Wahrnehmung*. Verlag

Hans Huber, 1982

[Gol02] Goldstein, E. B. ; Ritter, Manfred (Hrsg.): *Wahrnehmungspsychologie*. Spektrum Akademischer Verlag, 2002

[Gon02] Gonnet, Gaston: *Scientific Computation (Vorlesungsskriptum 2001/2002)*. Institute for Scientific Computing, ETH Zürich, 2002

[Gra69] Grawert, Gerald ; Wolter, Hans (Hrsg.): *Quantenmechanik 1*. Akademische Verlagsgesellschaft, Frankfurt am Main, 1969

[Gre98] Gregory, Richard L.: *Eye and Brain, The Psychology of Seeing*. Oxford : Oxford University Press, 1998

[GW92] Gonzalez, Rafael C. ; Woods, Richard E.: *Digital Image Processing*. 3rd. Addison-Wesley, 1992

[GZ05] Gerrig, Richard J. ; Zimbardo, Philip G.: *Pychology and Life*. Allyn & Bacon, 2005

[Haf98] Hafner, Robert: *Wahrscheinlichkeitsrechnung und Statistik*. Springer, 1998

[Hec91] Heckerman, David: *Probabilistic Similarity Networks*. MIT Press, 1991

[Her94] Hernández, Daniel: *Lecture Notes in Artificial Intelligence*. Bd. 804: *Qualitative Representation of Spatial Knowledge*. Berlin : Springer, 1994. – Revised version of a 1992 Ph.D. Thesis at the Institut für Informatik, Technische Universität München

[HGB95] Hoover, Adam ; Goldgof, Dmitry ; Bowyer, Kevin W.: Extracting a Valid Boundary Representation from a Segmented Range Image. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 17 (1995), Nr. 9, S. 920–924. – ISSN 0162–8828

[HJBJ+96] Hoover, Adam ; Jean-Baptiste, Gillian ; Jiang, Xiaoyi ; Flynn, Patrick J. ; Bunke, Horst ; Goldgof, Dmitry B. ; Bowyer, Kevin K. ; Eggert, David W. ; Fitzgibbon, Andrew W. ; Fisher, Robert B.: An Experimental Comparison of Range Image Segmentation Algorithms. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18 (1996), Nr. 7, S. 673–689

[HSB06] Hois, Joana ; Schill, Kerstin ; Bateman, John A.: Integrating Uncertain Knowledge in a Domain Ontology for Room Concept Classifications. In: Bramer, Max (Hrsg.) ; Coenen, Frans (Hrsg.) ; Tuson, Andrew (Hrsg.): *The Twenty-sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Cambridge, UK : Springer-Verlag, Dezember 2006 (Research and Development in Intelligent Systems XXIII), S. 245–258

[Hüb06] Hübner, Kai: *Symmetriesignaturen für bildbasierte Anwendungen in der Robotik*, Universität Bremen, Diss., 2006

[HWBR06] Hois, Joana ; Wünstel, Michael ; Bateman, John A. ; Röfer, Thomas: Dialog-Based 3D-Image Recognition Using a Domain Ontology. In: Barkowsky, Thomas (Hrsg.) ; Knauff, Markus (Hrsg.) ; Ligozat, Gérard (Hrsg.) ; Montello, Daniel R. (Hrsg.): *Spatial Cognition* Bd. 4387, Springer, 2006 (Lecture Notes in Computer Science). – ISBN 978–3–540–75665–1, 107–126

[HZ04] Hartley, R. I. ; Zisserman, A.: *Multiple View Geometry in Computer Vision*. 2nd. Cambridge University Press, 2004

[HZ06] Hübner, K. ; Zhang, J.: Stable Symmetry Feature Detection and Classification in Panoramic Robot Vision Systems. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, S. 3429 – 3434

[Jäh02] Jähne, Bernd: *Digitale Bildverarbeitung*. Springer, 2002

[Jan96] Jantschek, Thorsten: Wittgenstein über sehen und sehen-als. In: *Wittgenstein Studien 3(2).* (1996)

[Kan07] Kanthak, Christian: *Texturierte 3D-Rekonstruktion*, Universität Bremen, Diplomarbeit, 2007

[Kev02] Kevin, Murphy: *Dynamic Bayesian Network : Representation, Inference and Learning*, UC Berkeley, Diss., 2002

[Kle05] Klein, Rolf: *Algorithmische Geometrie*. Springer, 2005

[KPT+05] Kraats, E.B. van d. ; Penney, G.P. ; Tomazevic, D. ; Walsum, Th. van ; Niessen, W.J.: Standardized Evaluation Methodology for 2D-3D Registration. In: *IEEE Transactions on Medical Imaging* 24 (2005), Nr. 9, S. 1177–1190

[KWT87] Kass, M. ; Witkin, A. ; Terzopoulos, D.: Snakes: Active contour models. In: *International Journal of Computer Vision* 1 (1987), Nr. 4, S. 321–331

[LB98] Liter, J.C. ; Bülthoff, H.H.: An introduction to object recognition. In: *Zeitschrift für Naturforschung* (1998), Nr. 53c, S. 610–621

[Lev03] Levinson, Stephen C.: *Space in Language and Cognition*. Camebridge University Press, 2003

[Low87] Lowe, David G.: Three-Dimensional Object Recognition from Single Two-Dimensional Images. In: *Artificial Intelligence* 31 (1987), Nr. 3, S. 355–395

[Low99] Lowe, D. G.: Object Recognition from Local Scale-Invariant Features. In: *International Conference on Computer Vision*, 1999, 1150–1157

[Mal06] Mallot, Hanspeter A.: *Kognitive Neurobiologie (Grundlagen der Neuro- und Verhaltensbiologie III)*. Version: 2006. `http://www.uni-tuebingen.`

de/cog/teaching/ws2005_06/ham/vorl_cog/Skript.pdf, Abruf: 23.07.2006

[Mar82] Marr, D.: *Vision*. San Francisco : W. H. Freeman & Co., 1982

[Mey05] Meyer, Andreas: *Merkmals- und formbasierte 3D-Objekterkennung für Büroszenen*, Universität Bremen, Diplomarbeit, 2005

[MF07] Mandel, C. ; Frese, U.: Comparison of Wheelchair User Interfaces for the Paralysed: Head-Joystick vs. Verbal Path Selection from an offered Route-Set. In: *3rd European Conference on Mobile Robots (ECMR)*, 2007

[MHPM04] Michel, Jochen ; Hong, Vinh ; Paulus, Dietrich ; Münzenmayer, Christian: Farbkalibrierung mittels linearer Transformation. In: Droege, Detlev (Hrsg.) ; Paulus, Dietrich (Hrsg.): *Workshop Farbbildverarbeitung*, Der Andere Verlag, Tönning, 2004. – ISBN 3–89959–245–X, S. 83–90

[Min85] Minsky, Marvin: *The Society of Mind*. Simon and Schuster, 1985

[Min04] Min, Patrick: *A 3D Model Search Engine*, Princeton University, Diss., 2004

[Min06] Minsky, Marvin: *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster, 2006

[MK05] Muckli, Lars ; Kohler, Axel: Interpret und kreativer Lückenfüller. Wie optische Illusionen in der Großhirnrinde entstehen. In: *Forschung Frankfurt* 4 (2005), S. 14–20

[MMK+06] Meseth, J. ; Müller, G. ; Klein, R. ; Röder, F. ; Arnold, M.: Verification of Rendering Quality from Measured BTFs. In: *The 3rd Symposium on Applied Perception in Graphics and Visualization*, 2006

[MSB05] Mozos, Óscar M. ; Stachniss, Cyrill ; Burgard, Wolfram: Supervised Learning of Places from Range Data using AdaBoost. In: *IEEE International Conference on Robotics and Automation (ICRA)*, 2005

[MST00] Meer, P. ; Stewart, C. V. ; Tyler, D. E.: Robust Computer Vision: An Interdisciplinary Challenge. In: *Computer Vision and Image Understanding* 78 (2000), April, Nr. 1, S. 1–7

[MT06] Moratz, Reinhard ; Tenbrink, Thora: Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. In: *Spatial Cognition and Computation* 6 (2006), Nr. 1, S. 63–107

[Mur01] Murphy, Kevin P.: An introduction to graphical models, 2001

[MV87] Maturana, H. ; Varela, F.: *The Tree of Knowledge: A new look at the biological roots of human understanding*. Shambhala/New Science Library, Boston, 1987

[Neu08] Neugebauer, Peter: 3D-Digitalisierung realer Objekte / Polygon Technology GmbH. 2008. – Forschungsbericht

[NK03] Novotni, M. ; Klein, R.: 3D Zernike Descriptors for Content Based Shape Retrieval. In: *The 8th ACM Symposium on Solid Modeling and Applications*, 2003

[NK04] Novotni, M. ; Klein, R.: Shape Retrieval using 3D Zernike Descriptors. In: *Computer Aided Design* 36 (2004), Nr. 11, S. 1047–1062

[NS76] Newell, A. ; Simon, H.: Computer Science as Empirical Inquiry: Symbols and Search. In: *Communications of the ACM* 19 (1976), S. 113–126. – Reprinted in Haugeland, J., ed, *Mind Design II*, MIT Press, 1997

[OH06] Ohbuchi, Ryuatrou ; Hata, Yushin: Combining Multiresolution Shape Descriptors for 3D Model Retrieval. In: *WSCG'2006 Full Papers Proceedings*, 2006

[Pau01] Paulus, Dietrich: *Aktives Bildverstehen*. 20001. – Universität Erlangen–Nürnberg, Lehrstuhl für Mustererkennung

[Pau99] Paul, Harry: *Lexikon der Optik*. Spektrum Akademischer Verlag, 1999

[PCN98] Paulus, D. ; Csink, L. ; Niemann, H.: Color cluster rotation. In: *International Conference on Image Processing*, 1998, I: 161–165

[PE90] Poggio, T. ; Edelman, S.: A Network that Learns to Recognize 3D Objects. In: *Nature* 343 (1990), Nr. 6255, S. 263–266

[Pop94] Pope, Arthur R.: Model-Based Object Recognition - A Survey of Recent Research / Department of Computer Science, University of British Columbia. 1994 (TR-94-04). – Forschungsbericht

[Pot82] Potmesil, Michael: *Generating three-dimensional surface models of solid objects from multiple projections*, Rensselaer Polytechnic Institute (RPI), Troy, NY, Diss., 1982

[PTVF92] Press, William H. ; Teukolsky, Saul A. ; Vetterling, William T. ; Flannery, Brian P.: *Numerical Recipes in C*. Second Edition. Cambridge University Press, 1992

[Pyl84] Pylyshyn, Zenon W.: *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press, 1984. – ISBN 16098–6

[RDR95] Rivlin, Ehud ; Dickinson, Sven J. ; Rosenfeld, Azriel: Recognition by Functional Parts. In: *Computer Vision and Image Understanding: CVIU* 62 (1995), Nr. 2, S. 164–176

[Reg96] Regier, Terry: *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Cambridge, Massachusetts : The MIT Press,

1996. – ISBN 0–262–18173–8

[RGMJ02] Roy, Deb ; Gorniak, Peter ; Mukherjee, Niloy ; Juster, Josh: A Trainable Spoken Language Understanding System For Visual Object Selection. In: *International Conference of Spoken Language Processing*, 2002

[RL78] Rosch, E. (Hrsg.) ; Lloyd, B. (Hrsg.): *Cognition and Categorization*. Lawrence Erlbaum Assoc., 1978

[RL01] Rusinkiewicz, S. ; Levoy, M.: Efficient variants of the ICP algorithm. In: *3DIM01*, 2001, 145–152

[RN03] Russell, S. ; Norvig, P.: *Artificial Intelligence, A Modern Approach*. Prentice Hall, 2003 (AI)

[SA00] Stamos, I. ; Allen, P.: 3-D Model Construction using Range and Image Data. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-00)*. Los Alamitos : IEEE, Juni 13–15 2000, S. 531–536

[Sac04] Sachs, Lothar: *Angewandte Statistik*. Springer-Verlag, Berlin, 2004

[SB91] Stark, Louise ; Bowyer, Kevin: Achieving Generalized Object Recognition through Reasoning about Association of Function to Structure. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 13 (1991), Nr. 10, S. 1097–1104. – ISSN 0162–8828

[SB96] Stark, Louise ; Bowyer, Kevin: *Generic Object Recognition using Form and Function*. World Scientific, 1996. – ISBN 981–02–1508–8

[Sch98] Schütz, Christian: *Geometric point matching of free-form 3D objects*, Université de Neuchâtel, Diss., 1998

[Sch04] *Kapitel* Objekterkennung und Signaldetektion. In: Schwaninger, A.: *Praxisfelder der Wahrnehmungspsychologie*. B. Kersten and M.T. Groner (Hrsg.), Huber Verlag, Bern, CH, 2004, S. 106–130

[Sch06] Schultheis, Holger: Marvin Minsky on fundamental aspects of the field - The past, the present, and the future of Artificial Intelligence. In: *Künstliche Intelligenz* (2006), Nr. 4, S. 55–59

[SHGB93] Stark, L. ; Hoover, A.W. ; Goldgof, D.B. ; Bowyer, K.W.: Function-based recognition from incomplete knowledge of shape. In: *Proceedings of the IEEE Workshop on Qualitative Vision*, 1993, S. 11–22

[Sin03] Singh, Push: Examining the Society of Mind. In: *Computing and Informatics* 22 (2003), Nr. 6, S. 521–543

[Sou07] Sourceforge.Net: *Open Source Computer Vision Library*. Version: 2007. `http://sourceforge.net/projects/opencvlibrary/`, Abruf: 09.12.2007

[Str93] Stricker, Markus A.: Promising directions in active vision. In: *Int. J. Comput. Vision* 11 (1993), Nr. 2, S. 109–126. – ISSN 0920–5691

[Tan97] Tanaka, Keiji: Mechanisms of visual object recognition: monkey and human studies. In: *Current Opinion in Neurobiology* (1997), Nr. 7, S. 523–529

[The05] Thenbrink, Thora: Semantics and Application of Spatial Diensional Terms in English and German / Universität Bremen. 2005. – SFB/TR 8 Report No. 004-03/2005

[TV04] Tangelder, Johan W. H. ; Veltkamp, Remco C.: Survey of Content Based 3D Shape Retrieval Methods. In: *Shape Modeling International '04*, 2004, S. 145–156

[UB91] Ullman, Shimon ; Basri, Ronen: Recognition by linear combinations of models. In: *IEEE PAMI, October 1991* 13 (1991), Nr. 10, S. 992–1006

[V+05] Vernon, David u. a.: ECVision: A Research Roadmap of Cognitive Vision. (2005)

[VA05] Vladimier, Vezhnevets ; Alexander, Velizhev: GML Camera Calibration Toolbox. In: *Graphics ans Media Lab, Moscow State University* (2005)

[Ver05] Vernon, David: Cognitive Vision - The Development of a Discipline. In: *KI* 19 (2005), Nr. 2, S. 43–46

[Ver06a] Vernon, David: Cognitive Vision – The Development of a Discipline. (2006)

[Ver06b] *Kapitel* The Space of Cognitive Vision. In: Vernon, David: *Sampling th Spectrum of Approaches*. Springer, 2006, S. 7–24

[VJ01] Viola, Paul ; Jones, Michael J.: Robust Real-time Object Detection. (2001)

[Wah02] Wahlster, Wolfgang: Disambiguierung durch Wissensfusion: Grundprinzipien der Sprachtechnologie. In: *KI* 16 (2002), Nr. 1, S. 48–50

[Wah06] Wahlster, Wolfgang: *Einführung in die Methoden der Künstlichen Intelligenz (Vorlesungsskriptum WS 2006/07)*. Universität des Saarlandes, Saarbrücken, 2006

[WB00] Wallis, G. ; Bulthoff, H.: Learning to recognize objects / Max-Planck-Institut für biologische Kybernetik. 2000 (084). – Forschungsbericht

[WCH+95] Woods, Kevin ; Cook, Diane ; Hall, Lawrence ; Bowyer, Kevin W. ; Stark, Louise: Learning Membership Functions in a Function-Based Object Recognition System. In: *Journal of Artificial Intelligence Research* 3 (1995), S. 187–222

[WHK04] Wolff, M. P. H. ; Hauck, P. ; Küchlin, W.: *Mathematik für Informatik und Bioinformatik*. Springer-Verlag Berlin Heidelberg, 2004

[Wil07] Williams, Chris: *Probabilistic Modelling and Reasoning (Vorlesungsskriptum 2006/2007)*. Institute for Adaptive and Neural Computation, University of Edinburgh, 2007

[WR06] Wünstel, Michael ; Röfer, Thomas: A Probabilistic Approach for Object Recognition in a Real 3-D Office Environment. In: Skala, Vaclav (Hrsg.): *WSCG'2006 Posters Proceedings*, 2006, S. 41 – 42

[Zha98] Zhang, Zhengyou: A Flexible New Technique for Camera Calibration / Microsoft Research. 1998. – Technical Report MSR-TR-98-71

[ZL02] Zhang, Dengsheng ; Lu, Guojun: An Integrated Approach to Shape Based Image Retrieval. In: *ACCV2002* (2002)