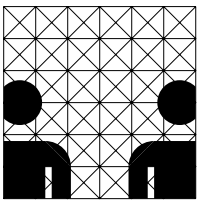
Entwurf und Evaluierung von Strategien zur 2D/3D-Objekterkennung in aktiven Sehsystemen

Dissertation
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

vorgelegt
am Fachbereich Informatik
der Universität Hamburg



von
Steffen Schmalz

Hamburg, April 2000

Gutachtende: Prof. Dr.-Ing. Bärbel Mertsching

Prof. Dr. Leonie Dreschler-Fischer

Tag der Disputation: 06.04.2001

Inhaltsverzeichnis

1 Einleitung **1**

2 Aktive Sehsysteme **3**
2.1 Philosophie 3
2.2 Aktuelle Forschungsansätze 5
2.3 Einsatz von aktiven Sehsystemen 13

3 Objekterkennungsansätze **15**
3.1 Grundlagen 15
3.1.1 Repräsentationsformen 15
3.1.2 Verarbeitungszweige 17
3.2 Einordnung 19
3.3 Geometrische Verfahren 21
3.3.1 Formenanalyse 21
3.3.2 Suchgraphen 30
3.3.3 Hashverfahren 38
3.3.4 Strukturbeschreibungen 40
3.4 Optimierungsverfahren 53
3.4.1 Hopfield-Netze 53
3.4.2 *Annealing*-Verfahren 58
3.4.3 Relaxationsverfahren 60
3.5 Biologienahe Systeme 65
3.5.1 Biologisch motivierte Systeme 66
3.5.2 Konnektionistische Systeme 69
3.6 Verfahren in aktiven Sehumgebungen 73
3.7 Wertung der Systeme 85

4 Aktives Sehen im NAVIS-Projekt **93**
4.1 Experimentierumgebungen 94
4.2 Merkmalsextraktion 94
4.3 2D-Objekterkennung 97

4.4 Stereobildverarbeitung 106

5 Erkennung mit Modellskalierungen 115
5.1 Erkennungssystem 115
5.2 Experimente und Ergebnisse 118

6 Erkennung mit zeitlichen Relationen 123
6.1 Repräsentationsschema 123
6.2 Erkennungssystem 124
6.2.1 Merkmalsbasis 125
6.2.2 Ansichtenklassifikation 126
6.2.3 Aufzeichnung zeitlicher Relationen 129
6.2.4 Objektklassifikation 130
6.3 Experimente und Ergebnisse 132

7 Erkennung mit Strukturbeschreibungen 137
7.1 Repräsentationsschema 138
7.1.1 Merkmalsbasis 140
7.1.2 selbstorganisierende Repräsentation 143
7.2 Matchstrategien 151
7.2.1 Hash-Verfahren 153
7.2.2 Relaxations-Verfahren 155
7.2.3 Graphensuch-Verfahren 157
7.2.4 Deformations-Verfahren 159
7.3 Experimente und Ergebnisse 163
7.3.1 Modellnetzgenerierung 164
7.3.2 Erkennung 167

8 Zusammenfassung und Ausblick 175

Literatur 179

Eidesstattliche Erklärung 199

Anhang 201

A Objektansichten 201

“Objekterkennung ... ist ein weites Feld”
frei nach v. Briest aus Fontane’s Effi Briest
[FONTANE 1896]

Kapitel 1

Einleitung

Die Objekterkennung als eines der anspruchsvollsten Probleme der rechnergestützten Bildverarbeitung ist seit dem Aufkommen des Computers als Rechenhilfsmittel eine Herausforderung für Ingenieure, Informatiker, Psychologen u.a. Die enorme Leistungsfähigkeit des menschlichen Sehsystems und die scheinbare Leichtigkeit, mit der wir in der Lage sind, auch komplexe Szenenkonfigurationen schnell und sicher zu analysieren, war dabei immer das zentrale biologische Vorbild.

Zwischenzeitlich lassen sich zwei prinzipiellen Herangehensweisen in der Forschung zum “Computersehen” unterscheiden: passives bzw. aktives Sehen. Ersteres hatte zum Ziel, möglichst viele Informationen über eine Szene aus einem Bild auszulesen. Die historisch jüngere Forschungsrichtung des aktiven Sehens versucht, motiviert durch die Arbeitsweise natürlicher Sehsysteme, jeweils nur die für eine aktuelle Aufgabe nötigen Informationen zu sammeln und bei Bedarf weitere Szenendaten zu generieren. Hier entsteht ein neues Systemverhalten durch die verkoppelte Arbeitsweise verschiedener verhaltensorientierter Module.

Ein Kernmodul jedes Sehsystems, ob nun passiv oder aktiv, ist die Objekterkennung. Aktive Sehsysteme bieten aber Möglichkeiten, um das äußerst schwierige Problem des Zuordnens von symbolischen Bezeichnern zu Pixelregionen zu vereinfachen. Durch Aufmerksamkeitsmechanismen kann sichergestellt werden, daß auch wirklich zu klassifizierende Objekte im Bild sind - eine wichtige Randbedingung für viele Erkennungssysteme. Weiterhin ist dadurch eine Translationsinvarianz bezüglich der Bildebene inhärent gegeben. Durch eine Stereobildverarbeitung wiederum lassen sich die Objekte aus dem Hintergrund herauslösen - eine deutliche Vereinfachung des Erkennungsproblems. Handhabungsmechanismen können bei unklarer Entscheidungslage weitere Objektansichten zur Ergebnisstabilisierung liefern. Schließlich lassen sich Trackingverfahren nutzen, um den Objekten typische Bewegungsabläufe zuzuordnen. Ein Erkennungssystem kann somit vielfältig von der Einbindung in ein aktives Sehsystem profitieren.

Die Arbeit beschäftigt sich mit der Entwicklung und Evaluierung von 2D/3D-Objekterkennungssystemen für den Einsatz in aktiven Sehsystemen. Hauptfrage hierbei ist natürlich zuerst die nach der erreichbaren Erkennungsrate. Zum anderen soll geklärt werden, in welche Richtung die Entwicklung von Objekterkennungssystemen gehen könnte. Sollte es angestrebt werden, langfristig jede Objektausprägung durch einen äußerst hochdimensionalen Merkmalsvektor eindeutig zu beschreiben? Ein Erkennungssystem wäre dann eine Datenbank aus den Merkmalsvektoren aller gelernten Objekte bzw. Objektansichten. Zur Zeit wäre ein solches System schon aus Gründen der Handhabbarkeit der Datenmengen undenkbar, aber mit der weiteren Erhöhung der Speicherintegration bzw. der Verwendung völlig neuartiger Speicherkonzepte erscheint es realisierbar. Dem gegenüber könnte ein Erkennungssystem stehen, das nicht das Ziel verfolgt "so viele wie möglich" Merkmale zu repräsentieren sondern "so viele wie nötig", d.h. die Merkmalsvektoren sollen hier nur wesentliche Objekteigenschaften repräsentieren. Hier tauchen sofort die nachfolgenden Fragen auf, wie: "Was sind die wesentlichen Objekteigenschaften?" und "Wieviel Information ist noch nötig, um Objekte stabil zu diskriminieren?". Im Laufe der fünfjährigen Forschungsarbeit wurde sich mit dem Aufbau von Erkennungssystemen aus beiden Gruppen beschäftigt. Die Aufgabenstellung wurde bewußt gleich gewählt, um eine gute Vergleichbarkeit zu erreichen.

Zunächst werden im folgenden Kapitel die grundsätzlichen Eigenschaften, der Aufbau und Einsatz von aktiven Sehsystemen erläutert. Im Kapitel 3 schließt sich eine umfassende Übersicht zu im internationalen Rahmen entstandenen Objekterkennungssystemen an. Die dabei vorgestellten Systeme wurden so ausgewählt, daß sie die gesamte Bandbreite von möglichen Erkennungssystemen darstellen. Eine besondere Aufführung erfahren dabei Systeme, die in aktive Sehsysteme integriert sind (Kapitel 3.6). Abschließend findet eine zusammenfassende Wertung der aufgeführten Systeme statt (Kapitel 3.7) von welcher schließlich allgemeine Kriterien zum Entwurf von Erkennungssystemen abgeleitet werden. Kapitel 4 gibt eine Übersicht über wesentliche im Kontext vom *N*euronalen-*A*ctive-*V*ision-*S*ystem NAVIS entstandene Arbeiten sowie über ein 2D-Erkennungssystem als Ausgangspunkt der hier beschriebenen Forschungen. Anschließend werden in drei Kapiteln die entworfenen Erkennungssysteme vorgestellt: die Erkennung mit Modellskalierungen mit dem Hauptaugenmerk auf die Integration von Stereobildverarbeitung zur Verbesserung der Skalierungsinvarianz in das 2D-Erkennungssystem (Kapitel 5), die Erkennung mit zeitlichen Relationen als ein 3D-Erkennungssystem, welches mittels hochdimensionalen Merkmalsvektoren Objektansichten repräsentiert (Kapitel 6) und die Erkennung mit Strukturbeschreibungen als alternatives System, das eine Charakterisierung wesentlicher Objekteigenschaften anstrebt (Kapitel 7). Abgeschlossen wird die Arbeit durch eine Zusammenfassung und einen Überblick über mögliche Weiterentwicklungen.

Kapitel 2

Aktive Sehsysteme

In diesem Kapitel soll sich der Vorstellung des relativ jungen Forschungsfeldes der aktiven Sehsysteme gewidmet werden (siehe auch [MERTSCHING und SCHMALZ 1999], [SCHMALZ 1999]). Zunächst erfolgt eine grundlegende Übersicht über die Ideen und die Motivation aktiver Sehsysteme. Im Anschluß daran wird ein Überblick über internationale Forschungsaktivitäten auf diesem Gebiet gegeben und sich abschließend mit den prinzipiellen Einsatzmöglichkeiten abseits heutiger Prototypen beschäftigt.

2.1 Philosophie

Die Bildverarbeitung als wissenschaftliche Disziplin findet ihren Ursprung u.a. in den Arbeiten von Marr [MARR 1982]. Er legte die Grundlagen zur zielgerichteten Analyse gegebener zweidimensionaler Abbildungen der Umwelt. Dabei bezog er auch Untersuchungen zur frühen visuellen Informationsverarbeitung bei Wirbeltieren ein, wodurch Erfolge bei der Bestimmung von sehrelevanten Gehirnarealen möglich wurden [HUBEL 1988]. Marr ließ in seinen Untersuchungen jedoch außer acht, daß jegliche Analyse von visuellen Daten immer ausgehend von einer Zielvorstellung, einer Absicht, durchgeführt wird und daß dazu der Beobachter u.U. seine Wahrnehmungsorgane erst entsprechend ausrichten muß. Wirbeltiere besitzen nur einen relativ kleinen Bereich absolut scharfen Sehens: die Fovea. Trotzdem sind sie in der Lage, auch komplexe Szenen schnell und robust zu bearbeiten. Eine Szene vollständig und akkurat zu rekonstruieren, würde unrealistisch hohe Anforderungen an die bildverarbeitenden Einheiten stellen. Aloimonos sagt dazu: *"There is simply too much that can be known about the world for a vision system to construct a general-purpose complete description"* [ALOIMONOS 1993]. Hochrechnungen ergaben, daß im Falle einer gleichhohen Auflösung der Retina das Gehirn des Menschen 15.000 kg wiegen würde [SCHWARTZ 1991]. Es ist so-

mit ein schrittweises Abtasten der Umgebung, eine partielle Rekonstruktion der Szene, unter Verwendung spezialisierter Aktorik nötig. Die Art und Weise dieses Abtastvorganges wird im hohen Maße durch die Intentionen des Beobachters beeinflußt. Hierfür hat sich auch der Begriff *purposive* bzw. *qualitative vision* herausgebildet [ALOIMONOS 1990]. Bajcsy sagt dazu: *"We do not just see, we look"* [BAJCSY 1988]. Zu einem gegebenen Zeitpunkt könnte die Suche nach Nahrung sinnvoll sein, zu einem anderen die Suche nach Fluchtwegen. Bei diesen Beispielen kann kaum mit der gleichen Szenen-Abtaststrategie gearbeitet werden. Im ersten Fall muß nach reifen Früchten gesucht werden, die sich evtl. durch ihre Farbe von unreifen unterscheiden, im zweiten Fall sind zunächst wegführende Wege zu finden und anschließend zu prüfen, ob ein Verfolger einen bestimmten Weg, z.B. auf Grund seiner Größe, nicht benutzen kann. Aloimonos stellt dazu heraus: *"... a vision system* is *a collection ... of various processes ... which solves a particular visual task"* [ALOIMONOS 1990]. Die zwei angedeuteten Beispiele beschränken sich auf die aktive Analyse einer Umgebung, d.h. es ist eine unterschiedliche Ausrichtung des Wahrnehmungsapparates nötig. Man kann aber noch weitergehen. U.U. ist auch eine aktive Veränderung von Szenenelementen nötig, um eine bestimmte Intention zu befriedigen. Man stelle sich exemplarisch Nahrung vor, die zunächst unter Blättern o.ä. verborgen ist. Das Sehsystem interagiert hier auch aktiv mit seiner Umwelt. Wie obige Beispiele zeigen, wird sich ständig um das Verständnis von bei Lebewesen ablaufenden sehsystemrelevanten Prozessen bemüht. Dementsprechend bildete sich der Begriff und damit die Forschungsrichtung *animate vision* mit Ballard als einem der Begründer heraus [BALLARD 1991], [BALLARD und BROWN 1992]. Durch die neurophysiologische und psychophysikalische Untersuchung biologischer bildverarbeitender Systeme hofft man, Hinweise über die Struktur und die Arbeitsweise dieser Systeme zu erlangen, um so technischen Sehsystemen neue Anwendungsfelder mit höheren Leistungsanforderungen zu erschließen.

Zusammenfassend kann gesagt werden, daß ausgehend von Marr's Feststellung *"Vision is knowing what is where"* [MARR 1982] sich die Schwerpunkte innerhalb der wissenschaftlichen Bildverarbeitung hin zu einem neuen Verständnis von Sehsystemen verschieben: *"visual systems are visual behaviors tightly integrated with the actions they support"* [SWAIN und STRICKER 1991]. Blake stellt dazu fest: *"... such shifts of emphasis require the development of new technical tools and, initially, pose more problems than they solve."* [BLAKE und YUILLE 1992].

Realisierte technische Systeme zielen aus Komplexitätsgründen bisher meist darauf ab, bestimmte Teilaspekte aktiver Sehsysteme umzusetzen. Hierbei wird häufig von der Realisierung von Perzeptions-Aktions-Zyklen gesprochen, um die sequentielle Verarbeitungsstruktur von Sehen auf der einen und Aktorik auf der anderen Seite hervorzuheben. Im einzelnen kann zwischen vier Hauptforschungs-

richtungen innerhalb aktiver Sehsysteme unterschieden werden:

- Bewegungsanalyse
- Aufmerksamkeitssteuerung
- Handlungsplanung
- Hardwareaufbau

mit einer Fülle von Teil- und Subproblematiken, wie z.B. der Objekt-Hintergrund-Trennung. Darauf aufbauend wird sich mit der Objekterkennung und Szenenrepräsentation auseinandergesetzt. Die Objekterkennung wird hier als ein von Mechanismen des aktiven Sehens profitierendes Modul betrachtet, d.h. Verfahren des aktiven Sehens übernehmen einen Teil der Aufgaben, um z.B. den Suchraum einzuschränken, Invarianzen bereitzustellen oder neue Objektdaten zu gewinnen.

Es ist bis heute weitgehend unklar, welche theoretischen Grundlagen aus welchen Teildisziplinen herangezogen werden müssen, um ein komplexes aktives Sehsystem aufzubauen, das sich hinsichtlich Stabilität, Zuverlässigkeit und Leistungsfähigkeit mit biologischen Systemen vergleichen läßt. Trotzdem ist international eine verstärkte Aktivität und Aufmerksamkeit hinsichtlich des aktiven Sehens zu verzeichnen, die schließlich zu einer größeren Akzeptanz dieser Forschungsrichtung und somit zum grundsätzlichen Lösen der Kernprobleme führen wird [NEUMANN und STIEHL 1993].

2.2 Aktuelle Forschungsansätze

Die besondere Eignung aktiver Sehsysteme wird vor allem innerhalb von Robotik-Anwendungen herausgestellt. Roboter u.ä. Geräte bieten eine einfache Möglichkeit zur Realisierung von Perzeptions-Aktions-Zyklen. Damit kann die sinnvolle Kopplung sensorischer Geräte wie Kamera, Sonar, Laser-Entfernungsmesser, Encoder u.ä. auf der einen Seite und den aktorischen Geräten wie Antriebsmotoren oder Greifvorrichtungen auf der anderen demonstriert werden. Bereits realisierte Applikationen bzw. laufende Forschungsarbeiten unterscheiden sich jedoch grundsätzlich in der Zielstellung der Integration von Methoden des aktiven Sehens. Auf dem Gebiet der Robotik wird Bildverarbeitung zumeist für Navigationszwecke verwendet, d.h. zur Detektion von Hindernissen, zur Suche nach Landmarken zur Selbst-Lokalisierung o.ä. Darauf aufbauend werden aber auch komplexere Problemfelder, wie eine einfache Objekterkennung oder -verfolgung, bearbeitet. Im folgenden wird ein Überblick über aktuelle Forschungsrichtungen

gegeben, der keinesfalls den Anspruch der Vollständigkeit erhebt. Ziel ist das Aufzeigen der Bandbreite aktueller Forschungen. Die ausgewählten Projekte sollen jeweils als Beispiel für typische Forschungs- und Entwicklungsrichtungen dienen, stellvertretend für die Vielzahl von existierenden und entstehenden Ansätzen und Konzepten. Die Forschungslandschaft auf dem Gebiet des aktiven Sehens bzw. des Computersehens als übergeordnetes Foschungsfeld ist dabei aber so vielfältig und aktiv, daß jede Auzählung immer nur temporär gültig ist. Zunächst werden Projekte vorgestellt, die Bildverarbeitung und andere Sensorinformationen hauptsächlich zur Navigation nutzen:

- Die Universität Bonn untersucht autonome Roboter innerhalb des *RHINO*-Projekts[1] [BUHMANN et al. 1995], [THRUN et al. 1998]. Der Rhino-Roboter bewegt sich autonom innerhalb von Büro-Umgebungen unter Nutzung von Mechanismen zur Kollisionsvermeidung und zur Fahrtplanung. Er benutzt künstliche neuronale Netze und Raumplan-Datenbanken, um Szenenrepräsentationen aufzubauen. Alle Algorithmen müssen in einer Echtzeitumgebung ablaufen. Als Sensorinformationen werden Kamerabilder, Sonarechos und Entfernungsdaten eines Lasermeßgerätes verwendet, um den aktuellen Weg zu analysieren und um Objekte grob zu klassifizieren. Rhino wurde als ein Museumsführer im Deutschen Museum Bonn eingesetzt [BURGARD et al. 1998]. Innerhalb dieser Applikation wurde die Kameraplattform allerdings nur zur Steigerung der Nutzerakzeptanz eingesetzt.

- An der Universität München werden visuell gesteuerte Fahrzeuge entwickelt [MAURER und DICKMANNS 1997], [DICKMANNS 1998][2]. Die Fahrzeuge können sich durch die aktive Steuerung von Gaspedal, Bremsen und Lenkung autonom fortbewegen. Das gesamte Steuerungssystem basiert auf visuellen Daten. Dementsprechend sind die Geräte mit mehreren beweglichen Kameras ausgestattet, die Bilder auch mit verschiedenen Zoom-Einstellungen liefern. Als Ergebnis entstanden aktive, binokulare Beobachter, die auch andere Fahrzeuge optisch verfolgen können. In Experimenten wurde die hohe Robustheit und Genauigkeit nachgewiesen.

- Die Universität Karlsruhe entwickelte einen mobilen Serviceroboter, der als Zimmerkellner in einem Hotel eingesetzt wird [HOFMEYER 1998], [GRAF und WECKESSER 1998]. Der Roboter übernimmt hier einfache Serviceaufgaben wie Gepäck- und Wäschetransport und Auslieferung von Getränken. Dabei ist das System vollständig in die Haustechnik integriert, d.h.

[1] http://www.informatik.uni-bonn.de/∼rhino

[2] http://www.unibw-muenchen.de/campus/LRT/LRT13/Fahrzeuge/VaMP_E.html

der Roboter kann Fahrstühle bedienen und Hotelgäste anrufen. Zur Navigation und Kollisionsvermeidung kommen ein Laser-Entfernungsmesser sowie Sonarsensoren zum Einsatz. Ein monokulares Sehsystem soll für einfache Objekterkennungsaufgaben eingesetzt werden. Insbesondere soll ein 3D-Modell der Umwelt aufgebaut werden. Dazu dienen zum einen die Meßwerte des Laser-Entfernungsmessers und zum anderen die visuellen Erkennungsergebnisse für natürliche Landmarken wie Türen und Säulen.

- Das Robotik-Institut der ETH Zürich erforscht autonome mobile Roboter als interaktive und kooperative Maschinen. Innerhalb des *MOPS*-Projekts[3] wird ein autonomer Roboter zur Verteilung der Hauspost verwendet [VESTLI und TSCHICHOLD 1996]. Der Roboter ist in der Lage, durch die Erkennung natürlicher Landmarken sehr genau zu navigieren. Es werden geometrische Strukturen durch die Suche nach geraden Linien, Zylindern, Ecken u.a. innerhalb von Tiefendaten analysiert. Der Roboter greift selbständig die Postboxen und liefert sie an die Adressaten. Dabei ist eine Interaktion über kabellose Verbindungen möglich, um neue Aufträge zu erteilen und Problemsituationen zu lösen.

- Das Robotik-Institut der Carnegie Mellon Universität nimmt am Aufbau eines Prototypen für ein *Automated Highway System (AHS)* teil [BAYOUTH und THORPE 1996][4]. Ziel des Projektes ist die Spezifikation von Fahrzeugen und Fahrbahnen, um eine völlig autonome, ohne menschliche Eingriffe funktionierende Steuerung unter Erhöhung von Sicherheit und Fahrzeugdurchsatz zu ermöglichen. PKWs, LKWs und Busse, die mit *AHS* ausgerüstet sind, sollen gemeinsam mit Fahrzeugen ohne *AHS* auf den gleichen Fahrwegen fahren können.

Sensordatenverarbeitung im Zusammenhang mit aktiven Sehsystemen kann darüber hinaus auch für komplexere Aufgaben eingesetzt werden. Somit dient sie nicht nur als Hilfsmittel für die Navigation sondern auch für die Lokalisierung, Identifizierung oder Verfolgung von Objekten. Im folgenden werden einige Beispiele vorgestellt:[5]

- An der TU Berlin wird der autonome Flugroboter *MARVIN*[6] entwickelt (siehe [MUSIAL et al. 1998] für das Vorgängerprojekt). Es handelt sich hierbei

[3] http://enterprise.ethz.ch/research/mops

[4] http://almond.srv.cs.cmu.edu/afs/cs/misc/mosaic/common/omega/Web/Groups/ahs

[5] verwiesen sei hier auch auf Kapitel 3.6, in dem einige Erkennungssysteme in aktiven Sehumgebungen detailliert vorgestellt werden

[6] http://pdv.cs.tu-berlin.de/MARVIN/index.html

um einen erweiterten Modellhubschrauber zur Teilnahme an Flugroboter-Wettkämpfen. Neben Sensoren zur Stabilisierung des Gerätes wird eine Kamera zum Finden und Klassifizieren von Objekten eingesetzt, wobei an den Objekten einfache Symbole zur Diskriminierung angebracht sind.

- An der Universität Bochum beschäftigte man sich schon früh mit mobilen Sehsystemen [VON SEELEN et al. 1994]. Die mobile Plattform *MARVIN* wurde aus einem kommerziellen System aufgebaut und um ein aktives binokulares Kamerasystem erweitert. Es wurde ein biologisch motiviertes *neural instruction set* für Aufgaben der frühen visuellen Wahrnehmung entworfen. Darauf aufbauend werden Verfahren des aktiven Sehens eingesetzt, um aufgabenorientiert Probleme wie Hindernisvermeidung, Fahrtplanung, Szenenerkennung, Tracking und 3D-Objekterkennung zu bearbeiten. Weiterer Forschungsbedarf wird in der Integration der verschiedenen Steuermodule in eine einheitliche Steuerumgebung gesehen.

  Weiterhin werden an der Universität Bochum Mechanismen zur Gesichtserkennung mit Methoden des aktiven Sehens gekoppelt [VETTER et al. 1997]. Beginnend mit einer Bildaufnahme ermöglicht eine Gesichts-Kandidaten-Segmentierung die Entscheidung über das Vorhandensein eines Gesichts. Erreicht wird dies durch simple Bildverarbeitungsprozesse, die parallel nach den Formen, typischen Bewegungen und Farben von Gesichtern suchen. Die 2D-Positionen und zugehörige Größendaten werden unter Verwendung eines Kalman-Filters verfolgt. Regionen, die Gesichtermerkmale enthalten und die über einen gewissen Zeitraum verfolgt wurden, werden mit Hilfe eines künstlichen neuronalen Netzes hinsichtlich Gesichtern getestet und bilden ein Suchfenster, das dem eigentlichen Gesichts-Klassifizierungsalgorithmus zugeführt wird. Somit wird der Suchraum des Klassifikators durch den Einsatz von Verfahren des aktiven Sehens erheblich verkleinert und damit die Fehlklassifikationsrate gesenkt.

  Innerhalb des *NEUROS*-Projekts[7] wurde der Roboter *ARNOLD* zur Erforschung von menschenähnlichen, autonomen Servicerobotern gebaut [BERGENER et al. 1997]. Der Roboter wird ausschließlich durch visuelle Sensordaten gesteuert, da visuell gesteuerte Roboter in vielen möglichen Anwendungsfällen eingesetzt werden können. Der benutzte Kamerakopf besitzt vier Kameras, die sowohl Übersichtsbilder als auch binokulare, fovealisierte Ansichten liefern. Eine beispielhafte Anwendung von *ARNOLD* besteht in der Lokalisierung von Menschen basierend auf deren Hautfarbe. Nach einer Erkennung soll er mit seiner Greifvorrichtung auf die Hand des Menschen zeigen.

[7] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/PROJECTS/NEUROS/NEUROS.html

- An der Universität Stuttgart werden autonome mobile Roboter zur Navigation in unbekanntem Gelände und zur Klassifizierung komplexer, sich untereinander sehr ähnlicher Objekte eingesetzt [OSWALD und LEVI 1997][8]. Innerhalb des Forschungsprojektes werden die im Zusammenhang mit kooperativer Bildverarbeitung auftretenden Probleme untersucht, d.h. mehrere Roboter sollen ein unbekanntes Objekt selbständig von verschiedenen Perspektiven aus betrachten. Dementsprechend wird das Systemverhalten eines Roboters nicht nur durch die eigene visuelle Wahrnehmung, sondern auch durch die Interaktion mit seinen Nachbarrobotern beeinflußt. Innerhalb des Beispielverhaltens "Verfolgen und Erkennen von Objekten" liefert eine Bewegungsanalyse ein Aufmerksamkeitsfenster, das dann durch ein verteilt arbeitendes Erkennungsmodul bearbeitet wird.

- Die Universität Tübingen beschäftigt sich ebenfalls mit der Entwicklung eines kooperativen Verhaltens von Robotern [ZELL et al. 1997a][9]. Als Beispiel sollen mehrere mobile Roboter sich selbständig innerhalb einer Schlange anordnen. Andere laufende Projekte zielen auf das Design von autonomen mobilen Robotern ab [ZELL et al. 1997b].

- An der TU Ilmenau werden biologisch motivierte neuronale Architekturkonzepte zur Handlungsorganisation in aktiv lernenden, verhaltensorientierten Systemen untersucht [GROSS und BÖHME 1997][10]. Der Roboter *MILVA* soll mit Menschen auf eine natürliche und direkte Weise kommunizieren. Als exemplarische Aufgabenstellung wird ein Bahnhofsszenario verwendet. Hierbei soll der Roboter auf Gesten von Reisenden reagieren und ihnen beim Gepäcktransport oder beim Finden eines Bahnsteiges helfen. Diese Aufgabenstellungen bedingen Fähigkeiten zum Erkennen von Händen und Gepäck, zum visuellen und motorischen Verfolgen von Personen und zur Kollisionsvermeidung bzw. Navigation.

- An der Universität Paderborn wird innerhalb des *DEMON*-Projektes ein System zur automatischen Demontage von Fahrzeugen entwickelt [GÖTZE et al. 1998][11]. Zunächst sollen PKW-Räder demontiert werden. Dazu muß das System die räumlichen Positionen der einzelnen Teile mit adäquater Genauigkeit bestimmen können [TRAPP et al. 1998] . Ein mit Spezialwerkzeugen und einem Stereo-Kamera-System ausgerüsteter Roboterarm soll die Fahrzeuge schrittweise zerlegen. Objekte bzw. Fahrzeugteile

[8] http://www.informatik.uni-stuttgart.de/ipvr/bv/comros
[9] http://www-ra.informatik.uni-tuebingen.de/forschung/welcome_e.html
[10] http://cortex.informatik.tu-ilmenau.de/forschung.html
[11] http://getwww.uni-paderborn.de/projekte/demon.html

werden hierbei holistisch als eine Menge charakteristischer Teilobjekte gelernt und später von einer beliebigen Distanz bei beliebiger Orientierung erkannt. Diese Invarianzen werden durch eine hypothesenbasierte Normalisierung erreicht. Die geometrischen Relationen der Teilobjekte sollen zur effektiven Modellierung komplexer Objekte in ein wissensbasiertes System einfließen.

- An der Universität Ulm wird die Interaktions-Organisation zwischen Modulen der symbolischen und subsymbolischen Ebene, konkret zwischen wissensbasierten Systemen und künstlichen neuronalen Netzwerken, für ein autonomes Fahrzeug untersucht [PALM und KRAETZSCHMAR 1997][12]. Ziel ist die Entwicklung eines lernfähigen kognitiven Apparates, einer zentralen Repräsentation von sensomotorischen Situationen, um eine Basis für den Entwurf eines Roboter-Verhaltens zu schaffen.

  An der Abteilung für Neuroinformatik werden u.a. biologisch motivierte Verfahren zur Roboternavigation und Hindernisvermeidung entwickelt [BARATOFF et al. 1999][13]. Benutzt wird der optische Fluß zur aktiven, visuell gestützten Verhaltenssteuerung. Weiterhin kommen log-polare Karten zum Einsatz, um zum einen eine Datenkompression bzw. Echtzeitfähigkeit zu erreichen und zum anderen, um die Fixation eines aktiven Sehsystems zu erleichtern.

- Die Universität Stanford beschäftigt sich ebenfalls mit der Entwicklung autonomer Roboter [BECKER et al. 1997], [GONZÁLEZ-BAÑOS et al. 1999][14]. Der "Intelligente Beobachter" soll in der Lage sein, Objekten zu folgen und mit Personen natürlichsprachlich, z.B. mit Hilfe des Kommandos "Folge dem nächsten sich bewegenden Objekt", zu kommunizieren. Das Projekt vereint Konzepte und Algorithmen der Bildverarbeitung, Bewegungsplanung und Computergrafik zum Aufbau eines robusten, integrierten Systems. Ein bedeutender Systemaspekt ist die Verwendung von Bildverarbeitung nicht nur zur Unterstützung der Navigation sondern als zentraler Mechanismus des Systems. Der Roboter besitzt mehrere Kameras, die es ihm erlauben, gleichzeitig Objekte visuell zu verfolgen und seine eigene Umgebung zu analysieren. Es werden fünf Module vorgestellt: Landmarken-Detektion, visuelle Objektverfolgung, Bewegungsplanung, Benutzerschnittstelle und Bewegungssteuerung. Die ersten beiden Module benutzen direkt Ergebnisse der Bildverarbeitung. Zur Landmarkendetektion werden spezielle Marken an bekannten Positionen

[12] http://www.uni-ulm.de/SMART/index.e.html
[13] http://www.informatik.uni-ulm.de/zoweg/Forschung/NInf.html
[14] http://underdog.stanford.edu

auf dem Fußboden angebracht. Wird eine Landmarke detektiert, werden insgesamt 16 Binärwerte aus den Grauwerten der inneren Markenstruktur ausgelesen. Diese Werte kodieren zum einen die Nummer der Marke als auch deren Orientierung. Visuelle Objektverfolgung bezieht sich hier auf die Detektion und Verfolgung von zylindrischen "Hüten" mit einer charakteristischen Streifenzeichnung, die auf einem Zielroboter montiert werden.

- An der Universität Rochester werden Rollstühle durch den Anbau von Sensoren und Verarbeitungseinheiten in mobile Roboter umgewandelt [BAYLISS et al. 1997][15]. Diese sollen u.a. ein Zielobjekt erkennen und motorisch verfolgen, welches an einem anderen, handgesteuerten Roboter befestigt ist. Auch hier wird ein speziell konstruiertes 3D-Kalibrierungsobjekt verwendet, um die Rechenkomplexität zu reduzieren. Damit das System echtzeitfähig wird, ist die Steuerung hierarchisch gestaltet. Untergeordnete Prozesse sind verantwortlich für die schnelle Lokalisierung eines Aufmerksamkeitsfensters, übergeordnete führen dann die geometrische Analyse und die Verfolgung durch. Das Verfolgungsmodul besteht aus vier Teilen: Vorhersage, Projektion, Messung und Rück-Projektion. Durch die bekannte Objekt-Konfigurationen können alle Verarbeitungsstufen erheblich vereinfacht werden.

- Das *CVAP*-Labor des *Royal Institut of Technology Stockholm* erforscht Systemansätze im Bereich des aktiven Sehens [CHRISTENSEN und CROWLEY 1998][16]. Folgende Schlüsselelemente eines aktiven Sehsystems werden herausgestellt: Die Benutzung von mehreren Typen visueller Information, von Aufmerksamkeitsmechanismen, von Objekt-Hintergrund-Trennung, von Hinweisen aus einer Bewegungsanalyse und von Tiefenhinweisen aus binokularen Kamerasystemen. Dadurch können Objekte effektiv detektiert und weiterverarbeitet werden. Aufmerksamkeitsmechanismen werden im Zusammenhang mit der Bewegungs- und Tiefenanalyse zur weiteren Erhöhung der Effektivität des Systems benutzt.

- Ein recht neues und äußerst spannendes Forschungsfeld für aktive Sehsysteme sind die Fußballweltmeisterschaften für mobile Roboter [KITANO 1998], [ASADA und KITANO 1999][17]. Hier werden die Anwendungsmöglichkeiten visuell gesteuerter autonomer Systeme in anschaulicher Weise deutlich gemacht. Daß sich dabei den Problemfeldern zunächst

[15] http://www.cs.rochester.edu/research/mobile/main.html
[16] http://www.nada.kth.se/cvap
[17] http://www.robocup.org

spielerisch genähert wird fördert zum einen eine breite gesellschaftliche Akzeptanz bzw. erregt weltweites Interesse und zum anderen können genügend einschränkende Bedingungen formuliert werden, so daß die Aufgaben überhaupt bearbeitbar werden. Beispielsweise steht die Ballfarbe im hohen Kontrast zum Untergrund und eine weiße Bande vermeidet Hintergrundeffekte. Nichtsdestoweniger werden typische Fragen aktiver Sehsysteme bearbeitet: schnelles aktiven Reagieren auf sensorischen Input unter Beachtung einer spezifischen Intention.

Die kurze Aufstellung aktueller Forschungsaktivitäten zeigt zunächst, daß die Entwicklung aktiver Sehsysteme noch starken Forschungsbezug besitzt. Es überwiegt kreatives, jedoch meist intuitives Handeln. Über den Grundaufbau von Perzeptions-Aktions-Zyklen herrscht weitgehend Konsens (siehe z.B. [SOMMER 1999]). Die konkrete Umsetzung und vor allem die Einschätzung der Leistungsfähigkeit der entwickelten Systeme ist aber nicht umfassend, vor allem im theoretischem Bereich, untersucht. Hier, wie ja ebenfalls auf dem Gebiet der klassischen Bildverarbeitung, spielen Erfahrungswerte die herausragende Rolle. Um langfristig erfolgreich zu sein, müssen die aktiven Sehsysteme aber hinaus aus den Laboren in, wenn auch zunächst nur einfache, Anwendungsbereiche. Es fehlt an Benchmarks für aktive Sehsysteme. Die Fähigkeit zur Lösung einer konkreten Aufgabe aus dem Laborumfeld reicht zur Einschätzung der allgemeinen Leistungsfähigkeit nicht aus. Durch die Fußballweltmeisterschaft für mobile Roboter steht erstmalig ein Testfeld zur Verfügung, durch das verschiedene Ansätze vergleichbar werden. Die Erfahrung zeigte hierbei, daß vor allem die gewählten Algorithmen über Sieg oder Niederlage entscheiden und weniger die z.T. sehr aufwendigen Ausrüstungen.

Die angestrebten Anwendungsfelder werden meist nur als Demonstration der Möglichkeiten, weniger als produktionsreife Entwicklungen betrachtet. Aktive Sehsysteme haben noch nicht Einzug in den industriellen oder privaten Alltag gefunden. Begründet ist dies u.a. auch durch die bisher ungenügenden visuellen Möglichkeiten der Systeme. Erfolgreiche Beispielanwendungen besitzen meist Sonar- oder Lasersensoren zur Navigation oder Hindernisvermeidung. Die visuelle Leistungsfähigkeit beschränkt sich i.a. auf die Detektion einfacher, speziell generierter, synthetischer Landmarken. Eine darüber hinausgehende komplexere Objekterkennung findet nicht statt. Diese Trennung von Sensorik zur Navigation und Zielerkennung ist häufig anzutreffen, da die Auswertung von Informationen von Sonar- oder Lasersensoren hardwaretechnisch gut unterstützt wird und somit meist in Echtzeit möglich ist. Durch ihre deutlich höheren Rechenzeitanforderungen haben visuelle Informationen z.Z. noch schlechtere Voraussetzungen für einen umfassenden Einsatz. Der Vorteil visueller Informationen ist aber, daß sie für alle anfallenden Aufgaben, von der Navigation bis zur Zielobjekterkennung,

verwendet werden können (siehe z.B. [BOLLMANN et al. 1999]).

2.3 Einsatz von aktiven Sehsystemen

Im vorherigen Kapitel wurden einige Forschungsarbeiten bezüglich aktiven Sehsystemen vorgestellt. Darauf aufbauend sind eine Reihe von Anwendungsfeldern denkbar. Mit der ständig steigenden Rechenleistung und die Verfügbarkeit von Sensorsystemen mit entsprechenden Analysealgorithmen können anspruchsvolle Problemkreise erschlossen werden. Prinzipiell können die Konzepte des aktiven Sehens in die folgenden Anwendungsaspekte eingebracht werden:

- *Industrie-Robotik:* Robotern innerhalb industrieller Anwendungen fehlt heute meist noch eine Art intelligentes Verhalten. Sie sind lediglich in der Lage, einmal programmierte Bewegungsabläufe beliebig oft und sehr genau zu wiederholen. Daraus ergibt sich aber auch eine grundlegende Voraussetzung: eine konstante Arbeitsumgebung mit gleichbleibenden Aufgabenstellungen. Schon kleine Veränderungen der Lage z.B. der Werkzeuge verursachen empfindliche Störungen. Aktive Sehsysteme bieten eine effektive Lösung dieser Probleme. Das langwierige und aufwendige Reprogrammieren der Roboter für neue Aufgabenstellungen kann erheblich reduziert werden. Der Roboter könnte sich auf der Basis von Zielvorgaben selbständig an neue Arbeitsumgebungen anpassen.

- *Service-Robotik:* Auf dem Gebiet der Dienstleistungssysteme entwickeln sich breite Anwendungsfelder für mit aktiven Sehsystemen gekoppelte Roboter. Service-Roboter sind i.a. mobile Systeme, die Menschen von einfachen Aufgaben entlasten. Beginnend mit einfachen Systemen, wie dem sichtgesteuerten Staubsaugen oder dem autonomen Zustellen der Hauspost, werden diese zukünftig ihren Platz im normalen Alltag einnehmen. Auf dem Gebiet der Pflege und Unterstützung von Behinderten wird besonders der Bedarf an leistungsfähigen, autonomen Systemen zur Steigerung der Lebensqualität bzw. zur Entlastung der Pflegepersonen deutlich.

- *Autonome Fahrzeuge:* Auf Fahrzeugen installierte aktive Sehsysteme bieten Möglichkeiten zur Unterstützung des Fahrers in schwierigen Situationen, zur Erhöhung des Fahrzeugdurchsatzes oder zum selbständigen Lösen von Transportaufgaben. Auch innerhalb unwirtlicher Umgebungen können solche Systeme erfolgreich eingesetzt werden. Durch Nutzung der Konzepte des aktiven Sehens, d.h. der Suche und Bearbeitung von bezüglich der Aufgabenstellung interessierenden Regionen, wird die Vielzahl der Stimuli

zur Navigation handhabbar. Zusammen mit Industrierobotern können autonome Fahrzeuge ganze Produktionsbereiche, wie z.B. Lagerhaltung und Hochbau, selbständig übernehmen.

- *Mensch-Maschine-Interaktion:* Durch aktives Sehen kann die Schnittstelle zwischen Mensch und Computer wesentlich nutzerfreundlicher gestaltet werden. In Verbindung mit Spracherkennern muß der Nutzer keine Kommandos mehr per Hand eingeben. Mit Hilfe von Bildverarbeitungsprozessen können auch sonst nur schwer erkennbare Wort- und Silbenkonstrukte aufgelöst werden, indem z.B. Mundbewegungen analysiert werden. Eine bildliche Verfolgung der Kopfpositionen liefert immer die genaue Position des Nutzers. Eine Interpretation der Blickrichtung wäre ebenfalls denkbar, um Bildschirmpositionen auszuwählen.

- *Überwachung/Authentisierung:* Aktive Beobachter können zum einen unautorisierte Aktionen feststellen und zum anderen bei Montage auf einem mobilen System autonom patrouillieren. Desweiteren sind auch Verifikationssysteme für Unterschriften durch den Einsatz einer Finger-Bewegungsanalyse möglich.

Kapitel 3

Einordnung von Objekterkennungsansätzen

Dieses Kapitel soll eine Übersicht über die verschiedenen Forschungsrichtungen zur Objekterkennung geben. Es folgt zunächst eine kurze Darstellung zweier wesentlicher Kriterien zum Entwurf bzw. zur Einordnung von Erkennungssystemen: das Repräsentationsschema und die Verarbeitungshierarchie. Besonders letzteres wird beim Einsatz von Erkennungssystemen in aktiven Sehumgebungen von Bedeutung sein. Anschließend werden eine Reihe von postulierten Erkennungsansätzen vorgestellt. Am Ende des Kapitels folgt eine zusammenfassende Wertung der vorgestellten Arbeiten sowie eine Analyse des derzeit erreichten Forschungsstandes verbunden mit einer Kritik. Darauf aufbauend werden schließlich Forschungsdesiderate für den Entwurf von Erkennungssystemen in aktiven Sehsystemen abgeleitet.

3.1 Grundlagen

3.1.1 Repräsentationsformen

Eine der Grundfragen beim Entwurf von Objekterkennungssystemen ist die nach der zu verwendenden Repräsentationsform. Man unterscheidet hierbei objektzentrierte und betrachtungszentrierte Darstellungen. Bei ersteren werden Objekte als Ganzes beschrieben, d.h. durch die Repräsentation ist eine Beschreibung aller Objektteile und damit aller Objektansichten implizit gegeben. Die betrachtungszentrierten Ansätze gehen jeweils von spezifischen Ansichten des Objektes aus, die sowohl zwei- als auch dreidimensional beschrieben werden können, und benötigen somit mehrere solcher Modellansichten, um ein Objekt vollständig zu repräsentieren. In Abbildung 3.1 sind die beiden Formen für ein einfaches Ob-

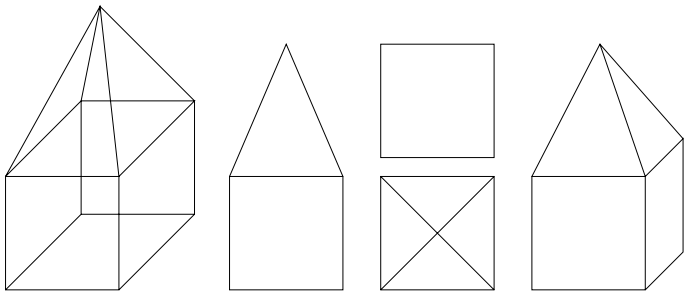


Abbildung 3.1: Gegenüberstellung von *a)* objektzentrierten und *b)* betrachtungszentrierten Repräsentationen

Tabelle 3.1: Qualitativer Vergleich von objektzentrierten und betrachtungszentrierten Repräsentationen

	REPRÄSENTATION IST	
	OBJEKTZENTRIERT	BETRACHTUNGSZENTRIERT
Vorteile	Die möglichen Ansichten der Objekte sind durch eine einzige Repräsentation gegeben. Durch den Einsatz von Grundformen, wie volumetrischen Primitiven, kann eine gute Generalisierung erreicht werden.	Die Generierung bzw. Analyse einer einzigen Objektansicht ist oft erheblich einfacher und schneller durchzuführen.
Nachteile	Erhebliche Komplexität bei der Berechnung des Modells und der Analyse einer gegebenen Szene.	Gesamtdarstellung eines komplexen Objektes benötigt u.U. eine große Anzahl von einzelnen Darstellungen.
Beispiele	[MARR und NISHIHARA 1978] [BIEDERMANN 1985, BIEDERMANN et al. 1993] [LOWE 1986] [THOMPSON und MUNDY 1987] [SHAO und KITTLER 1996] [LIU et al. 1998]	(siehe weiter in diesem Kapitel)

jekt dargestellt, und in Tabelle 3.1 wird ein qualitativer Vergleich vorgenommen. Weitergehende Darstellungen zu den objektzentrierten Repräsentationen sind z.B.

[MASSAD 1997] zu entnehmen.

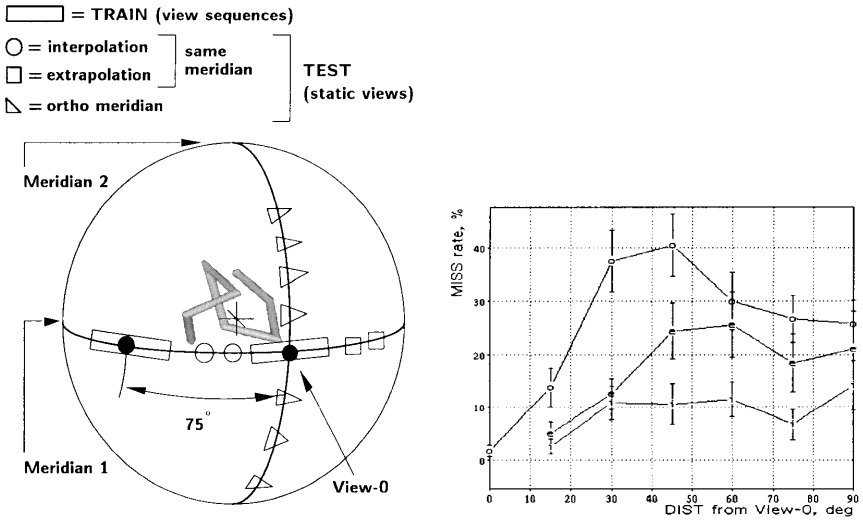
Innerhalb dieser Arbeit wird sich ausschließlich auf betrachtungszentrierte Darstellungen konzentriert, da diese zum einen Vorteile in ihrer Handhabbarkeit bzw. Berechenbarkeit besitzen und zum anderen neuropsychologische Untersuchungen auf das Verwenden dieser Repräsentationsform innerhalb des visuellen Systems von Primaten hindeuten. Experimente mit Versuchspersonen zeigen eine kontinuierliche Abnahme der Erkennungsleistung, je weiter die präsentierten Ansichten von den gelernten entfernt werden [BÜLTHOFF et al. 1995]. In Abbildung 3.2 ist der entsprechende Versuchsaufbau und die Fehlklassifizierungsrate angegeben. Präsentiert werden u.a. unvertraute "Draht-Objekte" in Ansichten zwischen bereits gesehenen (Abbildung 3.2a). In Abbildung 3.2b werden die Fehlklassifizierungsraten in drei Klassen angegeben: *o* für orthogonal zu gelernten Ansichten präsentierte Draht-Objekte (obere Kurve); *e* für präsentierte Ansichten, die innerhalb der *view-sphere* nicht zwischen zwei gelernten liegen (mittlere Kurve); *i* bezeichnen Ergebnisse für Ansichten, die sich zwischen zwei gelernten einordnen lassen. Die Ergebnisse geben deutliche Hinweise auf betrachtungszentrierte Objektrepräsentationen beim Menschen. Die Fehlklassifizierung ist umso stärker, je weiter eine Objektansicht sich innerhalb der *view-sphere* von einer bereits gesehenen Ansicht entfernt. Sowohl unter monokularer als auch unter stereoskopischer Betrachtung zeigen sich diese Ergebnisse.

In [PERRETT et al. 1991] werden diese Erkenntnisse auch auf neurophysiologischer Ebene nachgewiesen. Bei Einzelzellableitungen im superioren temporalen Sulcus von Makaken zeigen sich starke Abhängigkeiten von präsentierten Ansichten (siehe Abbildung 3.3a). Zellen, die auf die Gesamtheit aller Ansichten reagieren (objektzentrierte Darstellung), werden als Überlagerung aller betrachtungssensitiven Zellen interpretiert (Abbildung 3.3b).

Tanaka gelangt in [TANAKA 1996] zu ähnlichen Ergebnissen. Er weist auf gesichtssensitive Zellen hin, die nicht nur betrachtungszentriert reagieren, sondern auch systematisch angeordnet sind.

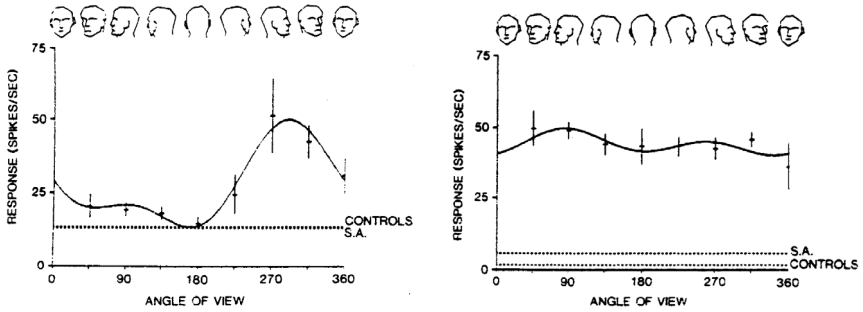
3.1.2 Verarbeitungszweige

Aus physiologischen und psychologischen Forschungen an visuellen Systemen von Primaten lassen sich Modelle für eine biologisch motivierte mehrstufige Objekterkennung von Sensorinformationen ableiten [TREISMAN 1986]. Auf Grund von Verhaltensstudien gibt es Hinweise darauf, daß auf höheren kortikalen Bereichen ein in Bezug auf ihre Funktionalität getrennte Informationsverarbeitung stattfindet. Die im inferioren temporalen Cortex (IT) konzentrierten Verarbeitungsmechanismen dienen hauptsächlich der Diskriminierung von Objekten untereinander, basierend auf deren Form, Farbe, Textur oder Größe. Probanden mit Läsionen in dieser Region können entspre-



a) b)

Abbildung 3.2: Psychophysikalische Experimente mit Drahtobjekten (nach [BÜLTHOFF et al. 1995]). *a)* Präsentation synthetischer Draht-Objekte; abgebildet ist die räumliche Lage der Trainings- und Testansichten. *b)* Fehlklassifizierungsraten in Abhängigkeit von der Entfernung der Testansicht von einer gelernten Ansicht.



a) b)

Abbildung 3.3: Einzelzellableitungen an Makaken (nach [PERRETT et al. 1991]). *a)* Antwortzeiten einer Zelle, die nur auf bestimmte Ansichten reagiert. *b)* Antwortzeiten einer objektzentrierten Zelle als Überlagerung aller betrachtungszentrierten Zellen.

chende Aufgaben nicht mehr zufriedenstellend lösen. Im Gegensatz dazu ist nach einer Beschädigung des posterioren parietalen Cortex (PP) das Lösen von Erkennungsaufgaben hinsichtlich der Wahrnehmung der räumlichen La-

ge von Objekten beeinträchtigt. Dementsprechend liegt der Schluß nahe, daß der IT vor allem mit Aufgaben zur Klassifizierung von Objekten (*Was* - Verarbeitungszweig) beschäftigt ist und der PP räumliche Beziehungen herstellt (*Wo* - Verarbeitungszweig) [UNGERLEIDER und MISHKIN 1982]. Zellableitungen geben ebenfalls Hinweise auf diese dedizierte Aufgabenverteilung [ANDERSEN et al. 1985], [FUJITA et al. 1992]. Desweiteren zeigt sich eine Verbindung zwischen selektiver Aufmerksamkeit und dem IT. Zum Beobachtungszeitpunkt nicht mit Aufmerksamkeit belegte Szenenregionen werden demzufolge ausgeblendet [MORAN und DESIMONE 1985]. Auf der anderen Seite kann auch eine Verstärkung zuvor bereits bearbeiteter Stimuli beobachtet werden [WISE und DESIMONE 1988] bzw. eine Abschwächung von bereits untersuchten Orten [KLEIN 1988]. Zu beachten ist, daß hierbei die Aufmerksamkeitsstrategien sowohl datengetrieben (*bottom-up*) als auch durch kognitive Prozesse wissensgetrieben (*top-down*) arbeiten [CHELAZZI et al. 1993].

3.2 Einordnung

Bei der Fülle der während der letzten Jahre im internationalen Rahmen entstandenen Arbeiten zur Objekterkennung ist die Suche nach einer möglichst umfassenden Taxonomie wünschenswert. Problematisch erscheint hierbei die Auswahl der Kriterien zur Einordnng der Erkennungsansätze. Zum einen könnte das gewählte Repräsentationsschema besonders interessant sein, zum anderen weisen eventuell die postulierten Matchstrategien Besonderheiten hinsichtlich Robustheit, Geschwindigkeit o.ä. auf. In Tabelle 3.2 wird eine grobe Klassifizierung vorgestellt. Gemeinsam ist allen Verfahren, soweit nicht anders betont, die Verwendung von betrachtungszentrierten Repräsentationen. Die Einordnung von Erkennungssystemen anhand nur eines Kriteriums stellt einen Kompromiß dar, da solche Systeme mindestens nach drei Kriterien hin untersucht werden sollten: die verwendete Merkmalsbasis, die gewählte Repräsentation und die Erkennungsstrategie. Aus Gründen der Übersichtlichkeit und um die besonderen Aspekte eines Systems hervorzuheben, wurde die hier dargestellte Einteilung gewählt. Im einzelnen wird nach dem verwendeten Grundprinzip der Erkennung unterschieden, d.h. die Tabelle ist im wesentlichen nach dem verwendeten Matchverfahren indiziert. Eine Ausnahme bilden die Ansätze unter Nutzung von “Strukturbeschreibungen” in der Klasse der “geometrischen Verfahren”, die vor allem Repräsentationsaspekte behandeln und u.a. als Grundlage zum Entwurf des im Kapitel 7 beschriebenen Objekterkennungssystems dienten. Eine weitere Sonderstellung nehmen die “Verfahren in aktiven Sehumgebungen” ein, die als letzte Gruppe vorgestellt werden, um zu verdeutlichen, daß sie die Zusammenführung verschiedenster Forschungsgebiete, wie die der Objekterkennung, Aufmerksamkeitssteuerung und Robotik,

darstellen. Es wird demonstriert, wie Objekterkennungssysteme durch die Nutzung von Methoden des aktiven Sehens profitieren. Die Aufführungen von international realisierten Erkennungssystemen erhebt keinesfalls den Anspruch der Vollständigkeit. Sind Verfahren nicht aufgeführt, bedeutet dies keinesfalls eine negative Wertung. Absicht ist einzig, die Spanne derzeit vorhandener Lösungen, ihre Leistungsfähigkeit aber auch Schwachstellen aufzuzeigen. Für die Auswahl der präsentierten Arbeiten wurde vor allem betrachtet, inwieweit diese Ansätze die typischen Eigenschaften und Prinzipien ihrer Klasse repräsentieren.

In den folgenden Kapiteln werden die in der Tabelle 3.2 referenzierten Arbeiten näher erläutert. Schwerpunkt der Darlegungen ist der grundsätzliche Ansatz des Systems sowie eine Einschätzung über dessen Leistungsfähigkeit. Eine abschließende, zusammenfassende Wertung findet im Kapitel 3.7 statt.

Tabelle 3.2: Grobklassifizierung von Objekterkennungsansätzen (wird fortgesetzt)

GEOMETRISCHE VERFAHREN	
	• Formenanalyse [DRYDEN und MARDIA 1998], [DICKINSON und METAXAS 1997], [NELSON 1995]
	• Suchgraphen [ESHERA und FU 1984], [GMÜR und BUNKE 1989], [BUNKE und MESSMER 1995], [CHUNG 1995], [BURGER et al. 1996], [LOURENS und WÜRTZ 1998]
	• Hashverfahren [BELLAIRE 1995], [BELLAIRE und LÜBBE 1995], [COSTA und SHAPIRO 1996], [BELLAIRE et al. 1996]
	• Strukturbeschreibungen [SHAPIRO und HARALICK 1981], [SHAPIRO et al. 1984], [ZHANG et al. 1992], [SHAMS 1995], [WILLIAMSON 1996b], [BENNAMOUN und BOASHASH 1997], [FISHER und MACKIRDY 1998]
OPTIMIERUNGSVERFAHREN	
	• Hopfield-Netze [LI und NASRABADI 1989], [GEE et al. 1991], [LIN et al. 1991], [CHAO und DHAWAN 1992], [KAWAGUSHI und SETOGUCHI 1994], [RAY und MAJUMDER 1994], [LEE et al. 1995], [SCHAFFER und CHEN 1995]
	• *Annealing*-Verfahren [GOLD und RANGARAJAN 1996], [AZENCOTT und YOUNES 1997]

Tabelle 3.2: Grobklassifizierung von Objekterkennungsansätzen

	• Relaxationsverfahren [FINCH und HANCOCK 1995], [SCHWARZINGER et al. 1995], [WILSON und HANCOCK 1995], [SHAO und KITTLER 1996],
BIOLOGIENAHE SYSTEME	
	• biologisch motivierte Systeme [BIESZCZAD 1994], [BASAK und PAL 1995], [GÖTZE et al. 1996], [MASSAD et al. 1998a]
	• konnektionistische Systeme [BALLARD 1984], [CRUZ et al. 1990], [ALEXANDRE et al. 1991], [KOLLIAS et al. 1991], [ANDO 1996]
VERFAHREN IN AKTIVEN SEHUMGEBUNGEN	
	[GIEFING et al. 1992a], [RYBAK et al. 1994], [RAO und BALLARD 1995], [FUKUNAGA et al. 1996], [HERBIN 1996], [SIPE und CASASENT 1997], [GVOZDJAK und LI 1998], [LIU et al. 1998], [ONISHI et al. 1998]

3.3 Geometrische Verfahren

3.3.1 Formenanalyse

Häufig werden einfache, aber leistungsfähige geometrische Zusammenhänge genutzt, um Formen bzw. Objekte normiert zu beschreiben und diese Beschreibungen zur Wiedererkennung zu verwenden. Grundidee ist das Beseitigen von Abhängigkeiten bezüglich einer affinen Transformation. In diesem Zusammenhang ist der Begriff der Formen-Koordinatensysteme üblich. Die Formen werden in ein Koordinatensystem transformiert, in dem die gewünschten Invarianzen integriert sind. Eine umfangreiche Aufstellung diesbezüglicher Verfahren und mathematischer Grundlagen findet sich in [DRYDEN und MARDIA 1998]. Dort werden ausgehend von dem Problem des Vergleichs von Landmarken einige Formen-Koordinatensysteme vorgestellt:

- *traditionelle Methoden:* In diese Kategorie fallen einfache Methoden, die Distanzen bzw. Winkel ins Verhältnis setzen, um so die Abhängigkeiten von der konkreten Merkmalsposition im 2D-Raum zu beseitigen. Weiterhin können die zu untersuchenden Objekte auch auf primitive geometrische Formen, wie z.B. Dreiecke, zurückgeführt werden. Diese einfacheren Objekte können dann invariant z.B. durch die eingeschlossenen Winkel beschrieben werden. Solche Beschreibungen werden z.B. in [LIN et al. 1991],

[SHAO und KITTLER 1996], [NELSON 1995] und [KOLLIAS et al. 1991] verwendet.[1]

- *Bookstein-Koordinaten:* Hierbei werden die ersten beiden Landmarken auf fest definierte Koordinaten transformiert und alle anderen Landmarken entsprechend der so gefundenen Transformationsvorschrift [BOOKSTEIN 1984]. Anschließend können die Landmarken von Modell und Szene direkt miteinander verglichen werden. Um Symmetrien beizubehalten, werden die Basispunkte meist auf $(-\frac{1}{2}, 0)$ und $(\frac{1}{2}, 0)$ gelegt. *Bookstein*-Koordinaten $u_j^B + iv_j^B$ ergeben sich aus den ursprünglich komplexen Koordinaten $z_1^0, \ldots, z_k^0$ für die k Landmarken wie folgt:

$$u_j^B + iv_j^B = \frac{z_j^0 - z_1^0}{z_2^0 - z_1^0} - \frac{1}{2}, \quad j = 3, \ldots, k \tag{3.1}$$

  Koordinaten werden hier und im folgenden durch komplexe Zahlen beschrieben, da sich damit Drehungen und Skalierungen einfacher ausdrücken lassen. Nachteil dieses besonders anschaulichen Verfahrens ist, daß Korrelationen in die Koordinaten eingebracht werden und die ersten beiden Merkmalspunkte sehr exakt detektiert werden müssen.

- *Kendall-Koordinaten:* Diese Koordinaten ähneln den Bookstein-Koordinaten, die absoluten Positionen werden nur in einer anderen Art und Weise beseitigt [KENDALL 1984]. Zunächst wird eine *Helmert*-Submatrix H definiert. H ist die $(k-1) \times k$ *Helmert*-Matrix ohne die erste Zeile. Die vollständige *Helmert*-Matrix H^F ist eine $k \times k$ quadratische Matrix, deren erste Zeile $1/\sqrt{k}$ entspricht. Die anderen Zeilen sind jeweils orthogonal zur ersten Zeile. Somit ergibt sich die $j$te Zeile von H zu:

$$(h_j, \ldots, h_j, -jh_j, 0, \ldots, 0), \quad h_j = -\{(j(j+1)\}^{-1/2} \tag{3.2}$$

  mit $j = 1, \ldots, k-1$ für k Landmarken. Die $j$te Zeile von H besteht somit aus j Wiederholungen von h_j, gefolgt von $-jh_j$ und $(k-j-1)$ Nullen. Die komplexen Landmarken $z^0 = (z_1^0, \ldots, z_k^0)^T$ werden zunächst durch eine Multiplikation mit H zu den invarianten Beschreibungen $z_H = Hz^0 = (z_1, \ldots, z_k)^T$ überführt. Die *Kendall*-Koordinaten $u_j^K + iv_j^K$ für k Landmarken ergeben sich aus:

[1] genauere Beschreibungen dieser Ansätze nachfolgend in diesem Kapitel

$$u_j^K + iv_j^K = \frac{z_j - 1}{z_1}, \quad j = 3, \ldots, k \tag{3.3}$$

- *Watsons-Dreiecks-Koordinaten:* Watson entwarf ein Koordinatensystem für Dreiecks-Formen [WATSON 1986]. Wenn $z = (z_1, z_2, z_3)^T \in \mathcal{C}^3$ die komplexen Landmarken sind, $\omega = e^{i\pi/3}$, $u = (1, \omega, \omega^2)^T$ und $\overline{u} = (1, \omega^2, \omega)^T$, dann sind 1_3 (repräsentiert das degenerierte Dreieck, d.h. alle Punkte liegen aufeinander), u und $\overline{u}$ orthogonal zueinander mit der Länge $\sqrt{3}$. Alle Dreiecke lassen sich damit als

$$z = \alpha 1_3 + \beta u + \gamma \overline{u} \tag{3.4}$$

  beschreiben. Da z dieselbe Form wie $cz + d1_3$ mit $c, d \in \mathcal{C}$ besitzt, kann die Form eines Dreiecks durch

$$z = u + b\overline{u}, \quad b \in \mathcal{C} \tag{3.5}$$

  angegeben werden. Damit lassen sich Dreiecke nur durch die Angabe der komplexen Zahl b beschreiben. Zerlegt man nun die zu beschreibenden nicht Dreiecksformen in Dreiecke, können beliebige Formen durch die Parameter b_i repräsentiert werden.

- *Tangentenraum-Koordinaten:* Diese Koordinaten bauen auf einen Formen-Raum auf, der alle möglichen Formen enthält, wobei die Formen die geometrischen Informationen der Landmarken invariant zur Position, Rotation und Skalierung enthalten. Der Tangentenraum ist dann eine linearisierte Version des Formen-Raums in der Nähe eines bestimmten, meist eine durchschnittliche Form repräsentierenden Punktes. Mit den komplexen Landmarken $z^0 = (z_1^0, \ldots, z_k^0)^T$ ergibt sich

$$z = (z_1, \ldots, z_{k-1})^T = \frac{Hz^0}{||Hz^0||} \tag{3.6}$$

  zur Beseitigung der Abhängigkeiten von der Position und Skalierung.

  Werden nun die Landmarken um θ rotiert und anschließend auf die Tangentenebene von γ, bezeichnet durch $T(\gamma)$, projiziert, ergibt sich der Abstand v zu (siehe auch Abbildung 3.4):

$$v \;=\; e^{i\hat{\theta}}\,[I_{k-1} - \gamma\gamma^*]\,z, \quad v \in T(\gamma),\, \hat{\theta} = \arg(-\gamma^* z) \tag{3.7}$$

γ ist ein komplexer Pol, d.h. die durchschnittliche, zu untersuchende Form. γ^* bezeichnet die transponierte der komplex konjugierten von γ. Da $v^*\gamma = 0$, bedeuten die komplexen Beziehungen, daß man den Tangentenraum als einen Subraum von $\mathcal{R}^{2k-2}$ mit der Dimension $2k-4$ betrachten kann. Die Matrix $[I_{k-1} - \gamma\gamma^*]$ ist die Matrix zur komplexen Projektion auf den orthogonalen Raum zu γ. Die euklidischen Entfernungen im Tangenten-Raum sind eine gute Approximation für die Ähnlichkeit der Formen bzw. für ihre Distanzen im Formen-Raum.

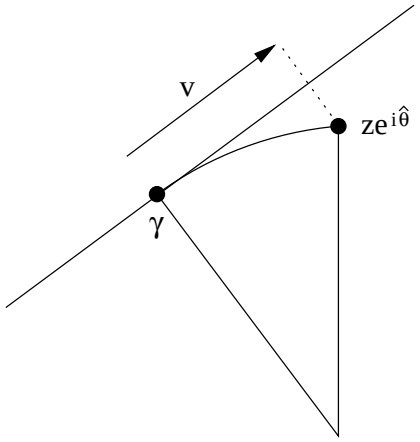


Abbildung 3.4: Bezug der Tangenten-Koordinaten v zu einem Formen-Raum (nach [DRYDEN und MARDIA 1998])

Die dargestellten Formen-Koordinatensysteme würden sich anbieten, um auch komplexe Objekte nach einer Zerlegung in geometrische Primitive, wie z.B. Dreiecke, zu bearbeiten. Nach der Transformation in einen invarianten Formen-Raum lassen sich einfache Abstandsmaße definieren, die die Ähnlichkeit von Modell- bzw. Szenenformen widerspiegeln.

In [DICKINSON und METAXAS 1997] werden Objekte unter Verwendung von deformierbaren Modellen und Aspektgraphen erkannt. Basis ist eine *recognition-by-parts*-Strategie, die zunächst 3D-Volumenanteile von Objekten bestimmt und anschließend diese durch die Analyse gegebener Aspekt-Graphen in Zusammenhang bringt. Als grundlegende Volumenmodelle kommen 10 verschiedene Modellierungsprimitive zum Einsatz. Innerhalb der Arbeit wird eine Kombination

von objekt- und betrachtungszentrierter Repräsentation verwendet. Volumetrische Teilobjekte werden mittels Volumenprimitive objektzentriert beschrieben, Teilaspekte der Objekte jedoch betrachtungszentriert. Als Vorteile dieses hybriden Ansatzes werden hervorgehoben, daß

- die Anzahl prinzipieller Aspekte relativ klein bleiben wird, da die Anzahl qualitativer Volumenprimitive ebenfalls klein ist. Dadurch bleibt eine Datenbasis auch für viele Objekte überschaubar.

- die Anzahl der Aspekte unabhängig von der Objektanzahl in der Datenbasis ist, da jeweils während der Trainingsphase nur die Ansichten der Volumenprimitive berechnet werden müssen.

Zur Repräsentation der Modellobjekte kommen Aspekthierarchien mit drei Ebenen zum Einsatz: der Ebene der Aspekte, die sich aus verschiedenen *faces* zusammensetzen und diese wiederum aus Begrenzungsprimitiven. Aus den Aspekten werden schließlich die Volumenprimitive gebildet (siehe Abbildung 3.5 für Details).

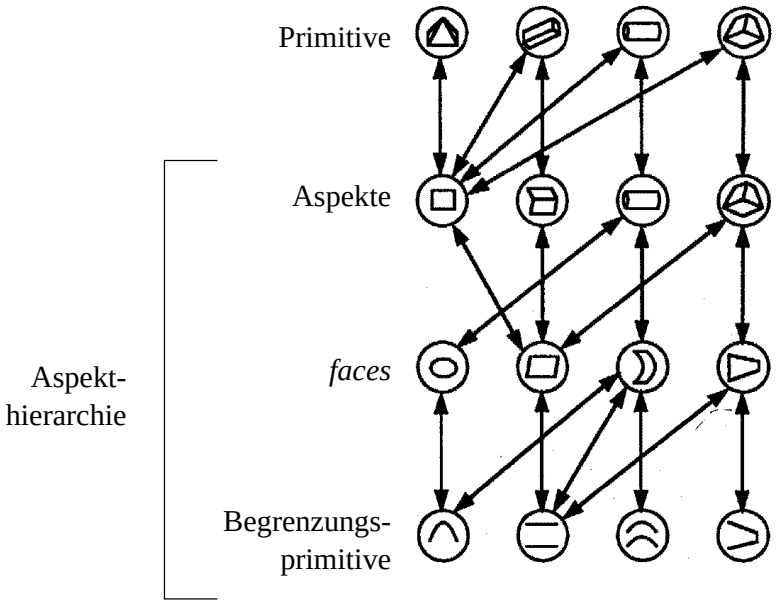


Abbildung 3.5: Die Aspekthierarchie (nach [DICKINSON und METAXAS 1997]); Aufwärtsverbindungen geben “Elternbeziehungen” an; Abwärtsverbindungen symbolisieren “Komponentenbeziehungen”

Die vertikalen Beziehungen werden durch *Teil-von*-Relationen beschrieben. Die Bestimmung der hierarchischen Beziehungen zwischen den Ebenen erwies sich als schwierig. Dementsprechend kommen bedingte Wahrscheinlichkeiten zum Einsatz. Die Wahrscheinlichkeiten werden durch eine Häufigkeitsanalyse von Merkmalsbeziehungen von in der *view-sphere* benachbarter Objektaufnahmen abgeleitet. Invarianzen werden durch eine grobe Kodierung von Merkmalen und durch die Verwendung qualitativer Merkmalsaussagen einbezogen. Die *faces* und Begrenzungsprimitive kodieren jeweils qualitative Beziehungen wie Koterminierung, Parallelität und Symmetrie zwischen qualitativ definierten Konturen.

Das Erkennungssystem besitzt zwei Erkennungsmodi: zum Matchen von beliebigen, unerwarteten Objekten und zur Suche nach einem bestimmten, vorher festgelegten Objekt. In den Abbildungen 3.6 und 3.7 werden beide Modi verdeutlicht.

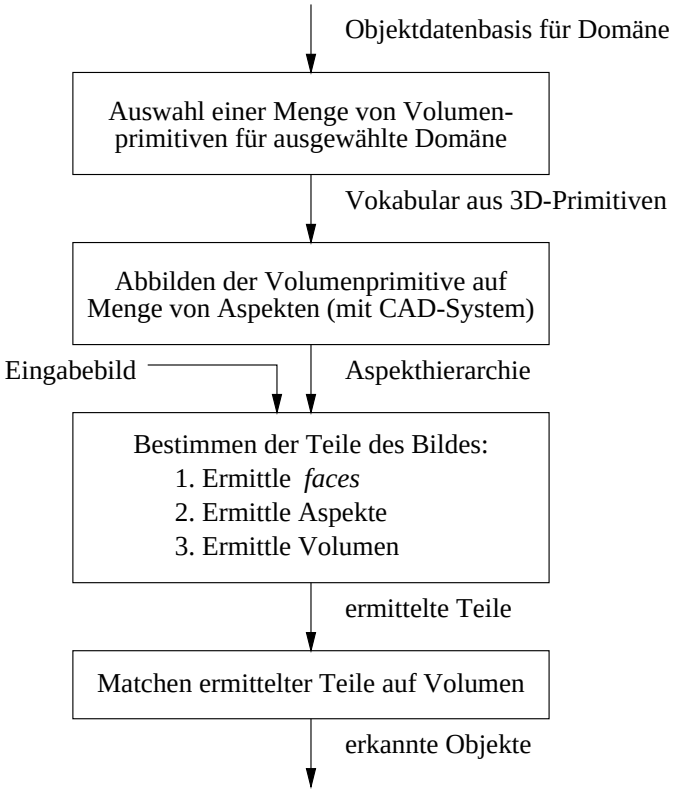


Abbildung 3.6: Erkennung von unerwarteten Objekten (nach [DICKINSON und METAXAS 1997])

Die gegebenen 3D-Objektmodelle werden auf eine 2D-Ebene projiziert. Diese Projektion wird mit der präsentierten Bildansicht in Beziehung gesetzt, wo-

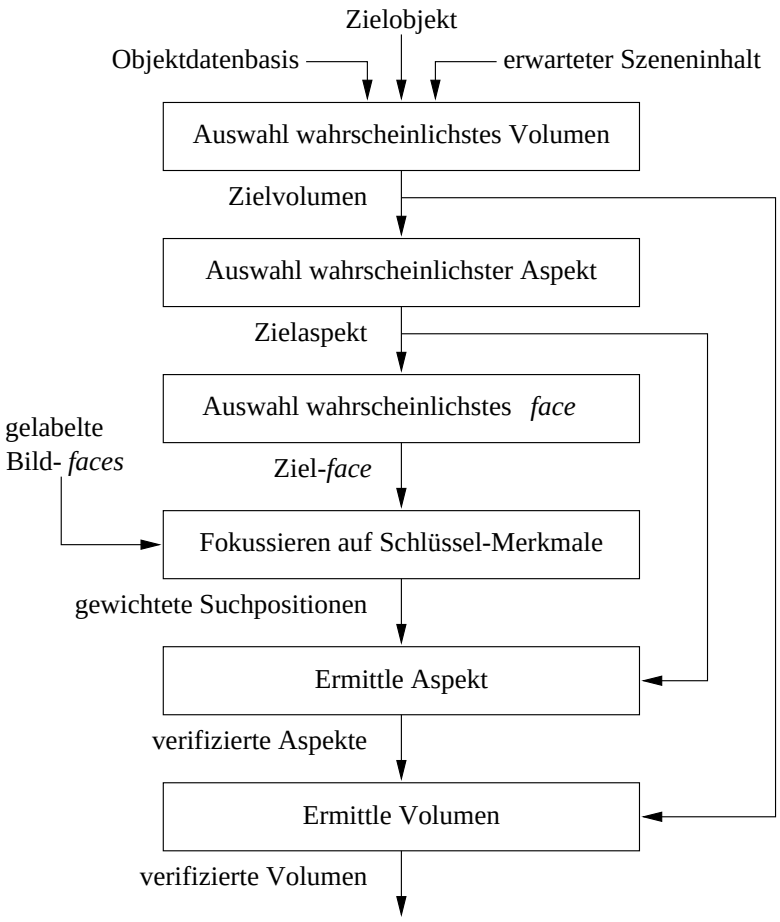


Abbildung 3.7: Erkennung von erwarteten Objekten (nach [DICKINSON und METAXAS 1997])

durch sich Kräfte zur Deformation des 3D-Modells ableiten lassen. Die Entfernungen zwischen den gegebenen bzw. den aus dem Szenenbild ermittelten Aspekten und den projizierten Modellaspekten werden zu “2D-Bildkräften” konvertiert. Das Modell wird entsprechend deformiert und die Projektion mit den ermittelten Aspekten in Übereinstimmung gebracht.

Die zehn Basisformen werden aus deformierten Ellipsoiden e abgeleitet:

$$e = a \begin{pmatrix} a_1 C_u^{\epsilon_1} C_v^{\epsilon_2} \\ a_2 C_u^{\epsilon_1} S_v^{\epsilon_2} \\ a_3 S_u^{\epsilon_1} \end{pmatrix}, \tag{3.8}$$

wobei $-\frac{\pi}{2} \leq u \leq \frac{\pi}{2}$, $-\pi \leq v < \pi$, $S_\omega^\epsilon = \mathrm{sgn}(\sin\omega)|\sin\omega|^\epsilon$, $C_\omega^\epsilon = \mathrm{sgn}(\cos\omega)|\cos\omega|^\epsilon$, $a \geq 0$ ein Skalierungsparameter, $0 \leq a_1, a_2, a_3 \leq 1$ Parameter für das Aspektverhältnis und $\epsilon_1, \epsilon_2 \geq 0$ Parameter für die Rechtwinkligkeit sind. Zur Deformation kommen translatorische, rotatorische, globale und lokale Kräfte zum Einsatz. Die Bewegungsgleichung ergibt sich zu:

$$\mathbf{M}\ddot{q} + \mathbf{D}\dot{q} + \mathbf{K}q = g_q + f_q, \tag{3.9}$$

wobei $\mathbf{M}$, $\mathbf{D}$ und $\mathbf{K}$ die Matrizen zur Begrenzung der Deformationen bezüglich Masse, Dämpfung und Steifheit sind. Der Vektor q kodiert die Parameter aller zugelassenen Deformationen des Modells, g_q gibt die wirkenden Trägheitskräfte an und f_q die von außen auf das Modell einwirkenden Deformationskräfte. Zur Vereinfachung des Aufwandes beim Matchen werden anschließend die Matrizen $\mathbf{M}$ und $\mathbf{K}$ zu 0 gesetzt. Zur Initialisierung des Deformationsprozesses wird von einer Segmentierung der Szene in Objektteile ausgegangen.

An einigen Beispielen wird die Funktionsweise des Verfahrens demonstriert. Jedoch kommen nur einfache, polyedrische Objekte zum Einsatz. Weder Segmentierungsfehler noch Hintergrundinformationen werden berücksichtigt. Bei Verwendung der einfachen Objekte konnte sich das ausgewählte Modell gut an die vorhandenen Szenendaten anpassen. Weiterhin bleibt aber ungeklärt, wie aus verrauschten, schlecht segmentierten Bildern die einzelnen Volumenanteile des Objektes extrahiert und anschließend das richtige Modell ausgewählt wird. Der Einsatz parametrisierbarer Ellipsoiden ist insofern günstig, da alle Volumenprimitive mit Hilfe einer Gleichung beschrieben werden können. Jedoch ergibt wiederum die hohe Anzahl der Parameter einen großen Suchraum und damit hohe Rechenkosten. Dadurch sind wieder starke Vereinfachungen nötig, die wiederum die Diskriminierung ähnlicher Objekte erschwert. Zusammenfassend ist festzustellen, daß die Autoren vordringlich an einer mathematischen Ausformulierung des Erkennungsproblems interessiert waren. Allerdings mußte diese auch für die Erkennung sehr einfacher Objekte unter optimalen Bedingungen schon stark vereinfacht werden. Der resultierende Parameterraum für alle Erkennungsschritte ohne entsprechendes Vorwissen wird sonst einfach zu groß und damit nur schwer handhabbar.

In [NELSON 1995] wird ein Erkennungssystem basierend auf Assoziativspeichern vorgestellt. Es wird explizit zwischen einer Hypothesen- und Validierungsphase unterschieden. Als Objektbasis dienen 3D-Objekte, die durch ihre Ansichten repräsentiert werden. Grundidee ist die Verwendung sog. *keys* zur Beschreibung von Merkmalen. Ein *key* ist ein robustes Merkmal mit ausreichend invarianter Information zur Diskriminierung der Objektansichten und Parametern, die ein

einfaches Indizieren ermöglichen. Folgende Anforderungen werden an die Eigenschaften der *keys* spezifiziert:

- Die Merkmale müssen komplex genug sein, um nicht nur die Konfigurationen der Objekte ausreichend zu beschreiben, sondern um auch genügend freie Parameter für eine Indizierung bereitzustellen.
- Die Wahrscheinlichkeit, daß dieses Merkmal detektiert werden kann, muß substantiell groß sein.
- Die Index-Parameter sollen sich nur gering ändern, wenn sich die Konfiguration eines Objekts ändert.

Durch diese strengen Anforderungen ergibt sich eine starke Domänenabhängigkeit. Innerhalb der Arbeit wird sich auf die Erkennung von Polyedern konzentriert. Dementsprechend werden *key*-Merkmale ausgewählt, die sich auf Ketten von Liniensegmenten beziehen. Es werden zunächst Liniensegmente im Bild gesucht, die anschließend zu perzeptuellen Gruppen aus drei Segmenten zusammengefaßt werden. Diese Liniensegmentgruppen sollen jeweils eine 3D-Begrenzung beschreiben. In Abbildung 3.8 sind zwei Beispiele angegeben.

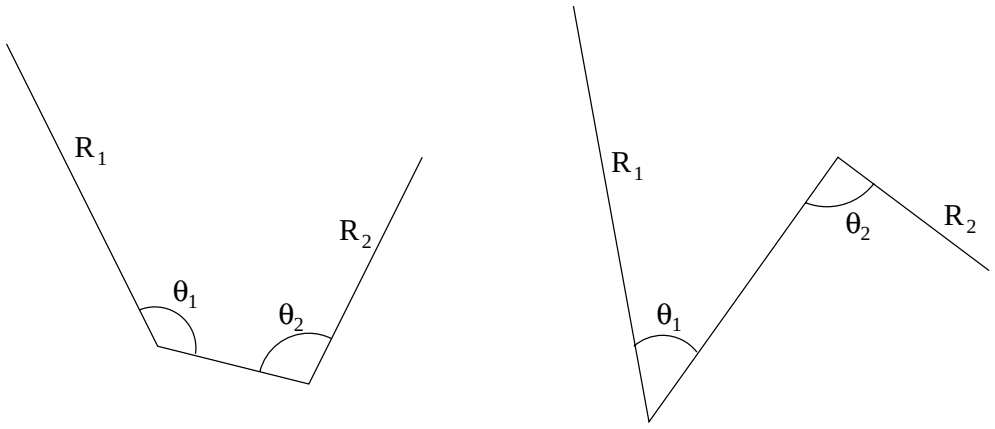


Abbildung 3.8: Beispiele für Liniensegmentgruppen mit invarianten Winkeln und Längenverhältnissen (nach [NELSON 1995])

Die *key*-Merkmale erzeugen assoziierte Hypothesen für die Identitäten und Konfigurationen aller Objekte, die die Merkmale hätten erzeugen können. Innerhalb einer zweiten assoziativen Stufe wird basierend auf der statistischen Analyse des Auftretens der *keys* die wahrscheinlichste Hypothese ausgewählt, indem die Evidenzen der einzelnen Hypothesen akkumuliert werden. Auf die assoziativen

Speicher wird über ein Hash-Regime zugegriffen. Das Problem, wie die einzelnen Merkmale zur Evidenz einer Hypothese beitragen, wird durch die Anwendung einer Merkmals-Häufigkeits-Verteilung gelöst. Evidenz wird proportional zu $\log(1 + \frac{1}{kx})$ akkumuliert, wobei x die Wahrscheinlichkeit ist, daß die Datenbank-Statistik zu der bestimmten Match-Beobachtung führt und k eine Proportionalitätskonstante, die eine Beziehung zwischen der Vorhersage der Ausprägungen aus den *key*-Merkmalen und der wirklichen geometrischen Ausprägungen herstellt. Zur Aquisition von Modellwissen wird die Verwendung eines aktiven Agenten vorgeschlagen, der selbständig seine Umgebung exploriert und Objektwissen sammelt. Weiterhin sollen Verfahren des Aktiven Sehens zur Aufmerksamkeitssteuerung eingesetzt werden. Durch die Verwendung eines Aufmerksamkeitsfensters sollen störende Merkmale ausgeblendet werden. Innerhalb von Experimenten konnte die Funktionsfähigkeit des Verfahrens demonstriert werden - allerdings nur für einfache, geometrische Grundformen (Dreiecke, Vierecke, Kreuze, u.a.) und mit fehlendem Hintergrund.

3.3.2 Suchgraphen

Bei der Verwendung von Suchgraphen sind i.a. sowohl die Modelle als auch die Szene als Graph gegeben. Meist werden durch die Knoten lokale Merkmale und durch die Kanten relationale Beziehungen kodiert. Das Matchen zweier Graphen beinhaltet im allgemeinsten Fall den Vergleich der beiden Graphen. Verschiedene Abstufungen lassen sich definieren (nach [BUNKE und MESSMER 1995]):

- *Isomorphismus:* Ein Isomorphismus ist eine bijektive Abbildung der Knoten des einen Graphen auf die Knoten des anderen, wobei die Struktur der Kanten erhalten bleiben soll. Zwei Graphen sind also isomorph, wenn sie strukturell identisch sind.

- *Subgraphisomorphismus:* Hierbei werden zwei Graphen G_1 und G_2 dahingehend untersucht, ob ein Subgraph aus G_2 isomorph zu G_1 ist.

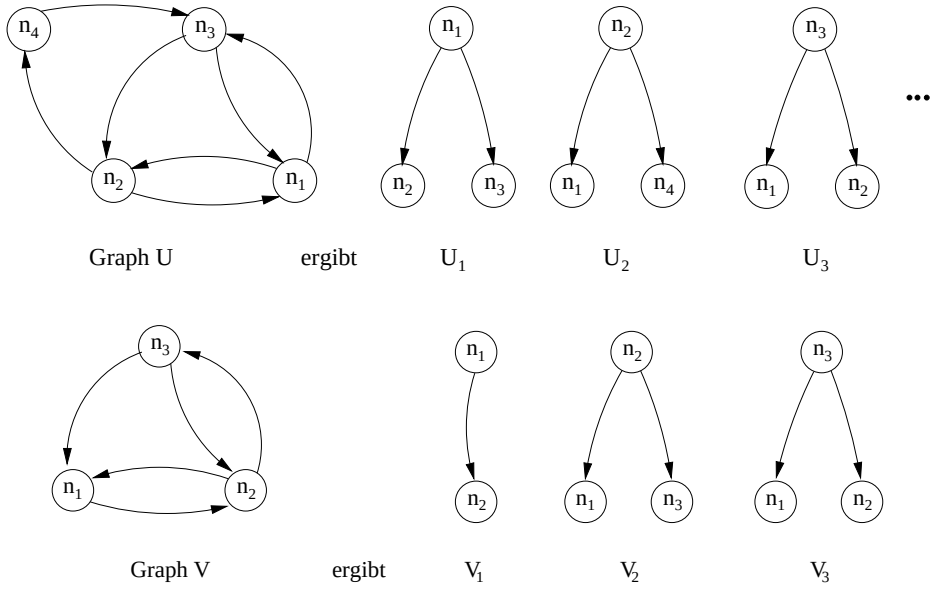
- *bidirektionaler Subgraphisomorphismus:* Aus den zwei zu matchenden Graphen werden jeweils Subgraphen herausgelöst, die auf Isomorphismus untersucht werden.

- *fehlertolerantes Graphmatchen:* Das auszuführende Graphmatchen soll gewisse Defekte innerhalb der Graphen tolerieren. Damit können bei der Segmentierung oder der Graphengenerierung auftretende Fehler kompensiert werden. Es ist die Definition eines Fehlermodells notwendig, das sog. *edit*-Operationen wie Einfügen, Löschen oder Substituieren von Knoten bzw.

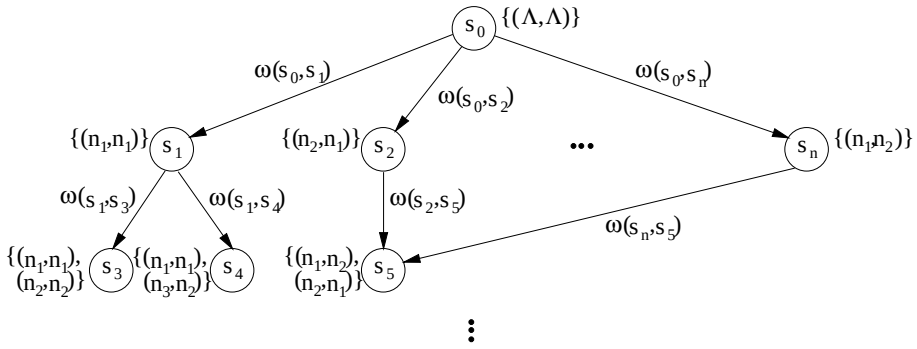
Kanten beinhaltet. Üblicherweise werden für die einzelnen Operationen Kosten definiert. Ein Graphmatchproblem stellt sich somit als die Suche nach einer Folge von *edit*-Operationen dar, wobei minimale *edit*-Kosten entstehen sollen.

In [ESHERA und FU 1984] wird die Zweckmäßigkeit der Nutzung von Graphenstrukturen in Erkennungssystemen hervorgehoben. Hauptaugenmerk liegt dann auf der Entwicklung von Abstandsmaßen für Objektgraphen. Die Autoren definieren ein Abstandsmaß auf der Basis einer Kostenfunktion, die die Transformation des einen Graphen, oder eines Subgraphen, in den anderen Graphen beschreibt. Der Abstand bzw. die Ähnlichkeit zweier Graphen wird bestimmt durch die Anzahl der nötigen Knotenentfernungen, -einfügungen und -substitutionen. Ein Erkennungsvorgang läuft in folgenden Stufen ab:

1. Zerlegung der Objektgraphen in primitive, gerichtete *basic*-Graphen, indem pro Knoten die wegführenden Kanten aufgeführt werden:



2. Generierung einer Zustandsraumdarstellung in Form eines gerichteten, azyklischen, markierten Gitters. Jeder Knoten s_i repräsentiert einen Rekonstruktionsschritt aus zwei Subgraphen der zu matchenden Objektgraphen U und V, wobei das Matchen der entsprechenden *basic*-Graphen mit kodiert wird. Die Verzweigungen innerhalb des Gitters werden mit den Kosten für den Rekonstruktionsprozeß gelabelt:



3. Suche des kürzesten Pfades über das Gitter. Ausgehend vom gewünschten Zielzustand wird rückwärts bis zum Startpunkt gesucht. Als Ergebnis wird zum einen der Weg über das Gitter geliefert und zum anderen die Distanz dieses Weges.

Der vorgestellte Ansatz zeichnet sich durch eine umfassende Komplexitätsanalyse und durch genaue Algorithmenbeschreibung in Form von Pseudocode aus. Leider werden keine praktischen Versuche angegeben, wodurch sich nur schwer die Leistungsfähigkeit abschätzen läßt. Problematisch erscheint allerdings, daß, sollte das System zur Erkennung realer Objekte eingesetzt werden, die Generierung der Graphenrepräsentationen im hohen Maße vom Ergebnis der Segmentierung des Bildes abhängt.

In [GMÜR und BUNKE 1989] wird ein Verfahren zur 3D-Objekt-Erkennung durch Graphmatchen vorgestellt. Eingesetzt wird es in dem Roboter-Seh-System *PHI-1*. In Abbildung 3.9 ist die System-Architektur abgebildet. Basierend auf 3D-Volumen-Daten von einem CAD-System werden die Modellgraphen aufgebaut. Jede Modellbeschreibung beinhaltet zwei Graphen: einen Ecken-Kanten-Graph und einen Flächen-Kanten-Graph, wobei letzterer nur während der Verifikationsphase benutzt wird.

Aus dem präsentierten Grauwertbild wird ein Bildgraph abgeleitet. Dazu werden zunächst Kanten detektiert und evtl. zu Kurven verbunden. Anschließend werden diese Strukturen in gerade Linien, Kreise und Ellipsen klassifiziert. Kanten an Kreuzungspunkten werden zu Ecken verbunden und ebenfalls gelabelt. Anschließend werden sich überlappende Objekte getrennt und ein Bildgraph generiert. Das Erkennungssystem arbeitet zweistufig. Zunächst wird durch eine Subgraphisomorphiesuche eine Hypothese gebildet, anschließend diese durch geometrische Rückprojektionen überprüft. Zur Bildung einer Hypothese wird ein Breite-Zuerst Suchalgorithmus basierend auf [ULLMANN 1976] eingesetzt. Es werden eine Reihe von *constraints* zur Komplexitätsverringerung der Graphensuche eingesetzt:

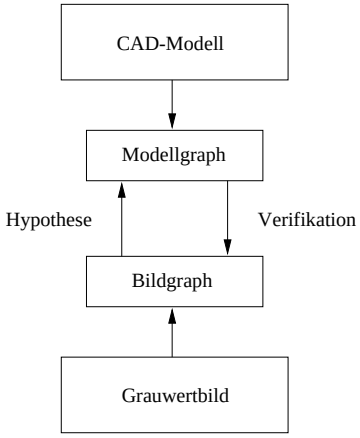


Abbildung 3.9: Systemarchitektur von *PHI-1* (nach [GMÜR und BUNKE 1989])

- *Nachbarschafts-constraints:* Es können Nachbarn eines Knotens im Szenengraph nur mit Nachbarn des matchenden Modellgraphen in Beziehung stehen.

- *Geometrische constraints:* Nach dem erfolgreichen Matchen von drei Knoten können bei gegebener Kameracharakteristik (Vergrößerungsfaktor) zu lange Kanten ausgefiltert werden. Nach dem Matchen von mindestens vier Knoten kann bei Annahme einer Parallelprojektion die Projektionsmatrix $\mathbf{P}$ mit

$$\vec{v_p} \;=\; \mathbf{P}\vec{v_m} \;=\; \mathbf{TRS}\vec{v_m} \tag{3.10}$$

  errechnet werden. $\vec{v_p}$ und $\vec{v_m}$ sind die Koordinaten eines Bild- bzw. Modellknotens. Die Projektionsmatrix **P** setzt sich dabei aus einer Translationsmatrix **T**, einer Rotationsmatrix **R** und einer Skalierungsmatrix **S** zusammen. Bei Kenntnis der Projektionsmatrix können die Positionen der weiteren Graphenelemente vorausberechnet werden.

- *Topologische constraints:* Nur passende Eckentypen dürfen gematcht werden und unter Benutzung der errechneten Projektionsmatrix kann die Sichtbarkeit von Kanten und Ecken in der Szene vorausberechnet werden.

- *Such-constraints:* Die Suchtiefe wird begrenzt. Dadurch verringert sich nicht nur die Komplexität des Algorithmus, sondern auch die Wahrscheinlichkeit von Fehlzuweisungen durch Segmentierungsfehler und Schatteneffekte.

Weiterhin erwies sich die Wahl des Startpunktes des Strukturvergleichs als kritisch. Die besten Ergebnisse wurden mit Startknoten, die möglichst viele Nachbarn unterhalb euklidischer Entfernungsschranken besitzen, erzielt. Nach Bildung einer Hypothese wird diese durch eine geometrische Projektion verifiziert. Dieser Prozeß stellt im Gegensatz zur Hypothesengenerierung ein *top-down*-Prozeß dar. Es wird die berechnete Projektionsmatrix auf den ausgewählten Modellgraphen angewendet. Der so erzeugte Projektionsgraph wird mittels Tiefe-Zuerst-Suche durchlaufen und dabei die Ähnlichkeit zu dem Objektgraphen ermittelt. Als Ähnlichkeitsfunktion dient die Distanz D_g zweier Graphen:

$$D_g = \sum_{V_{j_{mat}}} w_j d_j + \sum_{V_{k_{ins}}} w_k c_{ins} + \sum_{V_{l_{del}}} w_l c_{del} \tag{3.11}$$

$V_{j_{mat}}$ bezeichnet hierbei die Anzahl der Knoten, $V_{k_{ins}}$ die Anzahl der eingefügten und $V_{l_{del}}$ der gelöschten Knoten. Der erste Term charakterisiert geometrische Zuordnungsfehler durch die Distanz d_j zwischen Bild- und Projektionsgraph, die beiden folgenden beinhalten die Kosten zum Einfügen bzw. Löschen von Knoten in den Projektionsgraphen c_{ins} und c_{del}. Alle betrachteten Knoten können mit w_i spezifisch gewichtet werden. Entsprechend ergibt der Graph mit minimalen D_g die Matchentscheidung. Das dargestellte Verfahren zeichnet sich durch umfangreiche Komplexitätsanalysen sowie damit einhergehenden Vorschlägen zu deren Verringerung aus. Das Verfahren wird demonstriert an einem Beispiel mit industriellen Objekten mit Verdeckungen, allerdings ohne Hintergrund. Das System konnte die sich überlappenden Objekte trennen, labeln und Modellgraphen zuordnen. Obwohl zunächst nur anhand von künstlichen Objekten demonstriert[2] wird doch die Leistungsfähigkeit der auf Graphen basierenden Objekterkennung deutlich. Darauf aufbauend wird in [BUNKE und MESSMER 1995] nochmals die besondere Eignung von Graphen zur Lösung von Objekterkennungsaufgaben hervorgehoben. Vorgestellt wird nun ein Algorithmus zur Subgraphisomorphie-Suche, der sich durch eine effektive Handhabung auch großer Modelldatenbanken und durch die Möglichkeit des inkrementellen Erweiterns der Datenbank auszeichnet. Aus den Modellgraphen werden symbolische Beschreibungen, sog. *networks of prototypes*, abgeleitet. Dadurch wird eine komprimiertere Darstellung der Graphelemente erreicht, indem mehrfach auftretende Elemente nur einmal aufgeführt werden. Der Algorithmus folgt einem “Teile und Herrsche”-Ansatz. Zur Bestimmung einer Subgraphisomorphie zwischen dem Graphen G und einem anderen Graphen g wird G in zwei disjunkte Subgraphen G_1 und G_2 zerlegt und diese Subgraphen auf Subgraphisomorphie zu g überprüft. Ist diese gegeben und bleibt

[2]Hierbei muß auch das Erscheinungsjahr (1989) der Arbeit bezüglich der damals verfügbaren Rechenleistung berücksichtigt werden.

die Kantenstruktur in g erhalten wird auf eine Subgraphisomorphie zwischen G und g geschlossen. Konnte keine Isomorphie festgestellt werden, wird die Teilung von G fortgesetzt, maximal bis zum Erreichen der Knotenebene. Dieser Algorithmus wird um fehlertolerante Eigenschaften erweitert. Es werden auch nicht exakte Übereinstimmungen akzeptiert. Dazu werden *edit*-Kosten festgelegt, die innerhalb einer *best-first*-Suche berücksichtigt werden, um die Subgraphen mit den geringsten Abweichungen zuerst zu untersuchen. Weiterhin kommt ein *lookahead*-Ansatz zum Einsatz, der vorausschauend die *edit*-Kosten in einem ausgewählten Verarbeitungszweig schätzt. Die Leistungsfähigkeit des Ansatzes wird anhand der Erkennung von technischen Symbolen demonstriert.

Chung stellt ein Erkennungssystem vor, daß ebenfalls auf die Anwendung von Algorithmen zur Bestimmung von Subgraphisomorphismen beruht [CHUNG 1995]. In einer Modellansicht werden Punkte mir einer hohen Krümmung (Ecken) detektiert und in einem Graphen als Knoten eingetragen. Die Kanten werden durch Konturen mit geringer Krümmung gebildet. Dabei wird der Matchprozeß als eine sonst in der Stereobildverarbeitung üblichen Abbildung betrachtet (siehe Abbildung 3.10). Zur Steuerung der Abbildung von dem Modellgraphen auf einen Szenengraphen werden eine Reihe von *constraints* definiert:

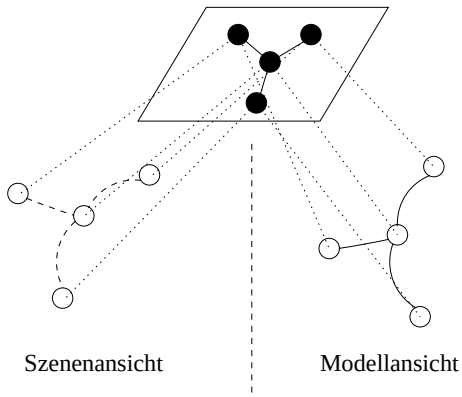


Abbildung 3.10: Matchprozeß nach [CHUNG 1995]

- *Topologie-constraints:* Die Abbildung soll topologieerhaltend sein, d.h. daß Modellgraphkanten gültige Szenengraphkanten werden. Dieses *constraint* wird zur Validierung einer Abbildung verwendet.

- *Starrheit-constraints:* Ein Merkmalspunkt eines starres Modellobjekts wird unter der Voraussetzung einer schwachen perspektivischen Projektion auf

die Epipolarlinie der Szene abgebildet. Für ein Modellpunkt $\vec{p}$ und dem korrespondierenden Szenenpunkt $\vec{p}$ gilt dann: $\vec{a} \cdot \vec{p} + \vec{b} \cdot \vec{p} + c = 0$ (für alle $\vec{a}$, $\vec{b}$ und c). Für praktische Anwendungen wird anstelle des Vergleiches mit 0 die ϵ-Umgebung getestet. Es kann gezeigt werden, daß nur vier Punktkorrespondenzen zur Bestimmung der epipolaren Geometrie zwischen zwei Objektansichten ausreichen.

- *Flachheit-constraints:* Bei drei gegebenen Punktkorrespondenzen $\{(\vec{p}_i, \vec{p}_i) : i = 1, 2, 3\}$, wobei diese Punkte in einer Ebene liegen und nicht kollinear sein dürfen, zwischen Modell und Szene kann zu einem Modellpunkt $\vec{p}$ der korrespondierende Szenenpunkt $\vec{p}$ in baryzentrischen Koordinaten angegeben werden als: $\vec{p} = \alpha\vec{p_1} + \beta\vec{p_2} + \gamma\vec{p_3}$. Die Koordinaten α, β und γ ergeben sich durch ein lineares Gleichungssystem: $\vec{p} = \alpha\vec{p}_1 + \beta\vec{p}_2 + \gamma\vec{p}_3$ und $\alpha + \beta + \gamma = 1$.

Zur effektiven Anwendung der *constraints* müssen gewisse Anforderungen an die zu verarbeitenden Objekte gestellt werden: Sie müssen starr und stückweise glatt sein; die Oberflächen dürfen nur eine geringe Krümmung besitzen. Der Autor bezeichnet solche Objekte als “Quasipolyeder”. Zunächst werden die Knoten nach dem Starr- und Flachheits-*constraint* gematcht. Dazu wird der Modell- und Szenengraph in Teilgraphen zu je vier Knoten zerlegt und diese Teilgraphen nach ihrer “Verschiedenheit” geordnet. Jetzt erfolgt ein lineares Durchsuchen der geordneten Menge nach bester Übereinstimmung. Darauf aufbauend werden Korrespondenzen zwischen den Graphenkanten hergestellt und mittels des Topologie-*constraints* überprüft. Es wird der Szenengraph als eine im Modellraum nach Kanten suchende dynamische Kontur betrachtet. Die Ecken sind dabei die festen Bezugspunkte. Die Bewegung der Kanten besitzt hierbei wegen der Epipolareigenschaften nur einen Freiheitsgrad. Entsprechend der maximalen Übereinstimmung zwischen Modell und Szene bei minimaler Bewegung der Kanten wird über die Klassifikation entschieden. Der Ansatz zeichnet sich durch hohe mathematische Absicherung aus. Auch konnten an einem Beispiel in einer realen Szene gute Ergebnisse gezeigt werden. Problematisch bleibt die effektive Bestimmung der nötigen initialen vier Punktekorrespondenzen und die tolerierbaren Szenentransformationen.

In [BURGER et al. 1996] werden Graphen eingesetzt, um den strukturellen Aufbau von Objekten zu beschreiben. Die Autoren betonen die besondere Eignung struktureller Beschreibungen bezüglich deren Robustheit gegenüber veränderlichen Beleuchtungs- und Betrachtungsbedingungen. Das Erkennungssystem stützt sich auf die Arbeiten von Bischof und Caelli, u.a. vorgestellt in [BISCHOF und CAELLI 1997], und arbeitet nach folgenden Prinzipien:

- Strukturelle Informationen werden nicht aus einzelnen Merkmalspunkten abgeleitet, sondern aus der Gesamtheit des Bildes nach einer Wavelet-Filterung.
- Die strukturellen Primitive müssen nicht präzise lokalisiert sein, d.h. nur deren grobe räumliche Position und Formeigenschaften sind wesentlich.
- Die Primitive müssen nicht mit Objektteilen korrespondieren, die ein menschlicher Beobachter als typisch charakterisieren würde.

Innerhalb der Vorverarbeitung wird das Bild einer Wavelet-Filterung unterzogen. Als Ergebnis entstehen 32-elementige Vektoren, die mittels Vektorquantisierung in 64 Strukturklassen klassifiziert werden. Basis der Repräsentation sind *Part Adjacency Graphen PAG* und *Part Compatibility Graphen PCG*. Erstere repräsentieren Primitive und deren räumliche Relation zueinander. Dabei werden nur Kanten zwischen benachbarten Knoten bzw. zwischen Knoten mit einer bestimmten maximalen Distanz berücksichtigt, um einen vollständig vernetzten Graphen zu vermeiden. Innerhalb des *PCG* werden Primitive und deren Ähnlichkeit kodiert, d.h. es sind nur Knoten mit einem hohen Ähnlichkeitsmaß über Kanten verbunden. Zu beachten ist hierbei die u.U. hohe Kardinalität der sich dabei ergebenden Kantenmenge. Innerhalb der Lern-Phase werden alle unären Merkmale aller Ansichten und aller Objektklassen in Cluster zerlegt und dadurch ein Entscheidungsbaum aufgebaut. Ergeben sich hierbei nicht eindeutige Clusterzuordnungen, werden binäre Merkmale zwischen diesen Primitiven berechnet und diese erneut einer Clusteranalyse unterzogen. Diese Zerlegung und Klassifizierung wird fortgesetzt, bis alle Cluster eindeutige Zuordnungen beinhalten oder der Entscheidungsbaum eine bestimmte Tiefe überschritten hat. Während der Erkennungsphase wird zunächst ein *PCG* aufgebaut und in kleinere, aus 3 bis 5 Knoten bestehende, azyklische Teilgraphen zerlegt. Pro Objekt ergeben sich so etwa 50-100 Teilgraphen. Diese werden dann entsprechend dem aufgebauten Entscheidungsbaum klassifiziert. Jede Klasse erhält einfach ein Zähler, der deren Auftrittswahrscheinlichkeit repräsentiert. Die Leistungsfähigkeit des Systems wird anhand der Unterscheidung künstlich generierter Instanzen von Stühlen und Tischen demonstriert. Allerdings finden sich keine quantitativen Ergebnisaussagen. Auch wird die Segmentierung der Bilder sowie eine Objekt-Hintergrund-Trennung per Hand vorgenommen.

In [LOURENS und WÜRTZ 1998] wird ein Objekterkennungssystem basierend auf einer optimierten Graphensuche vorgestellt. Ziel war das Verhindern einer kombinatorischen Explosion und einer möglichst invarianten Erkennung. Sowohl die Szene als auch die Modelle sind durch Graphen beschrieben. Graphknoten werden an Szenenecken plaziert, Graphenkanten an den Verbindungslinien

zwischen Ecken. Die Knoten besitzen als kennzeichnendes Merkmal den Winkel, den die Ecke einschloß, die Kanten die relative (auf die längste vorkommende Kante normierte) Länge. Bei der Graphensuche werden ständig die maximalen und mittleren Streuwerte der sich ergebenden Winkel- und Längendifferenzen kontrolliert und mit vorgegebenen Parametern verglichen. Darauf aufbauend wird entschieden, ob ein konkreter Suchpfad im Korrespondenzbaum nicht weiter verfolgt wird. Probleme mit im Szenenbild u.U. nicht vorhandenen Merkmalen werden durch das Einfügen von Graphenknoten mit entsprechenden Kosten abgefangen. Mit diesem Ansatz konnten gute Ergebnisse für die Erkennung von Büromaterial (Textmarker) erzielt werden. Probleme bereiten dem Ansatz gekrümmte Objekte bzw. Objektteile, da durch die Konzentration auf Eckenmerkmale diese Szenenbereiche nicht analysiert werden können. Ebenso ist Voraussetzung, daß das Modell sehr genau spezifiziert wird, da unterschiedliche Winkel- bzw. Größenänderungen pro Modell nicht toleriert werden. Die Vorgabe von maximalen globalen Änderungen schränkt hier die Invarianzleistungen des Systems ein.

3.3.3 Hashverfahren

Hashverfahren benutzen eine Hashtabelle zum Abspeichern von Informationen. Indiziert wird diese Tabelle durch einen Schlüssel, der durch eine gewählte Hashfunktion aus der zu speichernden Information abgeleitet wird. Dabei kann es zu Kollisionen kommen, d.h. mehrere Informationselemente können den gleichen Index für die Tabelle erzeugen. Hierbei sind verschiedene Auflösungsmöglichkeiten bekannt: Verwendung linear verketteter Listen pro Hasheintrag oder alternative Hashfunktionen, die andere Indizes erzeugen. Hashverfahren sind gute Methoden zum Auffinden von gespeicherten Informationen. Sie stellen einen Kompromiß zwischen Geschwindigkeit und Speicherbedarf dar.

In [BELLAIRE 1995] wird ein Erkennungssystem basierend auf der Hash-Technik vorgestellt. Es wird ein Objekt-Graph aus gegebenen CAD-Daten generiert. Aus diesem Graphen werden Subgraphen bestehend aus drei Knoten und zwei Kanten extrahiert, die dann die Basis für die Berechnung des Hash-Wertes bilden. Es werden fünfdimensionale Merkmalsvektoren aus der Lage der Kanten und Knoten zueinander abgeleitet, wobei eine Klassifizierung nach den Berührungspunkten der Kanten durchgeführt wird. Der Algorithmus wird in 5 Stufen unterteilt, wobei zwischen einem einmaligen Verarbeiten der Modelle in einer *offline*-Phase und der *online*-Phase als der eigentlichen Erkennung unterschieden wird:

1. Die Generierung der Modellgraphen, die Extraktion der Subgraphen und die Berechnung der Merkmalsvektoren. (*offline*)

2. Aufbau der Hashtabelle für die Modellansichten. Jeder Ansicht wird ein Hash-Akkumulator zugeordnet. Durch jeden Merkmalsvektor einer Ansicht wird eine Verbindung zu der betreffenden Ansicht hergestellt. (*offline*)

3. Merkmalsextraktion aus der Szene (hier Kantengenerierung) und Transformation in fünfdimensionalen Merkmalsvektor, wobei dieser Vektor noch gewichtet werden kann. Die Vektoren erhöhen den Wert einer spezifischen Hash-Akkumulator-Zelle. (*online*)

4. Ableitung der 16 wahrscheinlichsten Hypothesen durch die maximalen Einträge in den Akkumulatoren. Diese Hypothesen werden anschließend gewichtet nach der Summe der Gewichte der eingeordneten Merkmale, nach der Anzahl der Akkumulatoreinträge bei gleichen Gewichtssummen und dem relativen Anteil von fehlenden Modellkanten in Bezug zu gelernten Ansichten. (*online*)

5. Löschen von erkannten Objektansichten aus der Hashtabelle und weiter mit Schritt 4. Somit können auch evtl. verdeckte Objekte erkannt werden. (*online*)

In [BELLAIRE und LÜBBE 1995] werden konkretere Angaben zu den Wichtungen der Merkmalsvektoren gemacht: entscheidend ist die euklidische Distanz der Szenenmerkmalsvektoren zu den Modellvektoren, die entsprechend einer Gaußfunktion gewichtet wird. Die Verteilung der Funktion kann während der Laufzeit adaptiert werden. Weiterhin wird die Menge der Objektansichten durch statistische Verfahren minimiert. In [BELLAIRE et al. 1996] wird das Verfahren zur Müllsortierung eingesetzt und um die *Photometric Stereo Method* nach [WOODHAM 1980] zur Oberflächenrekonstruktion erweitert. Störende Oberflächenreflexionen werden durch das *Dicromatic Reflection Model* nach [SHAFER 1985] beseitigt.

Ein sehr ähnlicher Ansatz zu den oben dargestellten Arbeiten wird in [COSTA und SHAPIRO 1996] vorgestellt, der ebenfalls auf der Anwendung von Hash-Tabellen beruht. Es wird ebenso von einer Teilung eines kompletten Objektgraphen in Subgraphen ausgegangen. Parallelen zeigen sich auch bei der Unterteilung in eine *offline* und eine *online*-Phase. Es wird ebenfalls ein Akkumulator pro übereinstimmenden Hash-Index erhöht und entsprechend das Modell mit den maximalen Akkumulatoreinträgen weiter untersucht. Hier wird allerdings ein Subgraph für die Szene bestehend aus zwei Knoten und einer verbindenden Kante betrachtet. Ein weiterer Unterschied findet sich in den verwendeten Merkmalen und Relationen. Modellobjekte werden in geometrische Primitive wie Ellipsen und Linien zerlegt und in Beziehungen wie “schließt ein” und “parallel zu” gesetzt.

Das Verfahren wird demonstriert anhand der Erkennung industrieller Teile unter verschiedenen Beleuchtungsbedingungen. Es wird angeregt, die Hash-Ergebnisse durch Nutzung bestimmter Merkmale zu einer Lageschätzung des präsentierten Objekts zu benutzen.

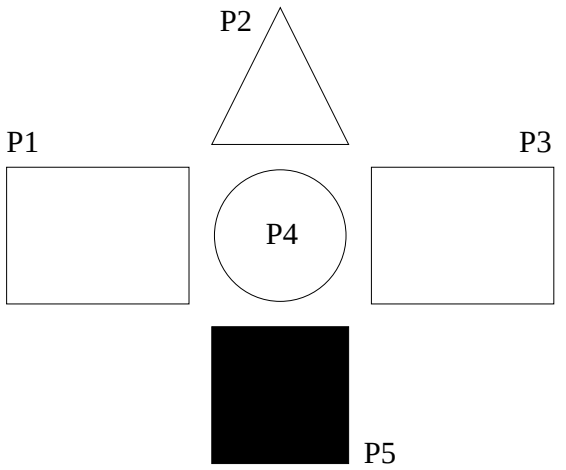
3.3.4 Strukturbeschreibungen

Strukturelle Beschreibungen dienen einer qualitativen, expliziten Repräsentation von Wissen über Objekte. Die Argumente dieser Datenstrukturen korrespondieren zu lokalen Merkmalen oder Komponenten der Objekte und die Prädikate zu deren Eigenschaften bzw. räumlichen Relationen. Vorteil bezüglich einer 3D-Objekterkennung ist vor allem, daß die Beschreibungen bei kleineren Wechseln des Betrachtungspunktes konstant bleiben. Als Repräsentationsform werden häufig Graphen (siehe auch Kapitel 3.3.2) oder neuronale Strukturen verwendet.

In [SHAPIRO und HARALICK 1981] wird ein Verfahren zur Repräsentation und zum Matchen von Objekten unter Nutzung von geometrischen Strukturbeschreibungen vorgestellt. Ein Objekt D wird durch das Paar $D = (P, R)$ beschrieben, wobei P die Menge der Primitive $P = \{P1, \cdots, Pn\}$ ist. Jedes Primitiv Pi steht für die binäre Relation $Pi \subseteq A \times V$ mit A der Menge der Attribute und V der Menge der möglichen Werte. $R = \{PR1, \cdots, PRK\}$ bezeichnet die Menge von benannten Nären Relationen über P. Für jedes $k = 1, \cdots, K$ ist $PRk = (NRk, Rk)$, wobei NRk der Name für die Relation Rk ist. Abbildung 3.11 zeigt ein Beispielobjekt und die dazugehörige, z.T. unvollständige, strukturelle Beschreibung.

Zum Matchen der strukturellen Beschreibungen von Modell- und Szenenobjekt kommt ein inexaktes Matchverfahren zum Einsatz, da die Generierung der Strukturbeschreibungen realer Bildobjekte immer fehlerbehaftet ist. Der Algorithmus verwendet ein Ähnlichkeitsmaß aus der Wahrscheinlichkeit, daß eine durch den Match festgestellte Veränderung der Modellbeschreibung wirklich stattfand und der Wahrscheinlichkeit, daß der Match nur zufällig zustandekam.

Zur Beurteilung der Matchgüte für Objektteile wird ein einfaches Distanzkriterium eingesetzt. Dieses Kriterium wird für jedes Attribut einzeln festgelegt und besteht in einem Schwellwert, bis zu dem die Attributswerte für matchende Objektteile differieren dürfen. Zur Berücksichtigung der unterschiedlichen Bedeutung von Objektteilen, wird die Beschreibung der Objekte um die Funktion wp mit $wp : \ P \rightarrow [0, 1]$ und $\sum_i wp(P_i) = 1$ erweitert, die allen Primitiven eine Wichtung zuordnet. Die Relationen R erhalten jeweils auch eine Menge von Wichtungsfunktionen WR mit $WR = \{w1, w2, \cdots, wK\}$. Für alle k mit $k = 1, \cdots, K$ weist wk mit $wk : \ Rk \rightarrow [0, 1]$ und $\sum_{k \in Rk} wk(r) = 1$ den Tupeln der Relation Rk Gewichte zu. Unter Verwendung der Gewichte wird nun die



a)

$$
\begin{array}{lcl}
Dp & = & \{P, RP\} \\
P & = & \{P1, P2, P3, P4, P5\} \\
RP & = & \{(\text{Left}, \text{Left\_}P), (\text{Above}, \text{Above\_}P)\} \\
\text{Left\_}P & = & \{(P1, P4), (P4, P3)\} \\
\text{Above\_}P & = & \{(P2, P4), (P4, P5)\} \\
P1 & = & \{(\text{shape}, \text{rectangular}), (\text{color}, \text{white})\} \\
P2 & = & \{(\text{shape}, \text{triangular})\} \\
P3 & = & \{(\text{shape}, \text{rectangular})\} \\
P4 & = & \{(\text{shape}, \text{circular})\} \\
P5 & = & \{(\text{color}, \text{black})\}
\end{array}
$$

b)

Abbildung 3.11: *a)* ein Modellobjekt *Dp*; *b)* strukturelle Beschreibung des Objekts *Dp* (nach [SHAPIRO und HARALICK 1981])

Matchbedingung, der ϵ-Homomorphismus, definiert:

$$\sum_{\substack{r \in R \\ h(r) \notin S}} w(r) \;\leq\; \epsilon \tag{3.12}$$

R ist wieder die Näre Relation über P, $w : \; R \rightarrow [0, 1]$ ist die Wichtungsfunktion für R und S ist die Näre Relation über eine zweite Primitivmenge Q. Mit h als eine Abbildung $h : \; P \rightarrow Q$ matcht das N-Tupel r aus R auf S, wenn $h(r)$ ein Element von S ist. Somit wird die Summe der Wichtungen für alle nicht matchenden Elemente bezüglich ϵ abgetestet. Zum konsistenten Labeln und

somit zur Generierung eines relationalen Homomorphismus wird ein Baum-Such-Algorithmus unter Verwendung von *look-ahead* und *forward checking* Operatoren eingesetzt.

Das Verfahren wird anhand statistisch generierter Strukturbeschreibungen evaluiert. Die besten Ergebnisse sind bei der Verwendung des *forward checking* Operator zu beobachten. Abschließend muß jedoch festgestellt werden, daß die Generierung von symbolischen Strukturbeschreibungen für reale bzw. komplexe Objekte nur schwer durchführbar ist. Somit ist die Datengrundlage des Suchalgorithmus u.U. sehr fehlerbehaftet und ungenau. Prinzipiell sollte jedoch die Kapazität des Suchverfahrens bei gegebenen strukturellen Objektbeschreibungen Berücksichtigung finden.

In [SHAPIRO et al. 1984] wird die Idee der relationalen Beschreibungen weiter ausgebaut. Hier sollen diese Modelle zu einer groben Modellierung von dreidimensionalen Objekten benutzt werden. Die groben Objektbeschreibungen werden bei einem initialen Match herangezogen und als Basis für einen Clusteralgorithmus verwendet, der ähnliche Objekte kombiniert. Auf hierarchisch tieferen Ebenen erfolgt dann eine präzisere Beschreibung mit mehr Details. Alle Objekte werden in folgende Primitive zerlegt:

- *sticks:* Diese Objektteile sind stangenähnliche, lange, dünne Teile mit nur einer signifikanten Dimension. Zur Vereinfachung werden alle *sticks* als Liniensegmente repräsentiert.

- *plates:* Flache, ausgedehnte Objektteile mit annähernd flachen Oberflächen werden als *plates* bezeichnet. Diese haben zwei signifikante Dimensionen und werden vereinfacht als Kreise repräsentiert.

- *blobs:* Mit *blobs* werden Objektteile mit drei signifikanten Dimensionen bezeichnet und durch Kugeln angenähert.

Alle Objektteile sind in ihrer Grundform annähernd konvex. Zum Zusammenfügen der Objektteile zu einem kompletten dreidimensionalen Modell werden eine Reihe von *constraints* definiert, die sowohl die Verbindungsstrukturen zwischen den Teilen als auch inhärente Objekteigenschaften spezifizieren:

- *binäre Verbindungen:* Die direkten Verbindungen zwischen zwei Objektteilen ist integraler Bestandteil dreidimensionaler Objektmodelle. Eine Verbindung wird erstens durch den Verbindungstyp zur Angabe der in Verbindung stehenden Elemente (Ende-Ende, Ende-Mitte, Kante-Kanten, $\cdots$) und zweitens durch einen Verbindungswinkel, angegeben durch konkrete Werte oder Intervalle, beschrieben.

- *ternäre Verbindungen:* Binäre Relationen reichen zur vollständigen Beschreibung dreidimensionaler Strukturen nicht aus. Zur exakten Beschreibung müßten gar Näre Relationen für beliebige N eingesetzt werden. Es wird sich aber auf den Einsatz von ternären Relationen beschränkt. Solche Relationen werden durch das Tripel $(s1, s2, s3)$ beschrieben, wobei $s1$ und $s3$ $s2$ berühren. Die räumliche Lage von $s1$ und $s3$ in Bezug zu $s2$ wird durch zwei Elemente angegeben. Das erste gibt an, ob $s1$ und $s3$ $s2$ am selben Ende bzw. an der selben Fläche berühren. Die zweite Komponente gibt den Winkel an, der sich durch die Verbindung der Schwerpunkte von $s1$, $s2$ und $s3$ ergibt.

- *Support-Struktur:* Hierbei wird Bezug auf die Funktion von Elementen eines Objekts genommen.

- *zusätzliche Constraints:* Weitergehende Informationen sind zur Objektbeschreibung notwendig, wenn sich die Objektteile z.B. nicht berühren. Aber auch globale Eigenschaften von Objektteilen, wie z.B. deren Längenbeziehung, können hier angegeben werden.

Während der Matchphase müssen die generierten relationalen Beschreibungen gegen eine Objektdatenbank gematcht werden. Zur Beschleunigung des Systemverhaltens wird hier nicht gegen jedes Modellobjekt gematcht, sondern zunächst gegen mehrere Objekte beschreibende *profile*-Daten, die durch Clusterung gewonnen werden. Anschließend werden die relationalen Daten gegen alle Objekte im matchenden Cluster verglichen. Als Clusterkriterium werden die euklidischen Distanzen der Attributwerte aller möglichen Objektpaare gebildet. Die Attribute werden dabei noch unterschiedlich gewichtet. Folgende Attribute werden eingesetzt:

- Anzahl aufrechter Teile (Wichtung 1)
- Anzahl horizontaler Teile (Wichtung 1)
- Anzahl schräger Teile (Wichtung 1)
- Anzahl *sticks* (Wichtung 10)
- Anzahl *plates* (Wichtung 10)
- Anzahl *blobs* (Wichtung 10)
- Anzahl Ebenen (Wichtung 10)
- Typ des oberen Objektteils (Wichtung 10)

- Lage des oberen Objektteils (Wichtung 10)
- Anzahl von Teilen, die das Basisteil stützen (Wichtung 10)

Beim Lernen eines neuen Objekts wird die Attributwert-Tabelle des Objekts mit denen der Clusterzentren verglichen und das neue Objekt dem ähnlichsten Cluster zugeordnet. Analog werden unbekannte Objekte klassifiziert und das Ergebnis als Hypothese für ein anschließendes Baumsuchverfahren mit einem *look ahead*-Operator verwendet. Zur Beurteilung des Matchergebnisses wird ein struktureller Fehler und ein Vollständigkeitsfehler berechnet, wobei dem strukturellen Fehler höhere Priorität eingeräumt wird.

Das Verfahren wird mit einer Möbel-Datenbasis getestet. Es ergeben sich durchweg gute Ergebnisse. Die grundsätzliche Idee, den Matchprozeß zunächst durch ein Clusterverfahren basierend auf globale Objekteigenschaften einzuleiten, ist ein gangbarer Weg, um Suchzeitenprobleme des Baumsuchverfahrens auszuräumen. Interessanterweise wird aber in einer späteren Arbeit ([COSTA und SHAPIRO 1996], beschrieben im Kapitel 3.3.3) von dem Clusterverfahren zugunsten eines Hashverfahrens wieder abgegangen. Problematisch scheint aber auch hier die Komplexität der Objektbeschreibungen. Die dargestellten Attribute und Relationen verlangen eine optimale Bildvorverarbeitung und sind ohne Nutzereingriff nur schwer aufzubauen.

In [ZHANG et al. 1992] wird sich der Frage nach den je nach Objektansicht sichtbaren Merkmalen im Rahmen eines Objekterkennungssystems gewidmet. Dabei wird ein ansichtenunabhängiges relationales Objektmodell verwendet. Dieses wird aus einem CAD-Modell abgeleitet. Knoten innerhalb dieser Graphbeschreibung werden geometrischen Merkmalen zugeordnet. Graphkanten bestimmen die Sichtbarkeit der Merkmale bei verschiedenen Ansichten und sind mit einer Wahrscheinlichkeit bewertet, daß die mit der Kante verbundenen Merkmale gleichzeitig sichtbar sind. Kanten werden weiterhin mit einer Merkmalsbeschreibung bewertet. Dazu werden Methoden vorgestellt, um qualitative Merkmale wie Parallelität, Kollinearität, Lagebeziehungen und relative Größe durch Skalare auszudrücken. Mit Hilfe des Graphmodells werden Hypothesen bezüglich der präsentierten Objekte und deren Lage abgeleitet. Nach Anwendung des Canny-Kantendetektors werden die Kantenmerkmale berechnet. Das Matchen mit dem Graphen erfolgt mit einer Tiefe-Zuerst-Suche, wobei der Suchprozeß durch die Bewertungen der Merkmals-Sichtbarkeit eingeschränkt wird. Nach der Hypothesengenerierung erfolgt eine Schätzung der Objektlage durch die Auswertung matchender, nicht paralleler, nicht kollinearer, aber koplanarer Liniensegmente. Das Verfahren wird anhand der Erkennung und Lageschätzung von PKWs in normalen Umweltbedingungen erfolgreich getestet. Das Matchen der Graphen resultiert immer in mehreren Hypothesen, womit sich ebenfalls mehrere mögliche

Lageschätzungen ergeben. Die Auswahl der “richtigen” Schätzung wird von den Autoren nicht weiter behandelt, stellt aber den kritischen Punkt im gesamten Erkennungssystem dar.

In [SHAMS 1995] wird ein neuronal orientiertes Erkennungssystem vorgestellt, das Objektmerkmale direkt in den Typen von Neuronen repräsentiert. Die Neurone werden dabei in einem Graphen angeordnet, dessen Knoten die Merkmale repräsentieren und die Graphenkanten die relative räumliche Lage der Merkmale angibt. In Abbildung 3.12 ist ein Beispiel angegeben. Das dreieckförmige Objekt wird durch vier Neuronen in dem dargestellten Graph angegeben. Jedes Neuron repräsentiert dabei ein bestimmtes Merkmal (hier: orientierte Kantenelemente und eine Textur).

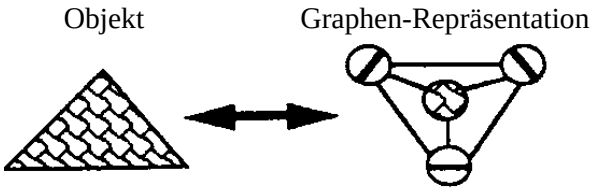


Abbildung 3.12: Prinzip der Objektrepräsentation (nach [SHAMS 1995])

Als Basismerkmale kommen Kanten verschiedener Orientierung und Größe zum Einsatz. Dazu wird das Szenenbild mit 12 unterschiedlich orientierten Gaborfiltern auf je zwei Skalierungsebenen bearbeitet. Dadurch ergeben sich insgesamt 24-dimensionale Merkmalsvektoren pro Bildpunkt. Die angeregte Verwendung von Texturen soll perspektivisch ebenfalls aus den Gabor-Filterantworten generiert werden. Das präsentierte System verwendet somit zwei Merkmalstypen, die durch zwei Neuronentypen repräsentiert werden:

- lokal arbeitende Neurone mit kleinen rezeptiven Feldern für kleine orientierte Kantenelemente und
- Neurone mit großen rezeptiven Feldern für große orientierte Kanten, die global agieren

Die Neurone können jeweils nur ein Merkmal innerhalb ihres rezeptiven Feldes speichern. Können an einer Bildposition mehrere Merkmale detektiert werden, so werden dort auch mehrere Neuronen mit einer räumlichen Entfernung von 0 generiert. Zur Ausbildung der Modellnetze werden CAD-Linienzeichnungen verwendet, auf denen heuristisch Neurone gleichmäßig verteilt werden. Angeregt

wird zwar eine selbständige Ausbildung der Modelle auf der Basis einer Mehrfachrepräsentation, allerdings kommt diese Strategie in den Experimenten nicht zum Einsatz.

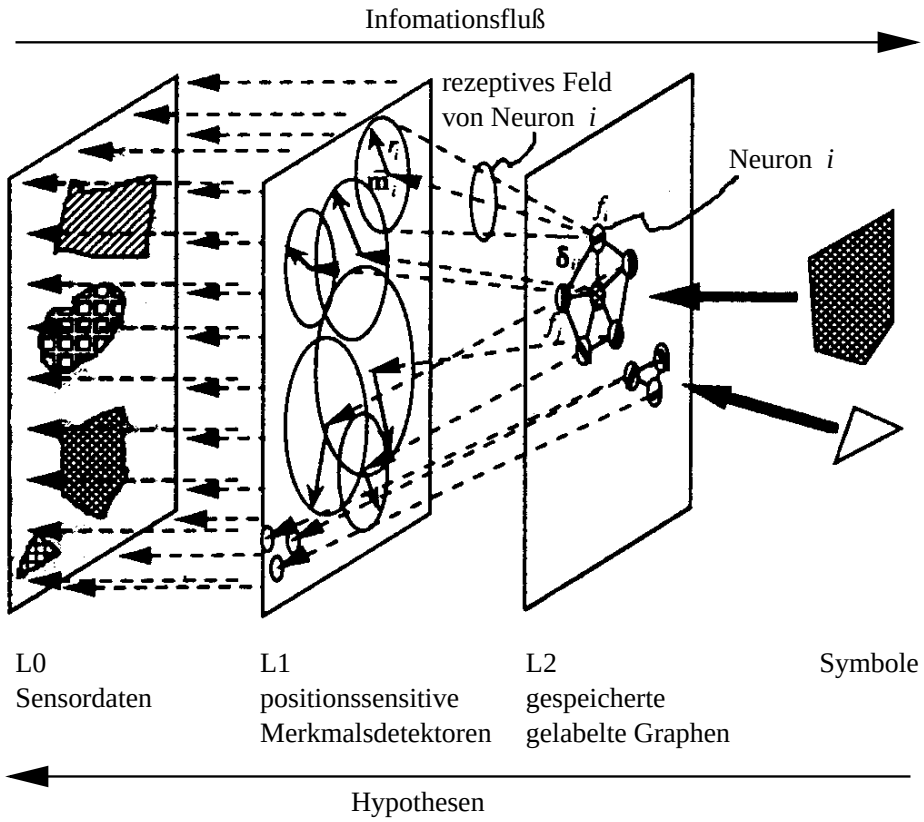


Abbildung 3.13: Hierarchische Objektrepräsentation (nach [SHAMS 1995])

In Abbildung 3.13 ist das Gesamtsystem mit den drei Verarbeitungsebenen dargestellt:

- *L0:* Hier wird das Eingabebild abgelegt.
- *L1:* Die Ergebnisse der Bildvorverarbeitung mit den Gabor-Filtern werden hier repräsentiert.
- *L2:* Dies ist die Ebene zur Aufnahme der Objektmodelle. Die Neurone i besitzen jeweils ein Merkmalsattribut f_i zur Kodierung des im rezeptiven Feld vorhandenen Merkmale. Durch einen Vektor δ_{ij} wird für zwei Neurone i und j deren räumliche relative Lage zueinander kodiert. Jedes Neuron i besitzt weiterhin drei weitere Attribute:

  - m_i: das Zentrum des rezeptiven Feldes,
  - r_i: der Radius des rezeptiven Feldes,
  - h_i: der Grad inwieweit das Neuron matchende Merkmale gefunden hat

Erkennungsrobustheit wird durch elastische Verbindungen zwischen Knoten bzw. Neuronen eingeführt. Während des Erkennungsvorganges wird ein Modellnetz an eine gegebene Szene so angepaßt, daß schließlich möglichst viele Neurone passende Szenenaktivitäten zu finden. Es wird eine lokale Dynamik der rezeptiven Felder in Abhängigkeit von h_i eingeführt, um Neurone an passenden Merkmalspositionen festzuhalten, wodurch sich auch der Suchbereich bei erfolgreichem Match immer weiter einschränkt, bzw. um im negativen Fall den Suchbereich weiter auszudehnen. Während des Matchens werden zufällig *L1*-Neurone der ausgewählt und die räumlich nächsten *L2*-Neurone gesucht, deren rezeptives Feld die *L1*-Neurone beinhaltet. Diese und deren Nachbarn werden in die Richtung der detektierten Merkmale bewegt.

Das Erkennungssystem wird demonstriert anhand der Erkennung eines Modelljeeps und eines Modellpanzers, wobei gute Ergebnisse erzielt werden konnten. Zu bemerken ist aber, daß beide Objekte sehr verschieden sind, wodurch wenig Verwechslungsgefahr untereinander besteht. Es konnte aber gezeigt werden, daß durch die Nutzung einfacher Kantenmerkmale und die Auswertung deren räumlicher Relationen eine stabile Objekterkennung möglich ist. Weiterhin wird eine Nutzung des System für Trackingaufgaben vorgeschlagen. Hierbei soll ausgenutzt werden, daß sich die Objekte von Bild zu Bild nur minimal ändern und somit auch nur minimale Änderungen der Modellneuronen nötig wird.

In [WILLIAMSON 1996b] wird sich intensiv der Frage einer geeigneten Repräsentationsform gewidmet. Es wird festgestellt, daß globale 3D-Strukturrelationen sich bei Blickrichtungsänderungen relativ konstant verhalten - im Gegensatz zu lokalen Merkmalen und deren räumliche Lage zueinander. Deshalb wird für eine ansichtenbasierte Objekterkennung die Kodierung von Merkmalen eingebettet in 3D-Strukturbeschreibungen vorgeschlagen. Zur Objektrepräsentation werden Graphen eingesetzt, wobei die Knoten lokale Merkmale und die Kanten deren räumliche Beziehung kodieren. Pro Knoten und Kante werden mehrere Neurone zur Merkmals- und Attributkodierung eingesetzt. Jeder Knoten kodiert mit je einem Neuron:

- die Position des Attributs (*P*-Neuron),
- die Art des Attributs (*FA*-Neuron),
- die Richtung der Verbindung zum einem benachbarten Knoten (*G*-Neuron) und

- Tiefeninformation (*D*-Neuron).

Ein Knoten kann auch mehrere *FA*-Neurone besitzen, die jeweils die unterschiedlichen Merkmale durch Templates kodieren. Das *FA*-Neuron mit bester Merkmalsübereinstimmung hemmt alle anderen für diesen Knoten. Pro Kante werden durch *L*-Neurone gespeichert:

- der Winkel und
- die Entfernung zwischen benachbarten Knoten,
- die Stärke der Verbindung, die in Abhängigkeit der Entfernung und der Übereinstimmung der *FA*-Neurone gesetzt wird, und
- die Richtung der Verbindung.

Jeweils eine Merkmalskarte für Konturen und Ecken bilden die Grundlage zur Bildung der Graphenrepräsentation. In Abbildung 3.14 wird die Herausbildung eines Objektgraphen demonstriert. Zum Aufbau der Objektgraphen konkurrieren sowohl die *P*-Neurone um Merkmalspositionen in den beiden Merkmalskarten als auch die Knoten-Verbindungen untereinander gemäß lokaler Merkmalsübereinstimmungen.

Die vorgestellte Objektrepräsentation wird von den Autoren außerdem zu einer Tiefenrekonstruktion eingesetzt. Basis dafür ist die Auswertung von Überdeckungen durch die Analyse “T”-förmiger Ecken. Die *D*-Neurone verwalten die so gewonnene Tiefeninformation.

Die Arbeit untersucht vor allem Fragen zur Objektrepräsentation. Die eigentliche Erkennung wird nur am Rande behandelt. So werden auch keine realen Objekte zu Demonstrationszwecken verwendet. Trotzdem überzeugt die gewählte Repräsentationsform sowie das Prinzip der Ausbildung durch selbstorganisierende Wettbewerbsprozesse vor allem auch wegen ihrer biologischen Plausibilität.

Ein Erkennungssystem basierend auf strukturellen Beschreibungen von Objekten, deren Prädikate die elementaren Objektteile und deren Argumente die räumlichen Beziehungen angeben, wird in [BENNAMOUN und BOASHASH 1997] erläutert. Basis ist die Zerlegung von Objekten in Objektteile, deren Annäherung mittels geometrischer Primitive und ein strukturelles Matchen von Modell- und Szenenobjekten. Hervorgehoben werden die Vorteile der strukturellen Beschreibungen für eine bezüglich Position, Orientierung, Größe und besonders fehlenden Objektteilen invariante Erkennung. Es lassen sich folgende Teilschritte unterscheiden:

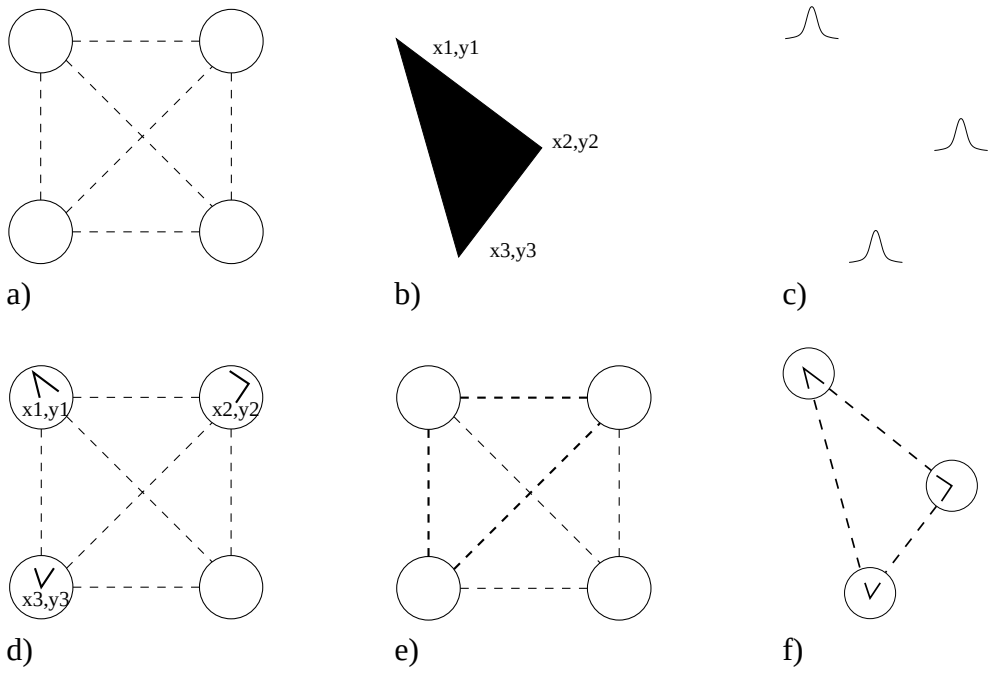


Abbildung 3.14: Ausbildung der Objektrepräsentation; *a)* vier nicht zugeordnete Knoten mit sechs Kanten ; *b)* Eingabebild; *c)* Merkmalsaktivitäten; *d)* drei Knoten werden den Maxima der Merkmalsaktivitäten zugeordnet, sie kodieren deren Position und lokale Formmerkmale; *e)* Graphenkanten werden aktiviert; *f)* die Knoten werden an die kodierten Positionen des Eingabebildes verschoben (nach [WILLIAMSON 1996b])

1. Kantenextraktion durch Kombination von Filtern erster und zweiter Ordnung. Einer möglichst optimalen Generierung von Objektkanten wurde große Bedeutung beigemessen, da die folgenden Algorithmen stark von “guten” Kanten abhängig sind. Die gefundene Lösung (siehe Abbildung 3.15) stellt ein Kompromiß bezüglich den konträren Anforderungen an Lokalität und Rauschunabhängigkeit der generierten Kanten dar. Der obere Verarbeitungszweig erzielt eine gute Lokalität der Kanten und der untere verringert das Rauschen. Beide Ergebnisse werden durch ein logisches UND verknüpft. Der starken Parameterabhängigkeit wird durch die Minimierung einer probabilistischen Kostenfunktion zur automatischen Festlegung aller Parameter begegnet. Als abschließende Verarbeitungsstufen findet ein *edge-closing* und eine Konturverfolgung statt.

2. Basierend auf *convex dominant points (CDP)* der Konturen findet eine Objektzerlegung statt. *CDPs* stellen Punkte mit einer hohen Krümmung ent-

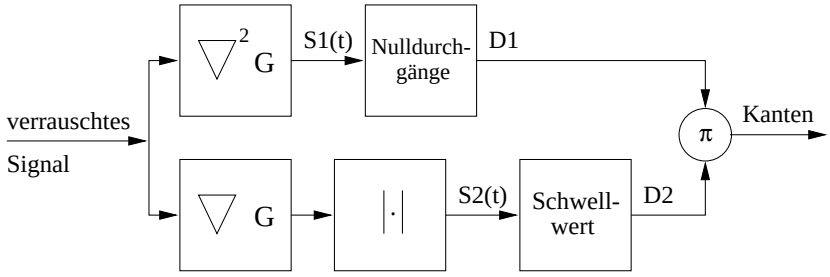


Abbildung 3.15: Kantendetektor nach [BENNAMOUN und BOASHASH 1997]

lang der Konturen dar. Diese Punkte werden entlang der Normalen bewegt, bis sie auf einen anderen sich bewegenden *CDP* oder auf eine Kontur treffen. Die Verbindung zwischen Anfangs- und Endposition stellt die Trennlinie für ein Objektteil dar.

3. Ein Abfahren der Begrenzungen aller Objektteile liefert die gesamte Zerlegung des Objekts in seine wesentlichen Teile.

4. Die isolierten Teile werden durch 2D-Superquadrics beschrieben. Durch diese parametrisierbaren Formen lassen sich mit Hilfe weniger Parameter die grundlegenden geometrischen Primitive generieren:

$$\vec{X}(\theta) = \begin{pmatrix} a_1 \cos\theta^{\varepsilon} \\ a_2 \sin\theta^{\varepsilon} \end{pmatrix} \quad \text{mit} \quad \begin{array}{rl} \vec{X}(\theta) & \text{Ortsvektor der Kontur} \\ \theta & \text{Orientierungswinkel} \\ \varepsilon & \text{Formparameter} \\ a_1, a_2 & \text{Größenparamter} \end{array}$$

Durch einen iterativen Minimierungsalgorithmus werden die Parameter der Superquadrics bestimmt und damit an die Objektformen angepaßt. Die Schwerpunkte und die räumliche Beziehung von den generierten Superquadrics bilden dann das Skelett des Objekts - die strukturelle Beschreibung (siehe Abbildung 3.16). Aus den Beschreibungen verschiedener Objekte wird schließlich eine Objektdatenbank aufgebaut.

5. Der Erkennungsvorgang besteht zunächst in der Suche nach dem Hauptobjektteil *MOP*, welches als das mit den meisten Nachbarn bzw. mit der größten Fläche angenommen wird. Als Matchmaß werden die Winkeldifferenzen zwischen den Objektskeletten des gelernten und präsentierten Objekts verglichen (siehe Abbildung 3.17). Daraus ergibt sich eine Abhängigkeit

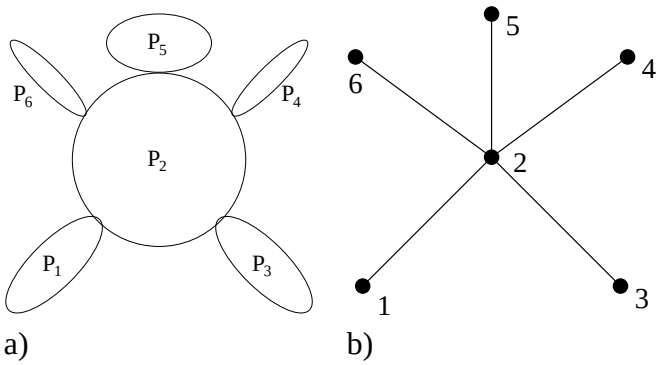


Abbildung 3.16: *a)* Zweidimensionale Superquadrics-Repräsentation; *b)* strukturelle Repräsentation (nach [BENNAMOUN und BOASHASH 1997])

von dem gewählten Referenzpunkt, d.h. je nach gewähltem initialem Match zwischen zwei Objektteilen ergeben sich verschiedene Winkeldifferenzen. Dementsprechend wird das Modellobjekt unter allen möglichen Rotationen gematcht - die minimalen Fehlerwinkel ergeben dann die Objektrotation. Die Größe des Objekts wird aus dem Größenverhältnis der *MOPs* abgeleitet.

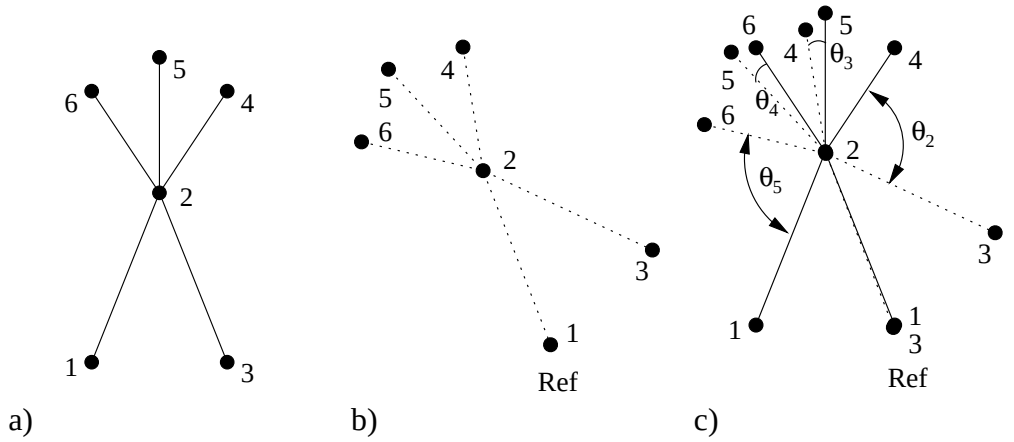


Abbildung 3.17: *a)* Modellstruktur; *b)* Rotation der Modellstruktur (Referenz 1-3); *c)* Berechnung der Fehlerwinkel (nach [BENNAMOUN und BOASHASH 1997])

Der Ansatz zeichnet sich durch die strikte Anwendung einfacher, wenig komplexer Algorithmen aus. Mit den angegebenen Beispielobjekten konnte eine 100% Erkennungsrate erzielt werden. Die Objekte werden immer zufriedenstellend zerlegt und durch Superquadrics beschrieben. Zum Einsatz kommen Spielzeug-

Objekte (Puppen, Autos), die allerdings untereinander nur eine geringe Ähnlichkeit besitzen. Hervorgehoben wird außerdem die Möglichkeit, das System auch zur Bildkomprimierung einzusetzen: nur die Strukturparameter sollen zur Beschreibung und Übertragung von Bildern ausreichen. Die Handhabung von Verdeckungen und Objektveränderungen ist durch die verwendete Objektrepräsentation ebenfalls flexibel. Wissen wird hier explizit innerhalb der Repräsentation und Erkennung genutzt, wodurch das System einfacher modifiziert und erweitert werden kann. Problematisch erscheint die starke Abhängigkeit von perfekten Objektkanten. Entsprechend werden die Testobjekte vor unstrukturiertem Hintergrund präsentiert. Weiterhin erscheint die Auswahl des Bezugsobjektteils (*MOP*) nach der Anzahl der vorhandenen Nachbarn fragwürdig. Abschließend kann gesagt werden, daß ein komplettes Objekterkennungssystem auf der Basis einer sehr einfachen, wenig komplexen Beschreibung realisiert werden konnte.

In [FISHER und MACKIRDY 1998] wird ein Objekterkennungssystem vorgestellt, das sich der Frage widmet, ob die Leistungsfähigkeit der Erkennung erhöht wird, wenn man zu den klassischen merkmalsbasierten Korrelationsverfahren noch eine strukturelle Beschreibung der Objekte hinzufügt. Grundidee ist, daß einzelne Merkmalslokalisierungen durch die Analyse benachbarten Merkmale stabiler werden. Das System wird aus zwei Teilkomponenten aufgebaut. Zunächst wird ein klassischer merkmalsbasierter Ansatz vorgestellt. Die zu untersuchenden Bilder werden log-polar-transformiert, um einfacher Skalierungen und Rotationen tolerieren zu können. Pro Bildpunkt werden 42 Merkmalsbilder generiert - gekennzeichnet durch drei Skalierungsebenen mit den folgenden 14 Merkmalen:

- rote, grüne und blaue Farbkanalwerte,
- zwei Merkmale mit *on-*, *off-center*-Verhalten, d.h. es werden sowohl schwarze als auch weiße einzelne Punkte einbezogen,
- vier radiale und orthogonale Balken mit *on-*, *off-center*-Verhalten, d.h. schwarze und weiße Kanten,
- vier Orientierungen für Kanten und
- ein Eckenmerkmal

Die Modelle sind durch diese Merkmalsbilder gekennzeichnet, wobei noch jedem Merkmal eine Wichtung über dessen Diskriminierungseigenschaften zugewiesen wird. Zur Erkennung wird pro Modell ein einfaches Korrelationsverfahren eingesetzt. In der zweiten Teilkomponente des Systems werden verschiedene Modelle mit je 42 Merkmalsbildern in räumliche Beziehungen zueinander gesetzt.

Hierbei wird eine symbolische Notation gewählt. Entsprechend wird ein Modell für Gesichter entworfen, das die Teilmodelle Augen, Nase, Mund und das ganzes Gesicht beinhaltet. Durch die Auswertung der räumlichen Lage der einzelnen Teilmodelle konnten die Matchergebnisse verbessert werden, wobei jedoch zu bemerken ist, daß die gestellte Erkennungsaufgabe wenig komplex ist. Es sollten in verschiedenen Gesichtern die entsprechenden Teilmodelle gefunden werden, d.h. eine Klassifikation der Gesichter untereinander fand nicht statt. Bemerkenswert an der Arbeit ist, daß hier Methoden des Aktiven Sehens eingesetzt werden, um die Erkennungsergebnisse zu stabilisieren. Es kommt eine Interessen-Karte zum Einsatz: eine Kombination der wohlbekannten Hemmungs- und Verstärkungskarten. Diese Karte wird benutzt, um sukzessiv Szenenpositionen anzufahren und dort ein Modellmatch unter translationsinvarianten Bedingungen einzuleiten. Die Autoren sprechen in diesem Zusammenhang auch von “Sakkaden”.

3.4 Optimierungsverfahren

Durch Optimierungsverfahren werden im allgemeinsten Fall Objektbeschreibungen durch schrittweise Veränderungen an Szenenbeschreibungen (oder umgekehrt) angepaßt. Das dabei erreichte Maß an Übereinstimmung zeigt das Vorhandensein dieser Objekte in der Szene an. Meist wird eine hochdimensionale Energiefunktion aufgestellt, die den Match repräsentiert. Innerhalb dieses oft als “Energiegebirge” bezeichneten Raumes wird nach “Tälern”, den die Matchergebnisse repräsentierenden Energieminima, gesucht. Zur Minimierung der Funktion kommen z.B. Hopfield-Netz-Modelle oder *simulated annealing*-Verfahren zum Einsatz.[3]

3.4.1 Hopfield-Netze

Hopfield-Netze minimieren eine spezifische Energiefunktion. Zur Objekterkennung werden die Neuronen dabei in einer zweidimensionalen Matrix angeordnet und die Reihen und Spalten mit den Modell- bzw. Szenenmerkmalen belegt (siehe Abbildung 3.18). Der Ausgabewert der Neurone repräsentiert hierbei ein Ähnlichkeitsmaß zwischen Modell- und Szenenmerkmal. Die zu minimierende Energiefunktion ist aus mehreren Termen zusammengesetzt:

$$E = -\sum_i \sum_k \sum_j \sum_l \mathcal{A} + \sum_i \mathcal{B} + \sum_k \mathcal{C} \tag{3.13}$$

[3]Umgekehrt wäre auch eine Beschreibung als Maxima denkbar.

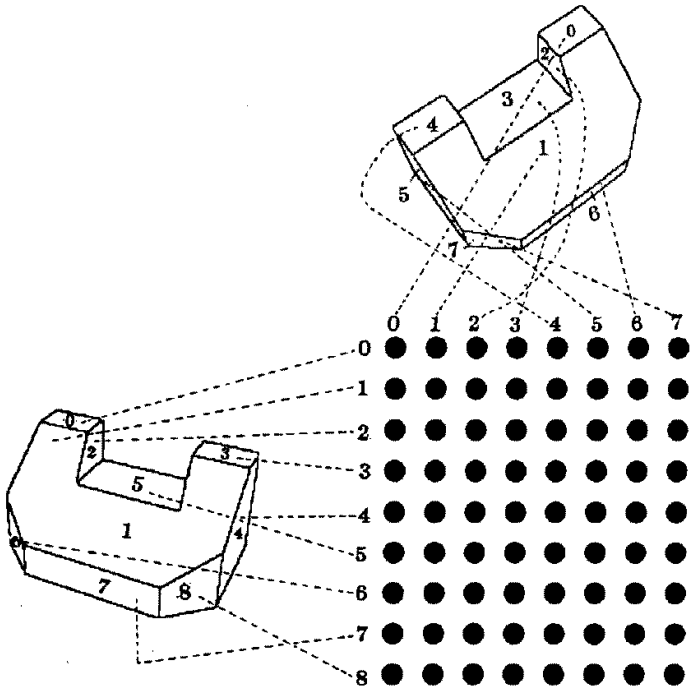


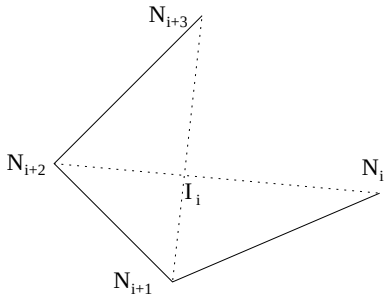
Abbildung 3.18: Belegung der Hopfield-Matrix (nach [LIN et al. 1991])

Der erste Term $\mathcal{A}$ beurteilt die Kompatibilität von Modellmerkmalen (Neuron (i,k)) und Objektmerkmalen (Neuron (j,l) und geht somit negativ in die Rechnung ein. Die beiden folgenden Terme $\mathcal{B}$ und $\mathcal{C}$ wichten das mehrfache Auftreten von aktiven Neuronen in Zeilen bzw. Spalten der Matchmatrix durch einen positiven Beitrag zur Energiefunktion. Am Ende des Optimierungsprozesses sollte pro Zeile und Spalte nur ein Neuron aktiv sein, welches die größte Ähnlichkeit zwischen zwei Modell- und Szenenmerkmalen angibt. Aufgabe der Optimierungsalgorithmen ist die zweckmäßige Auswahl von Neuronen und die sinnvolle Änderung ihrer Werte entsprechend den gegebenen Merkmalen. Weitere Ausführungen zur Theorie der Hopfield-Netze können z.B. [ZELL 1994] entnommen werden. Im folgenden werden einige Arbeiten vorgestellt, die sich mit der Objekterkennung durch Hopfieldnetze beschäftigen.

In [LI und NASRABADI 1989] kommt ein binäres Hopfieldmodell zum Einsatz, welches zur Klassifikation von Werkzeugen benutzt wird. Ausgangspunkt sind perfekt generierte Kantenbilder ohne störende Hintergrundeinflüsse. Als Repräsentationsschema kommen hier Graphen zum Einsatz. Die Knoten dieses Graphen werden durch spezifische Merkmale besetzt, die Graphenkanten stellen

meist wirkliche Kanten im Grauwertbild dar. Ein generierter Szenengraph wird gegen einen Modellgraphen gematcht. Dabei werden eine Reihe von Objekttransformationen zugelassen. Die präsentierten Objekte konnten überlagert, verschoben und in der Ebene rotiert erkannt werden. Skalierungen werden nicht unterstützt, da als Merkmale u.a. euklidische Distanzen eingesetzt werden.

Gee et al. haben sich besonders mit Möglichkeiten zur verbesserten Stabilität der Matchergebnisse bei Verwendung von Hopfieldnetzen beschäftigt [GEE et al. 1991]. Besonders heben sie dabei die Mehrfachzustands-Kodierung gekoppelt mit einem *annealing*-Prozeß [PETERSON und SÖDERBERG 1989] und die Eigenvektor- und Unterraumanalyse des Netzwerkverhaltens hervor [AIYER et al. 1990]. Letzteres ermöglicht eine effektive Entkopplung der Gültigkeits- und Kostenfunktionsterme in der Netzwerkgleichung, wodurch sich eine höhere Effektivität bei der Abarbeitung auf seriellen Rechnern ergeben soll. Beispielhaft soll die Klassifikation von handgeschriebenen Ziffern das Potential dieses Ansatzes zeigen. Es werden auf den skelettierten Ziffern gleichmäßig 20 Punkte angeordnet. Für einen Modellpunkt auf der Ziffer wird ein Modellraum definiert, auf den die Szenenpunkte projiziert werden. Zuvor werden alle Objekte so verschoben, daß ihr Schwerpunkt im Koordinatenursprung liegt. Letztendlich werden die Szenenpunkte solange linear transformiert, bis mit einem Modell eine bestmögliche Übereinstimmung erzielt worden ist. Diese Optimierungsaufgabe wird wieder einem Hopfield-Netzwerk übergeben. Trotz des interessanten Ansatzes konnten in den Experimenten keine überzeugenden Matchleistungen gezeigt werden. Viele Matchwerte unterschiedlichster Klassen liegen sehr dicht beieinander. Die Autoren stellten fest: *"We present the results ... not as evidence of the pattern recognitions system's current usefullness."*.



$$Sh_i = \frac{\overrightarrow{|N_i\, I_i|}}{\overrightarrow{|N_i\, N_{i+2}|}} = const$$

Abbildung 3.19: *Shape Number* (nach [LIN et al. 1991])

Lin et al. verwenden einen hierarchischen Ansatz [LIN et al. 1991]. Als Modellbasis werden charakteristische Ansichten von Polyedern eingesetzt. Zunächst

werden die gegebenen Polyederdarstellungen segmentiert und jeder Polygonfläche die Merkmale Fläche und normierte euklidische Distanzen zu den Nachbarflächen zugeordnet. Innerhalb des anschließenden *Surface Matching* wird mit Hilfe eines Hopfieldnetzes der Objektsuchraum eingegrenzt. Bei unklaren Matchergebnissen werden weitere Merkmale berechnet. Interessant ist hier das Merkmal *Shape Number*, welches sich für vier aufeinander folgende Ecken eines Polygonzuges invariant verhält (siehe Abbildung 3.19). Mit Hilfe der weiteren Merkmale wird ein erneuter Matchprozeß gestartet. Ausgangspunkt ist eine zyklische Rotation des Modells bis zum Feststellen einer maximalen Übereinstimmung.

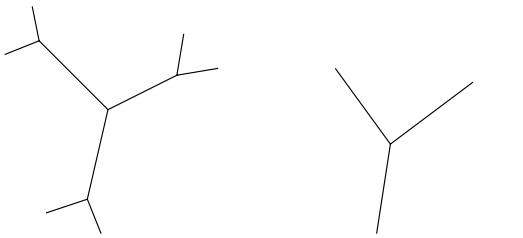
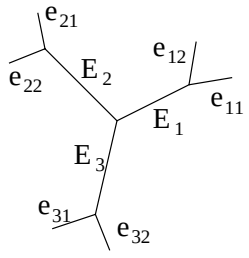


Abbildung 3.20: Unterschiedliche Priorisierungen von Verzweigungen (nach [CHAO und DHAWAN 1992])

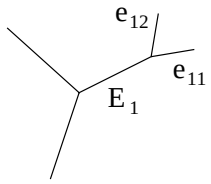
In [CHAO und DHAWAN 1992] wird sich besonders der Generierung möglichst optimaler Kantenbilder und robuster *key features* gewidmet. Zum Einsatz kommen Verfahren der Konturverfolgung, um sowohl unterbrochene als auch mehrfach auftretende Konturen exakt zu verarbeiten. Als *key features* werden Eckenmerkmale bezeichnet, wobei je nach Anzahl der einlaufenden weiteren Verzweigungen eine Priorisierung vorgenommen wird (siehe Abbildung 3.20). Verzweigungen mit höheren Prioritäten werden jeweils zuerst gematcht. Den *key features* werden drei Merkmale zugeordnet:

- *strukturelle:* Die Kanten innerhalb der gefundenen Verzweigungen werden entsprechend gelabelt:



ergibt $(E_1 e_{11} e_{12} E_2 e_{21} e_{22} E_3 e_{31} e_{32})$. Zur Invarianzbildung wird diese Menge entsprechend mehrfach rotiert: $(E_2 e_{21} e_{22} E_3 e_{31} e_{32} E_1 e_{11} e_{12})$, $(E_3 e_{31} e_{32} E_1 e_{11} e_{12} E_2 e_{21} e_{22})$.

- *relationale:* Eine Reihe von Regeln beschreibt die Zuordnung von Labeln zu der Lagebeziehungen von Kanten. Für



ergibt sich z.B.: $(E_1 \,\mathrm{C}\, e_{11} \,\mathrm{C}\, e_{12})$.

- *Winkel:* Es werden die Winkel zwischen den Kanten der Basisecken (oben mit E bezeichnet) in Bezug auf den Nullwinkel berechnet. Weiterhin werden auch die Winkeldifferenzen zwischen allen Kanten einbezogen.

Diese Merkmale werden wegen der Handhabbarkeit auf ganze Zahlen abgebildet. Die Objekterkennung erfolgt zweistufig unter Einsatz von Hopfieldnetzen. Zunächst wird eine Hypothese zur Auswahl der möglichen Objektklasse unter Nutzung der den *key features* zugeordneten Merkmalen generiert. Das so erhaltene Matchergebnis wird benutzt, um das Objektmodell entsprechend dem präsentierten Modell zu rotieren. Die Strukturbeschreibungen (erstes Merkmal) werden solange umsortiert, bis in der 2D-Anordnung der Neurone im Hopfieldnetz nur noch die Diagonale aktiv ist. Anschließend wird noch eine Skalierungskorrektur vorgenommen. Dazu werden zwei Basisecken überlagert und alle anderen Kanten des Modells solange verschoben, bis sie mit den präsentierten übereinstimmen. Es konnte gezeigt werden, daß dieser Ansatz auch zur Erkennung innerhalb realer Szenen eingesetzt werden kann. Von Vorteil wirkt sich hier die intensive Bildvorverarbeitung aus. Allerdings scheint die Anwendung auf quaderförmige Objekte beschränkt zu sein, da nur dort entsprechende Kanten und Verzweigungen in robuster Form detektierbar sind.

In [KAWAGUSHI und SETOGUCHI 1994] wird für die zweite Verarbeitungsstufe aus [LIN et al. 1991] eine alternative Strategie eingesetzt. Wieder wird eine Datenbasis der typischen Ansichten eines Objektes (auch hier Linienzeichnungen von Polyedern) verwendet. Die erste Verarbeitungsstufe bestimmt eine grobe Näherung, eine Matchhypothese, zwischen der Szene und den Modellen. Basis sind auch hier wieder Lagebeziehungen von angrenzenden Objektflächen. Treten nach Abschluß des initialen Optimierungsprozesses, dem *Surface*

Matching noch Mehrdeutigkeiten auf, d.h. fehlende eindeutige Zuordnungen innerhalb der Matchmatrix, wird eine zweite Verarbeitungsstufe eingesetzt. Innerhalb dieser werden Korrespondenzen zwischen Modell- und Szenenecken hergestellt und ein Fehlermaß auf der entsprechenden Transformationsmatrix berechnet. Dadurch konnten Hypothesen überprüft und Scheinlösungen ausgesondert werden. In [RAY und MAJUMDER 1994] werden sehr ähnliche Verfahren wie in [KAWAGUSHI und SETOGUCHI 1994] eingesetzt. Als Merkmale werden dort allerdings Krümmungsmerkmale zur Beschreibung von gebildeten Segmentflächen verwendet.

Lee et al. stellen heraus, daß auf Grund des Fehlens einer Konvergenzgarantie für ein globales Maximum das Matchergebnis sehr stark vom gewählten Ausgangspunkt und den verwendeten Merkmalen abhängt [LEE et al. 1995]. Sie verwenden die Extrempositionen und Amplituden der waveletttransformierten Kontur in mehreren Skalierungsebenen als Merkmalsbasis. Dadurch wird die Erkennung auch stark überdeckter Objekte möglich. Der Merkmalsmatchvorgang besteht wieder aus dem Standardvorgehen bei Nutzung von Hopfieldnetzen.

In [SCHAFFER und CHEN 1995] werden Hopfieldnetze eingesetzt, um das *Cyclic Ordered Assignment Problem* zu lösen. Um eine komplexe Objekterkennung zu vereinfachen, werden hier Objekte als aus primitiven Teilen aufgebaut betrachtet. Verwendet werden aber nur Objekte ohne innere Konturen. Die äußeren Konturen werden entsprechend gelabelt und gegen eine Modellbasis aus der Menge aller Konturlabels gematcht. Offensichtlich ist dabei der Anfangspunkt des Labels entscheidend, wodurch sich das Problem der zyklischen Rotation der Label ergibt. Dieses Problem wird als Optimierungsaufgabe formuliert und mittels Hopfieldnetzen gelöst. Jedes Neuron repräsentiert hierbei ein Objektteil. Durch geschicktes Formulieren der Energiefunktion und der Gewichtsbeschreibungen werden sowohl eins-zu-eins als auch keine-zu-eins Abbildungen zwischen Modellabel und Szenenlabel unterstützt.

3.4.2 *Annealing*-Verfahren

Das bereits seit längerer Zeit erfolgreich eingesetzte Verfahren des *simulated annealing* wird zur Lösung von Optimierungsaufgaben eingesetzt, wobei durch das "simulierte Abkühlen" stochastische Modifikationsoperatoren gesteuert werden: bei den anfangs hohen "Temperaturen" können dabei stärkere Veränderungen stattfinden als nach der schrittweisen "Abkühlung". Dadurch soll das evt. anfängliche Verharren in lokalen Minima verhindert werden.

In [GOLD und RANGARAJAN 1996] wird ein Graphmatchverfahren basierend auf einem neuronal und statistisch motivierten Energieminimierungsverfahren

vorgestellt. Durch den Einsatz der als *softassign* bezeichneten Technik kann ein *constraint*-Problem ohne die Nutzung von Strafwerten zum Matchen von zwei durch Graphen beschriebenen Objekten gelöst werden. Die Autoren haben Energiefunktionen definiert, die durch das explizite Kodieren von unvollständig besetzten Adjazenzmatrizen das Verarbeiten von Graphen mit zusätzlichen bzw. fehlenden Knoten oder Kanten ermöglichen. Die sich ergebenden Energiefunktionen erinnern sehr stark an die Formulierungen für Hopfield-Netze. Besonderheiten zeigen sich nur in der Handhabung von fehlenden Knoten bzw. Kanten und der Einbeziehung von Graph-Distanzmaßen in die Energiefunktion. Diese Funktion wird dann mittels *simulated annealing* minimiert. In jeder Iteration werden die Zeilen- und Spaltenwerte einer Matchmatrix verändert und anschließend neu normalisiert. Die Leistungsfähigkeit wird an einer Reihe von Graphmatchproblemen demonstriert. Leider konnte die Erkennung realer Objekte nur an per Hand konstruierten Graphen gezeigt werden. Es werden mehrere Knoten - und Kantentypen definiert, die verschiedene geometrische Merkmale kodieren. Durch die so konstruierten Objektgraphen konnte ein affin transformiertes Beispielobjekt innerhalb einer strukturierten Umgebung erfolgreich gematcht werden. Das System wird weiterhin an einer großen Anzahl zufällig erzeugter Graphen demonstriert.

In [AZENCOTT und YOUNES 1997] wird ein Erkennungssystem vorgestellt, welches auf der Basis von Segmentbeschreibungen den Vergleich zwischen Modell und Szene vornimmt. Als Optimierungsverfahren kommt *simulated annealing* zum Einsatz. Die Autoren stellen fest, daß ein direktes Matchen von Strukturbeschreibungen für Objekte problematisch ist, da das Ableiten einer entsprechenden Beschreibung aus einer Szene nicht perfekt durchgeführt werden kann und sich somit Differenzen ergeben. Sie formulieren deshalb die Objekterkennung folgendermaßen: Objektbeschreibungen sind dann ähnlich, wenn sich die vereinfachten, teilweise zusammengefaßten Beschreibungen ähneln. Innerhalb der definierten Kostenfunktion wird das Zusammenfassen von Segmenten berücksichtigt. Die Kostenfunktion beinhaltet drei Kriterien:

- Es wird die Ähnlichkeit der Deskriptoren der Vereinfachungen der Modell- und Szenenbeschreibung beurteilt.
- Die Menge der nicht matchenden Segmente sollte möglichst klein sein.
- Die Menge der zusammengefaßten, vereinfachten Segmente sollte möglichst klein sein.

Die problematische Parametrisierung bei Minimierungen unter Verwendung des *simulated annealing*-Algorithmus begegnen die Autoren mit der Feststellung,

daß bei günstiger Formulierung der elementaren Modifikationen und bei genügend kleinen Temperaturänderungen der Algorithmus ein globales Minimum bereitstellt. Leider können dem Artikel keine genaueren Angaben zur Wahl der Modifikationen entnommen werden. Als Merkmalsbasis kommen Segmentbeschreibungen von 2D-Ansichten zum Einsatz. Als Segment werden hinsichtlich ihrer Farbe homogene Regionen verwendet. Ziel ist eine skalierungs- und rotationsinvariante Beschreibung. Eingesetzt werden folgende Merkmale:

- Die Positionsdeskriptoren werden aus dem Schwerpunkt des Segments abgeleitet. Dabei wird ein Bezug zum Schwerpunkt des gesamten Objektes hergestellt. Durch die Verwendung von normierten euklidischen Distanzen zwischen den Schwerpunkten erhält man invariante Beschreibungen.

- Die Formdeskriptoren setzen sich aus zwei Beschreibungen zusammen: Der relativen Größe eines Segments im Vergleich zur Gesamtfläche des Objekts und Distanzen zwischen den durch Ellipsen beschriebenen Trägheitsmomenten zweier Regionen.

Mit diesen Merkmalen konnten die Autoren gute Ergebnisse bei fehlendem Hintergrund erzielen. Zu beachten ist, daß allerdings keine vollständige Invarianz erreicht werden konnte.

3.4.3 Relaxationsverfahren

Relaxationsverfahren werden eingesetzt, um eine sich iterativ entwickelnde Zuordnung zwischen Merkmalen bzw. Komponenten der Szenen- und Modellobjekte zu finden. Relaxation bedeuted in diesem Umfeld die schrittweise Veränderung von parametrischen, lokalen Beschreibungen, um globale Objektbeziehungen herzustellen.

Die Arbeitsgruppe um W. von Seelen nutzt Relaxationsverfahren im Rahmen eines auf Verkehrssituationen angepaßten Erkennungssystems ([SCHWARZINGER et al. 1992], [NOLL et al. 1993], [SCHWARZINGER et al. 1995]). Durch diese Einschränkung konnte auf die Einbeziehung von Rotationsinvarianz in das System bereits auf der Entwurfsebene verzichtet werden. Als Basismerkmale werden geometrische Primitive wie Linien, Ecken und deren eingeschlossene Winkel eingesetzt. Durch eine modifizierte Hough-Transformation werden initiale Hypothesen, sowie eine Schätzung für die Position und Skalierung generiert. Für die Validierung kommt ein elastisches Matchverfahren nach [DURBIN und WILLSHAW 1987], dort wird es zur Lösung von *Travelling Salesman*-Problemen verwendet, zum Einsatz. Innerhalb eines iterativen Matchprozesses paßt sich das Modell an die gegebene

Merkmalsverteilung an. Dabei wird das Modell deformiert, wobei die globale Form des Modells erhalten bleiben soll. Es werden zwei Objektmodelle getrennt an die Szenenmerkmale angepaßt: Ein generisches Modell P wird benutzt, um die Modellmerkmale an die Szene anzupassen. Dazu wird P elastisch deformiert, wobei es die globale Form beibehalten soll. Das initiale Modell P^I wird ebenso wie P verschoben und skaliert, aber nicht deformiert, d.h. P^I wird zur globalen Deformationssteuerung verwendet. Jeder Iterationsschritt ist in drei Teile gegliedert:

1. Das Modell P wird verschoben, skaliert und deformiert gemäß lokalen und globalen Ähnlichkeitsmaßen:

   - *lokal:* Berechnet werden die Differenzen der eingeschlossenen Winkel von matchenden Eckenmerkmalen, die Längen- und Winkeldifferenzen von Geradenstücken, wobei die Winkelinformation von Geraden höher gewichtet wird als die unsichere Längeninformation.
   - *global:* Es werden sog. *chord distributions* als ein Konzept zur Formenkodierung geschlossener Konturen eingesetzt. Eine *chord distribution* $h(r, \phi)$ beschreibt die Verteilung der Längen r und Winkel ϕ aller *chords*, die zwei äußere Punkte einer geschlossenen Kontur verbinden (siehe Abbildung 3.21 für ein Beispiel). Eine Verteilung $h_j^s(r, \phi)$ der Längen und Winkel aller *chords*, die das Merkmal j mit allen anderen Merkmalen kodiert, beschreibt die relative Position von j zu den anderen Merkmalen des Modells. Diese Verteilung ist nicht skalierungsinvariant, weshalb eine solche Verteilung für alle möglichen Skalierungen s berechnet werden muß.

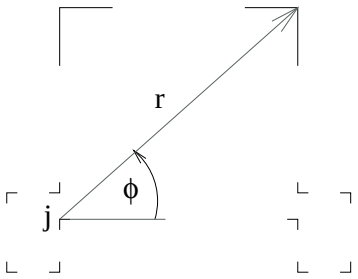


Abbildung 3.21: Beispielhafte *chord distribution* für ein Merkmal j (nach [SCHWARZINGER et al. 1995])

   Die Verteilungen der Szenenmerkmale werden dann fuzzifiziert, d.h. die im (r, ϕ)-Raum ausgebildeten Stützstellen werden räumlich ausgedehnt. Eine anschließende UND-Verknüpfung mit einer Modell-

verteilung ergibt im positiven Fall eine Übereinstimmung der resultierenden Stützstellenpositionen zwischen Szene und Modell. Die Höhe der Stützstellen ist ein Maß für die Ähnlichkeit der Verteilungen. Dieses Verfahren ist skalierungsvariant, d.h. das Ähnlichkeitsmaß muß für mehrere Skalierungen berechnet werden.

2. Das Modell P^I wird nun nach den errechneten Werten für P verschoben und skaliert.

3. Die Differenz der Positionen von P und P^I ergibt ein Deformationsvektor, der zu einer teilweisen Rücknahme der durchgeführten Deformation benutzt wird, um ein besseres dynamisches Verhalten zu erreichen.

Nach jedem Iterationsschritt werden vier Größen begutachtet, um evtl. eine Matchentscheidung abzuleiten:

- Der Translationsvektor während der Anpassung des Modells P^I.
- Der Skalierungsfaktor für die Anpassung des Modells P^I.
- Ein Maß für den Umfang der aktuellen Deformation, der aus den Positionsdifferenzen von P und P^I abgeleitet wird.
- Ein Maß für die aktuelle Ähnlichkeit von Modell- und Szenenmerkmalen.

Das System wird entsprechend der Aufgabenstellung für die Erkennung von Fahrzeugen im Straßenverkehr eingesetzt. Dabei kann es durch robuste und schnelle Ergebnisse überzeugen.[4] Auch in Extremsituationen (kein Objekt bzw. kein gelerntes in der Szene) liefert das System korrekte Ergebnisse. Die Repräsentation von relativen Positionen von Merkmalen zueinander ist hier der Schlüssel zu einem einfachen aber leistungsfähigen Erkennungssystem. Der Nachteil, daß mehrere Skalierungen getrennt repräsentiert werden müssen, erwies sich in der Praxis (zumindest für den gewählten Anwendungsfall) nicht als nachteilig. Eine Anpassung zum sukzessiven Erkennen mehrerer Objekte wird ebenfalls vorgestellt.

Die Arbeitsgruppe um E. R. Hancock untersucht die Nutzung von Graphenrepräsentationen innerhalb von Objekterkennungssystemen. In [FINCH und HANCOCK 1995] wird die Verwendung von *Delaunay*-Graphen als Objektrepräsentation eingesetzt. Diese entstehen bei einer Zergliederung von polygonalen Objekten in elementare Dreiecke, der *Voronoi*-Zerlegung:

[4] durchschnittlich werden 100 Iterationen bis zu einer Entscheidung benötigt

Jeder Punkt des Polygons wird auf die minimale euklidische Entfernung zu den Eckpunkten abgetestet. Ergeben sich gleiche Entfernungen für mehrere Ecken, bilden diese Punkte zusammen mit den Polygonseiten die Grenzen einer *Voronoi*-Zelle. Durch die Verbindung der Eckpunkte von benachbarten *Voronoi*-Zellen ergibt sich die *Delaunay*-Zerlegung. Die Autoren benutzen die Mittelpunkte von extrahierten Kantensegmenten als Eckpunkte der polygonalen Objekte und als Knoten der *Delaunay*-Graphen. Durch die geschickte Verbindung von benachbarten Knoten ergibt sich die Objektzerlegung. Die besondere Eignung der *Delaunay*-Graphen für Aufgaben der 2D- und 3D-Objekterkennung durch eine hohe Robustheit gegenüber Rauschen und Veränderungen des Beobachtungspunktes wird herausgestellt. Es werden jeweils drei untereinander verbundene Knoten zu einem sog. *face* zusammengefaßt. Die zu matchenden Graphen werden notiert als $G = (V, E, F)$, wobei V die Knoten, E die Kanten und F die *faces* repräsentieren. Die entwickelte Matchstrategie beruht auf die Ausnutzung der durch die *face*-Menge repräsentierten *constraints*. Während des Matchens wird der betrachtete Knoten innerhalb des Modellgraphen immer zusammen mit den kontextuellen Nachbar-*faces* untersucht (siehe Abbildung 3.22). Der eigentliche Matchprozeß wird durch eine probabilistische Relaxation realisiert. Es werden die Wahrscheinlichkeiten für die verschiedenen kombinatorischen Labelkonfigurationen geschätzt. Dazu werden zu einer Reihe von Labelmöglichkeiten Wahrscheinlichkeiten angegeben. Dadurch können alle Wahrscheinlichkeiten ohne freie Parameter geschätzt werden. Als Beispielanwendung werden Satellitenaufnahmen mit Kartenmaterial gematcht. Dabei werden die Vorteile aus der Nutzung von *faces* anstatt Kantenelementen als grundlegende Matcheinheit deutlich. Allerdings werden immer noch 5 von 25 Elementen (20%) falsch zugeordnet.

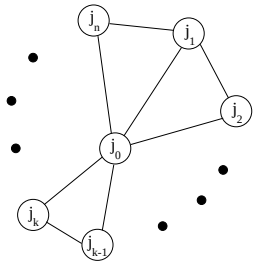


Abbildung 3.22: Betrachteter Knoten j_0 mit kontextuellen Nachbarn (nach [FINCH und HANCOCK 1995])

In [WILSON und HANCOCK 1995] werden ebenfalls die oben beschriebenen *Delaunay*-Graphen eingesetzt, obwohl nicht explizit erwähnt. Abgegangen wird vom Begriff *face*. Das Erkennungsproblem wird nun als der Match zwischen

zwei Graphen $G = (V, E, A)$ betrachtet, wobei V wieder die Knoten, E wieder die Kanten und A nun allgemein die Menge unärer Merkmale der Knoten repräsentieren. Die grundlegende Matcheinheit wird nun als *superclique* bezeichnet und stellt wiederum den zu matchenden Knoten mit den verbundenen Nachbarn dar (siehe Abbildung 3.22). Neu ist der Gedanke, den Erkennungsprozeß nicht strikt zwischen Vorverarbeitung, dem Aufbau relationaler Graphenstrukturen, und dem Matchen zu trennen. Vielmehr sollen beide Prozesse in intensiver Interaktion stehen, wodurch z.B. das Löschen und Wiedereinfügen von Knoten möglich wird. Der eigentliche Match ist wieder ein diskreter Relaxationsprozeß und wird gesteuert durch eine *posteriori*-Wahrscheinlichkeitsschätzung. Die durch diese Schätzung unberücksichtigten Knoten werden entfernt. Die Leistungsfähigkeit wird wieder am Beispiel des Matchens von Satellitenaufnahmen gezeigt. Der Ansatz wird in [CROSS und HANCOCK 1996] dahingehend erweitert, daß genetische Algorithmen zur Suche des globalen Minimums eines Bayes'schen Konsistenzmaßes verwendet werden. Damit konnten gute Ergebnisse auch bei sich überlappenden Graphen erzielt werden. Die bei genetischen Algorithmen notwendigen Festlegungen der Wahrscheinlichkeiten für Einfüge- oder Löschvorgänge und über den Initialzustand sind allerdings problematisch. In [WILSON und HANCOCK 1996] und [WILSON et al. 1996] werden die Verfahren und Ergebnisse nochmals zusammengefaßt. Allerdings finden sich auch hier keine weiteren Demonstrationen von Erkennungsleistungen jenseits des Satellitenbildmatchens.

In [SHAO und KITTLER 1996] werden deterministische Relaxationsverfahren zum Labeln von Szenen bzw. zur Korrespondenzanalyse für objektzentrierte Repräsentationen eingesetzt. Es wird festgestellt, daß die häufig eingesetzten probabilistischen Verfahren zwar gute Ergebnisse liefern, aber einige fundamentale Probleme nicht zufriedenstellend lösen können. Dementsprechend wird ein deterministisches Verfahren unter Verwendung der *Fuzzy*-Logik vorgestellt. Durch die Formulierung der Evidenz-Kombinierung mit Hilfe der *Fuzzy*-Logik ist es den Autoren möglich, die kombinatorische Evaluation von Nachbarschaftsverhältnissen in ein *Maximum Weight Bipartite Matching*-Problem zu überführen, wofür effiziente Algorithmen existieren. Damit kann ein deterministischer, nicht iterativer Algorithmus entwickelt werden. Innerhalb eines Beispiels zur Erkennung von Straßen in handgezeichneten Karten wird die Leistungsfähigkeit demonstriert. Dazu werden die Liniensegmente in eine invariante binäre Relation überführt. Die Relation x_{ij} besteht aus einem vierdimensionalen Vektor, wobei $p1_i$ und $p2_i$ der Start- bzw. Endpunkt des $i$ten Liniensegmentes und $p1_j$ und $p2_j$ des $j$ten Liniensegmentes ist (siehe auch Abbildung 3.23):

$$x_{ij} = s_{ij} \left[\frac{d(p1_i, p1_j)}{d(p1_i, p2_i)}, \frac{d(p1_i, p2_j)}{d(p1_i, p2_i)}, \frac{d(p2_i, p1_j)}{d(p1_i, p2_i)}, \frac{d(p2_i, p2_j)}{d(p1_i, p2_i)} \right] \quad (3.14)$$

mit

- d euklidische Distanz zwischen zwei Punkten
- s Vorzeichenfunktion: 1, wenn $p1_i$, $p2_i$ und $p1_j$ in Uhrzeigerrichtung bzw. kollinear angeordnet sind; 0, wenn nicht

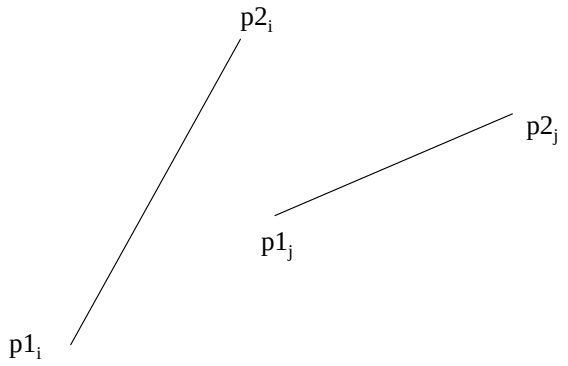


Abbildung 3.23: Anordnung der Merkmalspunkte zur Berechnung von x_{ij}

Grundidee ist die Beschreibung des a priori Wissens und der Szene durch *Fuzzy*-Mengen und die anschließende “geschickte” Vereinigung der beiden Mengen für die Label-Funktion.

3.5 Biologienahe Systeme

In diesem Kapitel sollen Systeme vorgestellt werden, die sich als nah im biologischen Vorbild betrachten. Ausgangspunkt sind die jeweils aktuellen Erkenntnisse aus Neurophysiologie und Psychophysik. Ausdrücklich wird eine direkte Umsetzung dieser Forschungsergebnisse hinsichtlich Struktur und Verarbeitungsregime in technische Systeme angestrebt. Die folgenden Kapitel unterscheiden sich im Grad des Festhaltens am biologischen Vorbild: biologisch motivierte Systeme versuchen eine entsprechende Verarbeitung inklusive angepaßter Vorverarbeitung umzusetzen, wohingegen konnektionistische Systeme von eigentlichen biologischen Verarbeitungsprinzipien mehr abstrahieren.

3.5.1 Biologisch motivierte Systeme

Die im Rahmen des NAVIS-Projektes entstandenen Erkennungssysteme mit starkem biologischen Bezug werden in den Kapiteln 4.3 bzw. 6 ausführlich beschrieben.

In [BIESZCZAD 1994] wird der *Neurosolver* als Rechnerumsetzung biologischer kortikaler Prozesse vorgestellt. Hierbei werden Problemlösungen durch das "Abfahren" von Trajektorien, die die temporalen Beziehungen zwischen Objekten einer Domäne beschreiben, generiert. Basiselement ist die in Abbildung 3.24 dargestellte *column*-Einheit. Globale Informationsverarbeitung findet durch das Zusammenwirken vieler in einer Achternachbarschaft angeordneter *column*-Einheiten statt. Die internen "Oben-Unten"-Verbindungen übergeben Aktivitäten von der oberen Schicht zur unteren. Die "Unten"-Schicht inhibiert die "Oben"-Schicht durch ihre Aktivität und generiert eine schwellwertbehaftete Ausgabe der Einheit. Die externen Verbindungen stellen Beziehungen zu Eingabegrößen und zu anderen *column*-Einheiten her und besitzen eine spezifische Verbindungsstärke. Die Aktivierung einer "Oben"-Schicht drückt ein Ziel, Konzept oder einen angestrebten Zustand aus. Ein Ziel gilt dann als erreicht, wenn die Aktivität der Einheit unterdrückt wird und die Einheit "feuert". Dann entspricht der angestrebte Zielzustand den angelegten externen Eingabedaten. Es werden eine Reihe von Regeln zur Aktivitätsausbreitung und zur Bestimmung der Verbindungsstärke in einem Verband aus mehreren *column*-Einheiten angegeben. Eine Problemlösung findet nun durch eine initiale Erregung einer Einheit und der anschließenden Aktivitätsausbreitung auf einer gelernten Trajektorie statt. Dabei wird eine "Breite Zuerst"-Suche realisiert. Leider konnte in der Arbeit das Potential des *Neurosolver*-Ansatzes durch das Fehlen konkreter Anwendungen nicht demonstriert werden. Angedacht ist die Anwendung für eine sichtgesteuerte Roboternavigation.

In [BASAK und PAL 1995] wird das System *PsyCOP* zur Erkennung flacher industrieller Objekte vorgestellt. Explizit wird ein *Was*- und *Wo*-Zweig (siehe Kapitel 3.1.2) unterschieden. Unter Mitwirkung von Mechanismen der selektiven Aufmerksamkeit werden initiale Hypothesen generiert, die anschließend durch ein konnektionistisches, rückgekoppeltes System verifiziert werden. Als Merkmale werden Ecken zusammen mit der Orientierung der Winkelhalbierenden benutzt (siehe Abbildung 3.25). Durch dieses Merkmal wird ein Koordinatensystem aufgespannt, welches zur rotations- und translationsinvarianten Lokalisierung von Objektpositionen verwendet wird. Der Ansatz benutzt zwei getrennte Kanäle zur Verarbeitung von Objektidentitäten und -positionen (siehe Abbildung 3.26). Die verdeckte Schicht kodiert Merkmals-Objekt-Assoziationen, die Ausgabeschicht gefundene Objekte. Zwischen den einzelnen Schichten sind eine Viel-

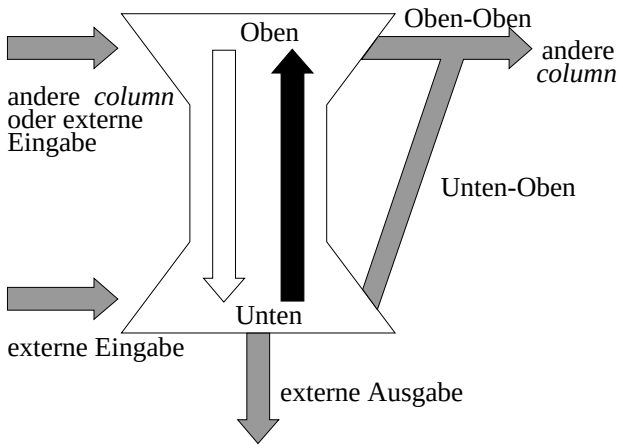


Abbildung 3.24: Blockdiagramm einer *column*-Einheit (nach [BIESZCZAD 1994])

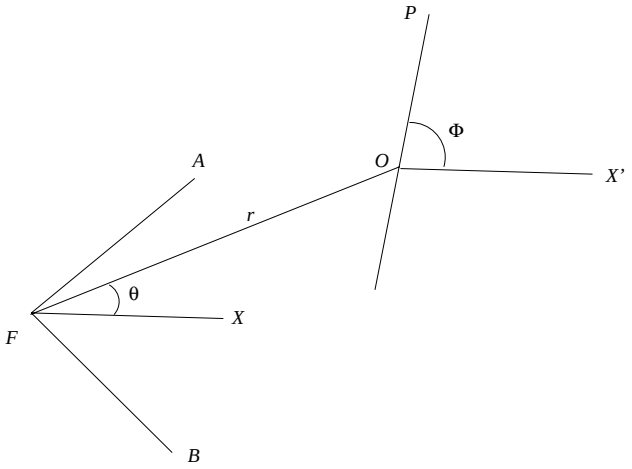


Abbildung 3.25: Objektpositionskodierung (nach [BASAK und PAL 1995]); AFB ist eine Ecke; FX die Winkelhalbierende; OP die Hauptachse des Objekts; (r, θ, ϕ) beschreibt die Position des Objekts in Bezug zur Merkmalsecke

zahl verschiedener Verbindungen vorgesehen (angedeutet durch die unterschiedlichen Verbindungslinien in Abbildung 3.26), die sowohl verstärkend als auch inhibierend wirken können, um eindeutige Assoziationen zu erzeugen.

Durch die *A*-Zellen wird der *Was*-Zweig realisiert, durch die *B*-Zellen der *Wo*-Zweig. Durch die Neuronen der Eingabeschicht werden alle Merkmale kodiert, d.h. alle Merkmalskombinationen von (r, θ, ϕ) werden auf die Menge der Neurone abgebildet. Jedes Merkmal generiert eine Menge von hypothetischen Objekten mit

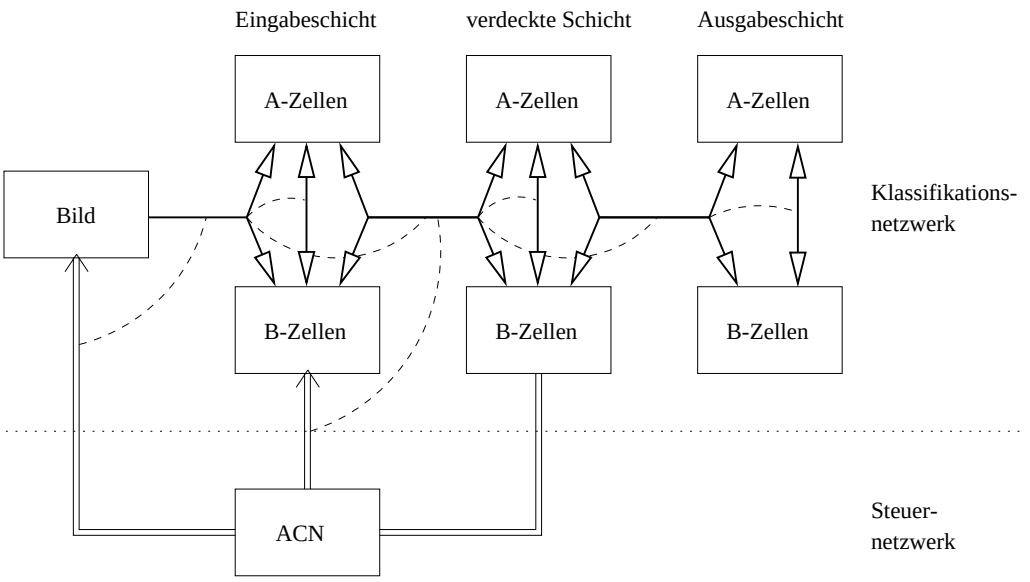


Abbildung 3.26: Blockdiagramm von *PsyCOP* (nach [BASAK und PAL 1995]); das Klassifikationsnetzwerk bestimmt Position und Identität eines Objekts, während das Steuernetzwerk selektive Aufmerksamkeitsmechanismen bereitstellt.

ihren Positionen durch die in den Neuronenverbindungen gespeicherten (r, θ, ϕ)-Werte. Dabei werden jeweils nur kleine Regionen des Bildes ausgewertet. Die Auswahl findet durch ein Aufmerksamkeitsmechanismus statt, der die innerhalb eines Bildes vorkommenden Merkmale sequentiell und in Gruppen dem Netzwerk zuführt. Innerhalb der Verifikationsphase werden die Aktivitäten von Hypothesen auf die Merkmalsschicht zurückgekoppelt. Entsprechend dem Verhältnis von rückgekoppelter und ursprünglicher Merkmalsaktivität wird die Hypothese verstärkt oder verringert. Die Aktivitäten in den *A*-Zellen sind binär: entweder 1 oder 0 entsprechend dem Vorhandensein des zugehörigen Elements. Die *A*-Neurone sind als Tabellen angeordnet, entsprechend der maximalen Anzahl von Merkmalen (Eingabeschicht), der maximalen Anzahl von Merkmal-Objekt-Assoziationen (verdeckte Schicht) und der maximalen Anzahl von Objekten, die das System lernen kann (Ausgabeschicht). Die *B*-Zellen kodieren die Wahrscheinlichkeit, daß ein bestimmtes Element an einer bestimmten Position zu finden ist. Positionen von Objekten werden prinzipiell als die Position ihres Schwerpunktes angenommen, die Orientierung ist durch die Hauptachse gegeben. Die *B*-Neurone sind in Form eines Würfels angeordnet, wobei jede Zelle eine spezifische Position und die Orientierung eines Elements angibt. Das *PsyCOP*-System ist in der Lage, industrielle, flache Objekte auch bei Überlagerungen zu erkennen. Problematisch könnte sich das Vorhandensein von störenden Hintergrundinformationen bemerkbar machen,

das System würde den Hintergrund mitlernen. Ungeklärt ist weiterhin die Frage nach einer Skalierungsinvarianz, die zunächst nicht implementiert ist. Nichtsdestotrotz zeichnet sich das System durch eine hohe biologische Relevanz und durch präzise Beschreibungen der Netzwerkstruktur aus.

3.5.2 Konnektionistische Systeme

Ballard nahm in [BALLARD 1984] die Idee der konnektionistischen Verarbeitungsmodelle von Barrow und Tenenbaum [BARROW und TENENBAUM 1978], [BARROW und TENENBAUM 1981] auf. Er beschreibt den grundlegenden Aufbau und die prinzipielle Funktionsweise von konnektionistischen Sehsystemen. In diesem Zusammenhang wird der Begriff der *intrinsic images* geprägt, die physikalische Eigenschaften der Grauwertbilder explizit kennzeichnen. Aus diesem *intrinsic images* werden dann die Merkmalsräume abgeleitet, die raum- und zeitunabhängig organisiert sind. Beide Räume werden durch sog. *Parameter Netzwerke* dargestellt, wobei jedes Element dieser Netzwerke einen bestimmten Parameterwert zusammen mit einem zugeordneten Sicherheitswert repräsentiert. Es werden Prinzipien formuliert, nach denen konnektionistische Sehsysteme arbeiten:

- Auf der niedrigsten Abstraktionsebene werden die *intrinsic images* angelegt. Als Basismerkmale werden vorgeschlagen:
  - Geschwindigkeitsvektoren,
  - Oberflächenorientierungen,
  - verdeckte Konturen und
  - Disparitätswerte.
- Unter Nutzung der verallgemeinerten Hough-Transformation werden die Merkmalsräume, eine weitere Ebene der Abstraktion, berechnet.
- Nutzung von Hierarchien zur Handhabung komplexer Strukturen.
- Einsatz von Techniken zur “dünnen” Kodierung. Abstrakte Repräsentationen sind i.a. nur dünn mit Merkmalen besetzt, wodurch sich Methoden wie die Anwendung von Subräumen oder Projektionen bzw. der “dünnen” Kodierung zur Abbildung hochdimensionaler Merkmalsräume anbieten.
- Segmentierung ist der Konvergenzzustand von Netzwerken. Dabei sind die *intrinsic images* und die Merkmalsbilder eng über globale Parameter gekoppelt. Segmentierung wird begriffen als eine massiv parallele Verrechnung von *intrinsic images* (lokalen Eigenschaften) und einer Menge von Merkmalsräumen (globale Eigenschaften).

- Aufmerksamkeit wird interpretiert durch räumliche Kopplungen und wird bestimmt durch das Zusammenwirken von Sensordatentransformationen und dem sequentiellen Charakter dieser Transformationen.

Grundidee des Ansatzes ist die Umsetzung von Verarbeitungsmechanismen der frühen visuellen Verarbeitung bei Wirbeltieren in Netzwerke aus primitiven Elementen mit spezifischen exzitatorischen bzw. inhibitorischen Kopplungen. Durch die Kopplungen werden sowohl physikalische wie semantische *constraints* abgebildet. Die Lösung ergibt sich dann durch die Konvergenz des Netzwerkes. Es werden Verfahren angegeben, wie die u.U. enorme Menge an Informationen durch verschiedene Abstraktionsebenen verarbeitet werden können. An einfachen Beispielen wird die Arbeitsweise demonstriert.

In [CRUZ et al. 1990] wird ein Erkennungssystem basierend auf Mehrebenennetzwerken vorgestellt. Grundelement ist das *Madaline*, welches aus einem Netzwerk mehrerer *Adaline*-Elemente aufgebaut ist, die in verdeckten Schichten angeordnet sind. Neuronenaktivitäten werden nicht rückgekoppelt (*feedforward*-Netze). Als Lernmechanismus wird die Backpropagation-Regel angewendet. Das Netzwerk soll durch den Lernvorgang selbständig invariante Bildmerkmale generieren. Dazu sollen aus einem 256×256 großen Eingabebild die "charakteristischen"-Werte von 16×16-Subbildern ermittelt werden. Es ergeben sich somit 16×16 Pixel große invariante Beschreibungen. Anschließend wird das so gewonnene invariante Bild in Beziehung zu einem gelernten invarianten Bild gesetzt und eine Rückgenerierung des Eingabebildes, jetzt ohne Störungen und Transformationen, durchgeführt. Das System wird an zwei Beispielen demonstriert, die (leider) nur schwer nachvollziehbar sind. Der Ansatz kann aber als Studie über den Einsatz von künstlichen neuronalen Netzwerken verstanden werden.

In [GUYOT et al. 1990] und [ALEXANDRE et al. 1991] wird die *Cortical Column* als neues konnektionistisches Modell vorgestellt. Es werden folgende Entwurfsprinzipien zugrunde gelegt:

- Die Knoten des Netzwerks sind nicht vollständig miteinander vernetzt. Die Verarbeitungseinheiten sind meist mit einer konstanten Zahl von weiteren Einheiten verbunden.

- Es werden drei Arten von Informationen verarbeitet: externe Stimuli, Aktionen an die externe Umgebung und interne Zustände oder Konzepte.

- Durchführung einer *Top-Down* Analyse durch die Generierung von Subzielen und deren rekursive Prüfung.

Vorbild sind die biologischen *Cortical Columns* im Kortex, die sich nicht durch anatomische Grenzen erkennen lassen, sondern durch ihr homogenes Verhalten unter Anlegen eines spezifischen Stimulus. Höhere kognitive Fähigkeiten werden als Resultat des parallelen Zusammenwirkens lateral verbundener *Columns* gesehen. Die Autoren fassen die neurophysiologisch festgestellten 6 Schichten in drei zusammen:

1. Die Pyramiden-Zellen der oberen Schichten simulieren die oberen Schichten 2 und 3. Sie empfangen interne Eingaben und stehen direkt oder indirekt mit allen Teilen des Kortex in Verbindung (interne Ausgabe).

2. Die Zwischenschicht empfängt Sensorsignale vom Thalamus (externe Eingabe) und von anderen kortikalen Gebieten, die in die Vorverarbeitung von Sensordaten involviert sind. Sie simuliert Schicht 4.

3. Die Pyramiden-Zellen der unteren Schichten simulieren die Schichten 5 und 6 und lösen Aktionen zur externen Welt aus und erzeugen interne Rückkopplungen.

Verbindungen zwischen den Einheiten existieren jeweils zu den Nachbarn innerhalb einer Schicht und zu entfernteren Einheiten funktional ähnlicher Schichten. Treffen mehrere Aktivitäten gleichzeitig auf eine Einheit, setzt sich die mit maximaler Aktivität durch. Es werden drei Aktivitätsstufen unterschieden, durch die kortikale Vorgänge wie selektive Aufmerksamkeit, Erwartungen oder das Auslösen von Aktionen repräsentiert werden sollen:

1. *low level* zur Repräsentation von moderaten Neuronenaktivitäten, durch die keine Ausgabeaktivitäten erzeugt werden, die aber Bedeutung für Aufmerksamkeits- und Erwartungsprozesse besitzen.

2. *higher level* zur Darstellung von Aktionspotentialen mit einer Frequenz von 50-100 Hz, wie sie bei sensomotorischen Interaktionen beim Matchen von internen Filtern (z.B. orientierungsspezifischen) entstehen.

3. Inhibitorische Prozesse werden durch eine Null-Aktivität dargestellt.

Mit Hilfe obiger Vereinbarungen stellen die Autoren eine Aktivierungstabelle auf, die auf spezifische Aktivitäten in Abhängigkeit vom globalen Zustand der *Cortical Column* mit internen und externen Aktivitätsstufen reagiert. Die Lernmechanismen werden ebenfalls über tabellarische Zuordnungen realisiert. Das Gesamtsystem wird u.a. an einfachen Problemen der Bildverarbeitung getestet. Es werden binäre, in einem 8×8 Raster dargestellte Buchstaben gelernt und verrauschte wiedererkannt. Transformationen der Buchstaben kommen dabei nicht

vor. Weitergehende komplexere Aufgabenstellungen werden nicht untersucht. Für diese Aufgabe sind insgesamt 1200 Einheiten mit 6 Verbindungen pro Einheit nötig.

In [KOLLIAS et al. 1991] wird die herkömmliche Nutzung von Multilayer-Netzen für Anwendungen der Objekterkennung kritisiert. Es wird festgestellt, daß solche Netze nicht die inhärenten Beziehungen zwischen den Knoten bzw. Neuronen berücksichtigen. Dementsprechend wollen die Autoren künstliche neuronale Netze höherer Ordnung (*Higher Order Neural Networks (HONN)*) einsetzen. Konkret kommen *HONN* dritter Ordnung zum Einsatz (siehe Abbildung 3.27), um Invarianzen bezüglich Translation, Rotation und Skalierung bereitzustellen. Typischerweise berechnet sich dabei der $i$te Ausgang y_i wie folgt:

$$y_i = f\left(\sum_j \sum_k \sum_l w_{ijkl} x_i x_k x_l\right) \tag{3.15}$$

x_i bezeichnet hierbei das $i$te Element der Eingangsdaten; w_{ijkl} ist die Wichtung für die Verbindung der multiplikativen Verknüpfung des $j$ten, $k$ten und $l$ten Eingangselements mit dem Ausgabeelement; f ist die Sigmoidfunktion. Die Summationsgrenzen decken den gesamten Eingangsraum ab.

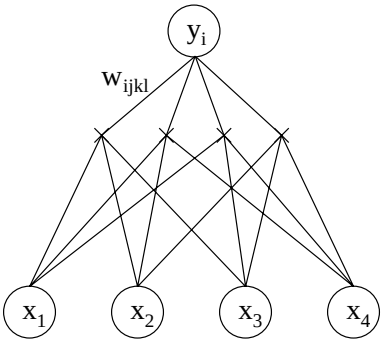


Abbildung 3.27: Grundaufbau eines künstlichen neuronalen Netzes dritter Ordnung (nach [KOLLIAS et al. 1991])

Als Merkmalsbasis wird eine geometrische Analyse der Eingabedaten verwendet. Jeweils drei Merkmalspunkte können als Dreiecke betrachtet werden. Die Innenwinkel dieser Dreiecke sind dabei invariant bezüglich Translation, Rotation in der Ebene und Skalierung. Problematisch hierbei ist jedoch die u.U. enorme Zahl an möglichen Dreiecken. Die Autoren stellen deshalb einen Algorithmus vor, der die Anzahl der zu berücksichtigenden Dreiecke reduziert. Zunächst wird die Beschreibung der Dreiecke auf zwei Winkel a, b reduziert, wobei

$a \leq b \leq c$ gelten soll. Dementsprechend wird die Menge der möglichen Dreiecke auf einen euklidischen, zweidimensionalen Raum abgebildet. Weiter werden dann einfache, zweidimensionale Operatoren eingesetzt werden, um die Anzahl der Dreiecke z.B. durch Tiefpaßfilter zu reduzieren. Zur weiteren Beschleunigung des Backpropagation-Lernvorgangs wird jeder Synapse eine eigene Lernrate zugeordnet. Mit Hilfe des dargestellten Verfahrens konnten handgeschriebene Ziffern erkannt werden.

Ein interessanten Objekterkennungsansatz unter Verwendung autoassoziativer Speicher wird in [ANDO 1996] vorgestellt. Dort werden nicht wie allgemein üblich Merkmalswerte miteinander verglichen, sondern es wird versucht, ein gegebenes Objekt so zu verändern, daß es mit dem Szenenbild übereinstimmt. Das System iteriert zwischen zwei Prozessen: dem Generieren eines Bildes und dem Vergleich dieses Bildes mit dem Eingabebild. Innerhalb der Vorverarbeitung findet nur eine Gauß-Filterung statt, wobei eine vollständige Objekt-Hintergrund-Trennung vorausgesetzt wird. Das so geglättete Bild wird auf ein regelmäßiges Raster abgebildet. Als Assoziativspeicher kommt ein neuronales Netz mit fünf Ebenen zum Einsatz, um so nichtlineare Dimensionalitätsreduktionen ausführen zu können. Genaueres über die Verbindungsstruktur ist nicht zu erfahren. Jedes Modellobjekt wird durch ein Assoziativ-Netz repräsentiert. Während der Klassifikationsphase treten alle Netze in Konkurrenz zueinander. Dasjenige Netz mit der besten Übereinstimmung mit dem Eingabebild bestimmt das Matchergebnis. Kann kein Match gefunden werden, so wird das Raster des Bildes verändert und damit das Bild deformiert. Für die Raster-Veränderung werden zwei Verfahren eingesetzt: zum einen wird versucht, die Schwerpunkte der Objekte in Übereinstimmung zu bringen, was einer Verschiebung des gesamten Objekts gleichkommt; zum anderen werden unabhängige Verschiebungen einzelner Rasterpunkte durchgeführt - begrenzt nur durch ein *smoothness constraint*, welches zu große Deformationen verhindern soll. An Beispielszenen konnten gute Ergebnisse mit diesem Ansatz erzielt werden. Allerdings werden nur in der Tiefe rotierte Abbilder zuvor gelernter Objekte repräsentiert. Weitere Invarianzen werden nicht gezeigt. Zu bemerken ist weiterhin, daß die gelernten Objekte in einem Rotationsabstand von 1° repräsentiert werden, ein nicht unerheblicher Aufwand bei größeren Objektdatenbanken.

3.6 Verfahren in aktiven Sehumgebungen

Im folgenden werden Ansätze beschrieben, die ihren Schwerpunkt auf die Ausnutzung der durch das aktive Sehen gegebenen Möglichkeiten zur Stabilisierung von Klassifikationsergebnissen legen. Neben den hier aufgeführten sei noch auf

[GÖTZE et al. 1996] im Kapitel 4.3 und [FISHER und MACKIRDY 1998] im Kapitel 3.3.4 hingewiesen.

Ein aktives Sehsystem, basierend auf präattentiver Merkmalsextraktion und attentiver Identifikation, wird in [GIEFING et al. 1991], [GIEFING et al. 1992b] bzw. [GIEFING et al. 1992a] vorgestellt (Abbildung 3.28 zeigt eine Systemübersicht). Objekte werden durch eine Menge von charakteristischen fovealen Ansichten repräsentiert, die durch einen überwachten Lernprozeß ausgewählt werden. Gewünschte auszuführende Kamerabewegungen (Sakkaden) werden in eine egozentrische *interest*-Karte eingetragen. Dabei wird zwischen *bottom-up*-Sakkaden (bestimmt durch Merkmalspunkte) und *top-down*-Sakkaden zur Objekthypothesenvalidierung unterschieden. Durch Wettbewerbsprozesse wird die endgültige Sakkade bestimmt. Im Gegensatz dazu verhindert eine Hemmungskarte die Rückkehr zu bereits untersuchten Szenenbereichen.

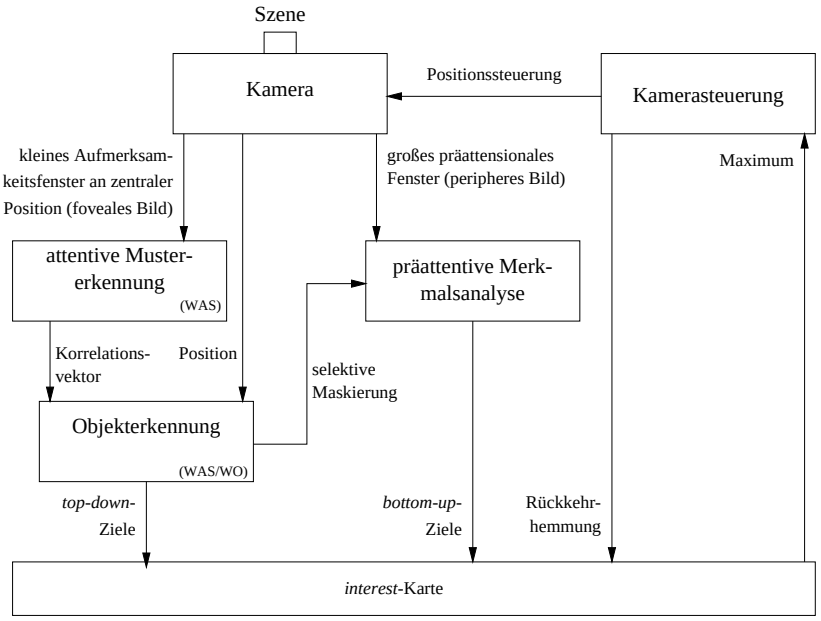


Abbildung 3.28: Systemübersicht (nach [GIEFING et al. 1992a])

Zur Objektcharakterisierung kommen folgende geometrische Merkmale zum Einsatz:

- Linien,
- gekrümmte Konturen,
- Kreuzungspunkte

Es wird die besondere biologische Relevanz durch ihre Entsprechung mit beobachtetem Verhalten der kortikalen Region V1 hervorgehoben. Objekthypothesen werden unter Verwendung eines Korrelationsverfahrens aufgestellt, welches das aktuelle foveale Bild mit allen gespeicherten fovealen Mustern vergleicht. Hierbei wird sowohl die Identität als auch die Position des Objekts gespeichert. Diese Informationen werden zeitlich in einem "Objektakkumulator" aufintegriert, der schließlich zur Ableitung einer Objekthypothese verwendet wird. Die Validierung erfolgt durch das "Anfahren" typischer, gelernter Objektkoordinaten.

Das Erkennungssystem wird für die Erkennung von 2D-Objekten getestet, wobei auch gestörte Muster, Distraktoren und Verdeckungen einbezogen werden. Eine Erweiterung auf eine 3D-Objekterkennung soll durch eine stereobasierte Tiefenbestimmung möglich werden. Der Ansatz zeichnet sich durch einen sehr durchdachten Aufbau und einer direkten Aktorik-Einbeziehung aus. Leider fehlt der Nachweis über die Funktionalität durch entsprechende Beispielbilder o.ä. quantitative Aussagen.

In [RYBAK et al. 1994] wird ein Erkennungssystem vorgestellt, welches auf einer *scanpath*-Analyse aufbaut, d.h. auf der Analyse einer Sequenz von Objektmerkmalen und Bewegungen zum Erreichen der nächsten Merkmalsposition [NOTON und STARK 1971]. Basiselement ist ein "retinales Bild", welches durch den augenblicklichen Aufmerksamkeitspunkt gegeben ist. Hier werden primäre Merkmale durch neuronal orientierte Verarbeitungseinheiten generiert. Zum Einsatz kommen orientierte Kantenelemente. Diese werden im Zentrum des retinalen Bildes (am sog. Basispunkt) und an einigen ausgewählten anderen Positionen (an sog. Kontextpunkten) extrahiert. Die Lage der Kontextpunkte soll die unterschiedlichen Auflösungseigenschaften der natürlichen Retina widerspiegeln (konzentrische Anordnung um den Basispunkt mit größer werdender Distanz). Die Menge der primären Merkmale wird in eine Menge invarianter Merkmale zweiter Ordnung transformiert. Eingesetzt werden relative Orientierungen und relative Winkel in Bezug zu einer im Zentrum des retinalen Bildes sich befindenden Grauwertkante.

Das gesamte Modell kann in drei verschiedenen Modi arbeiten:

1. Lernen von Bildern
2. Bildsuche
3. Bilderkennung

Beim Lernen von Bildern werden die Vektoren der invarianten Merkmale sukzessiver Fixationspunkte in einem neuronalen Netz gelernt (*Was*-Verarbeitungszweig). Die Position des nächsten Fixationspunktes wird aus der

Menge der Kontextpunkte abgeleitet. Jede relative Verschiebung der Aufmerksamkeit wird als ein invariantes Merkmal im sog. "Motorspeicher" abgelegt (*Wo*-Verarbeitungszweig). Als Ergebnis dieses Modus entsteht eine *Was*-Struktur mit Sensorinformationen und eine *Wo*-Struktur mit motorischen Informationen.

Im Modus Bildsuche werden die generierten *Was*- bzw. *Wo*-Fragmente mit den gespeicherten Strukturen verglichen. Bei Übereinstimmung wird eine entsprechende Hypothese gebildet und anschließend im Erkennungsmodus validiert.

Im Erkennungsmodus werden ausgehend vom Motorspeicher aufeinander folgende Verschiebungen der Aufmerksamkeit generiert, verbunden mit einem Vergleich von präsentierten und gespeicherten Merkmalen. Bei Übereinstimmung wird der *scanpath* während der Lernphase sukzessiv rekonstruiert. Ein Objekt gilt als erkannt, wenn eine Serie von Übereinstimmungen erzielt werden konnte.

Das System wird anhand der Erkennung von drei Gesichtern erfolgreich demonstriert. Dabei konnten auch Invarianzen bezüglich Verschiebung, Rotation in der Ebene und Skalierung nachgewiesen werden. Doch bleibt die gesamte Darstellung des Erkennungssystems sehr oberflächlich. Die sehr robuste Klassifikation der Gesichter ist nicht ohne weiteres nachvollziehbar. Auch fehlt die direkte Anbindung an Aktorik. Nichtsdestotrotz wird die Struktur eines typischen aktiven Sehsystems mit einer äußerst interessanten Repräsentationsform angegeben.

In [RAO und BALLARD 1995] wird ein vollständiges 3D-Objekterkennungssystem unter Verwendung eines aktiven Sehsystems vorgestellt. Ebenso wie später der Ansatz aus [GÖTZE et al. 1996] wird die Objekterkennung als zweistufiger Prozeß modelliert: zunächst Bildung einer Hypothese über ein Objekt an bestimmten Positionen und eine anschließende Verifikation des Objekttypes. Als Basismerkmale kommen auch hier Detektorfilter für verschiedene Skalierungsstufen und Winkelauflösungen zum Einsatz. Es werden Gaußfilter verschiedener Ordnungen eingesetzt und deren neurophysiologische Plausibilität hervorgehoben. Aus dem Ergebnis der Faltungen wird ein *iconic index* abgeleitet, d.h. ein normalisierter 45-dimensionaler Merkmalsraum aufgebaut. Diesen äußert großen Suchraum begründen die Autoren folgendermaßen: "*The relatively large number of measurements at an image point makes its characteristic vector practically unique*" und "*... a representation of an object of interest in the form of a high-dimensional vector can be subjected to considerable noise before it is confused with the vectorial representation of other objects.*". Pro Objekt wird nur eine geringe Anzahl von Merkmalspunkten verwendet. Ausschlaggebend zur Auswahl sind folgende Verfahren:

- Pro Objekt werden durch den Schwerpunkt virtuelle Linien gezeichnet. An den Schnittpunkten mit im Radius exponentiell größer werdenden Kreisen um den Schwerpunkt werden die Merkmalswerte ausgelesen (siehe Abbildung 3.29).

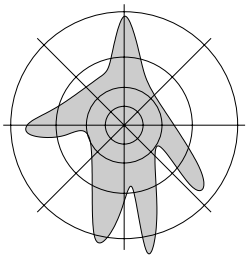


Abbildung 3.29: Verteilung der Merkmalspunkte auf einem Beispielobjekt (nach [RAO und BALLARD 1995])

- Unter Verwendung der Filterantworten werden alle Punkte eines Objekts hinsichtlich ihrer Besonderheit (*saliency*) untersucht. Dazu werden die Quadrate der Filterergebnisse über alle Orientierungen, Filterordnungen und Skalierungen aufsummiert. Dieses Merkmal wird als *photometric energy* bezeichnet und nach einer Schwellwertbehandlung werden die sich so ergebenden Punkte als Objektpunkte gespeichert.

Eine 3D-Erkennung wird durch die Verwendung mehrerer charakteristischer Ansichten pro Objekt realisiert. Durch die Verwendung eines binokularen Kamerasystems wird eine Objekt-Hintergrund-Trennung durchgeführt. Verwendet wird hierfür eine Disparitätsanalyse der untersuchten Szene. Zur Steuerung der motorischen Freiheitsgrade (Lösung der inversen Kinematik) des aktiven Sehsystems wird eine selbständig gelernte *motor map* nach Kohonen eingesetzt. Die beiden Teilprozesse Objektlokalisierung und -identifizierung gestalten sich folgendermaßen:

- Objektlokalisierung: Es werden die Repräsentanten, d.h. die Merkmalsvektoren des zu suchenden Objekts, gegen alle Objektpositionen gematcht. Dazu wird ein Distanzmaß basierend auf einer normalisierten Korrelation definiert. Eine globale Suche nach maximaler Übereinstimmung für alle Merkmalsvektoren an allen Bildpunkten ergibt dann die Position des gesuchten Objekts.

- Objektidentifizierung: Hier werden Szenenmerkmalsvektoren gegen alle gespeicherten Modellvektoren gematcht. Von dem zu matchenden Objekt werden die Merkmalsvektoren extrahiert und geprüft, welche Modellobjekte einen Szenenvektor in ihrer Merkmalsvektormenge besitzen. Entsprechend wird ein Zähler über die matchenden Vektoren pro Modellobjekt geführt und daraus die Klassifikationsentscheidung abgeleitet.

Der enorme Berechnungsaufwand zum Aufbau des 45-dimensionalen Merkmalsraumes und zum Matchen der Modellvektoren gegen alle Szenenpositionen ist nur durch den Einsatz spezieller Pipelinerechner handhabbar. Damit konnten laut Aussage der Autoren echtzeitnahe Bearbeitungszeiten erreicht werden. Der Lokalisierungsalgorithmus konnte auch in stark strukturierten Umgebungen Objekte zuverlässig finden. Bei Verwendung entsprechend vieler Punkte (bis zu 25) pro Szenenobjekt konnten sehr gute Erkennungsergebnisse erzielt werden. Auch das erfolgreiche Tolerieren von Verdeckungen konnte gezeigt werden. Invarianzen bezüglich Skalierungen von 5-10% konnten ebenfalls nachgewiesen werden. Bei größeren Skalierungen wird die Benutzung von Zoom-Objektiven vorgeschlagen. Weitergehende Invarianzen bezüglich Rotation in der Ebene bzw. im Raum werden nicht explizit erwähnt.

In [FUKUNAGA et al. 1996] bzw. [ONISHI et al. 1998] wird sich dem Problem der nicht eindeutigen Erkennung von Objekten unter Verwendung einer einzigen Ansicht gewidmet. Dementsprechend wird die Verwendung mehrerer aufeinander folgender Ansichten favorisiert, wobei die Objektwahrscheinlichkeiten aufintegriert werden. Modellansichten werden aus 3D-CAD Daten generiert. Jede Modellansicht wird um einen Koordinatenvektor erweitert, der die verschiedenen Koordinatensysteme (Welt-, Kamera- und Objektkoordinatensystem) in Beziehung setzt. Klassifiziert wird mit einem wahrscheinlichkeitstheoretischen Ansatz, der die Zugehörigkeit eines Merkmals zu einem Objekt angibt. Als Merkmalsbasis werden räumlich ausgedehnte Kantenbilder verwendet. Diese Bilder, Potentialbilder genannt, bestimmen die Klassifikationsentscheidung für eine Ansicht unter Verwendung einer Methode der kleinsten Fehlerquadrate durch eine Analyse des Residuums der Unterschiede der Potentialbilder von Modell und Szene. Die Objektwahrscheinlichkeiten aufeinander folgender Ansichten werden durch die *Dempster*-Kombinationsregel verrechnet. Für die Bestimmung neuer Kamerapositionen werden Verfahren des aktiven Sehens (Aufmerksamkeitssteuerungen) vorgeschlagen. Dazu wird ein Aktions-Plan generiert, der durch Kamerabewegungen zu einer eindeutigen Klassifizierung führen soll.

In [HERBIN 1996] werden theoretische Grundlagen für eine Objekterkennung mit aktiven Sehsystemen untersucht. Es wird ein formales Modell auf einer abstrakten Basis entworfen und innerhalb einer Simulationsumgebung validiert. Es wird die Möglichkeit für aktive Sehsysteme hervorgehoben, neue, evtl. diskriminierendere Objektansichten durch Kamerabewegungen zu generieren. Es wird festgestellt, daß die Sequenz von Merkmalen für verschiedene, aufeinander folgende Ansichten charakteristisch ist. Dementsprechend muß jeweils entschieden werden, ob zu einem bestimmten Zeitpunkt die vorhandenen Daten zur Entscheidungsfindung ausreichen, oder ob neue Daten und damit neue Objektansichten ge-

neriert werden sollen. Damit ergibt sich das für aktive Objekterkennungssysteme typische Systemverhalten: generiere solange neue Daten, bis das präsentierte Objekte eindeutig erkannt werden kann. Die Handlungen sind somit von allen alten Systemzuständen abhängig. Die eigentliche Klassifikationsentscheidung wird durch einen wahrscheinlichkeitstheoretischen Ansatz geliefert. Berechnet wird die a posteriori Wahrscheinlichkeit für das Vorhandensein eines bestimmten Objekts. Als Merkmale werden Ecken für verschiedene Auflösungsstufen verwendet. Es wird jeweils zwischen drei grundlegenden Systementscheidungen unterschieden (siehe auch Abbildung 3.30):

1. Berechnung neuer Merkmale auf einer anderen Auflösungsstufe,

2. Bewegung der Kamera zur Generierung einer weiteren Objektansicht, und

3. Feststellung des Klassifikationsergebnisses, wenn a posteriori Wahrscheinlichkeit groß genug.

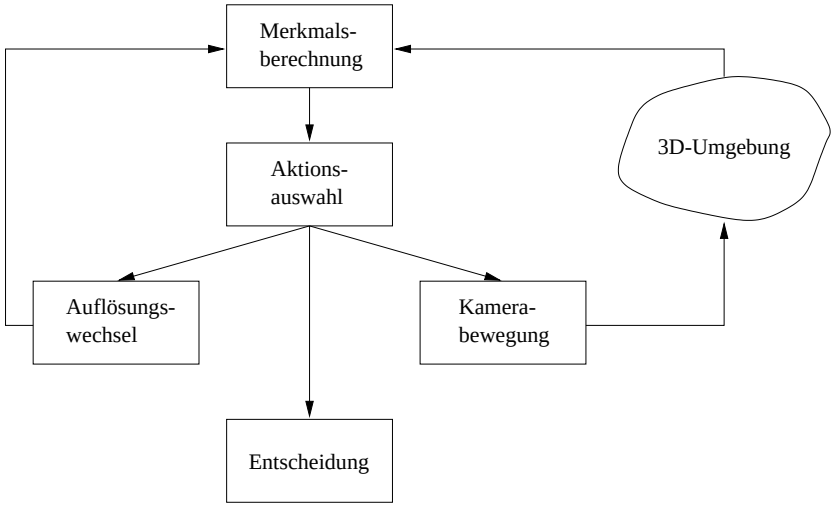


Abbildung 3.30: 3D-Objekterkennung in einer aktiven Sehumgebung (nach [HERBIN 1996])

Zu Demonstrationszwecken werden simulierte Schachfiguren ohne Hintergrund herangezogen. Eine simulierte Kamera kann sich dabei sowohl auf- und abwärts entlang der *view-sphere* bewegen. Die Länge und Richtung der Bewegungen ist zufällig. Eine seitliche Bewegung wird nicht mit einbezogen, da die eingesetzten Figuren rotationssymmetrisch sind.

Es konnte gezeigt werden, daß durch die selbständige Generierung neuer Daten stabilere Objekterkennungsergebnisse erzielt werden können. Allerdings beschränkte man sich nur auf Simulationen eines einfachen Erkennungsproblems. Trotzdem konnte die Überlegenheit aktiver Erkennungssysteme gegenüber klassischen passiven Systemen gezeigt werden.

Der Gedanke der selbständigen Generierung von weiteren Objektdaten wird in [SIPE und CASASENT 1997] weiter ausgebaut. Auch dort soll sich der Sensor bewegen, um nicht eindeutige Klassifikationen zu verbessern. Ebenso wird ein wahrscheinlichkeitstheoretischer Ansatz als Klassifikationsverfahren gewählt. Als Merkmale kommen die 10 maximalen Eigenwerte der Bilder zum Einsatz. Somit wird jedes Objekt als ein Punkt in einem zehndimensionalen Merkmalsraum betrachtet. Verschiedene Objektansichten ergeben mehrere solche Punkte, die durch Kanten verbunden werden, wodurch sich die grundlegende Repräsentationsform, die Merkmalsraum-Trajektorie (*feature space trajectory (FST)*), ergibt. Objektansichten stellen die Ecken dieser Trajektorie dar und Kamerabewegungen die Kanten. Zur Klassifikation wird der euklidische Abstand zu den einzelnen Objektansichten gebildet. Eine Wichtung findet dahingehend statt, inwieweit die präsentierte Ansicht sich in der Nähe von Entscheidungsgrenzen befindet. Diese Grenzen befinden sich im Merkmalsraum und repräsentieren die Bereiche, die zu mehreren Objektansichten den gleichen euklidischen Abstand im Merkmalsraum besitzen. Weiterhin wird eine Lageschätzung des präsentierten Objekts berechnet. Dazu wird die Lage zwischen den gelernten Ansichten ausgewertet und deren bekannten Lageparameter zur Interpolation der Unbekannten benutzt. Durch die globale Repräsentation aller Objekte und deren Ansichten besteht jetzt hier die Möglichkeit, auch die evtl. nötigen Kamerabewegungen genau zu bestimmen. Im Gegensatz zu [HERBIN 1996] wird also nicht nur bestimmt, ob eine Sensorbewegung und damit eine neue Objektansicht nötig ist, sondern auch die Art und der Umfang dieser Bewegung. Dazu wird wieder der Modell-Merkmalsraum ausgewertet. Abbildung 3.31 verdeutlicht die Konzepte.

Eine präsentierte Ansicht $\times$ befinde sich in der Nähe der Entscheidungsgrenze EG. Somit soll zur Validierung eine neue Ansicht des wahrscheinlichen Objekts #1 generiert werden. Gesucht wird nun eine gelernte Ansicht, die maximal von EG entfernt ist (hier Ansicht $20°$). Innerhalb der dargestellten Beispielumgebung wird ausschließlich eine Objektdrehung, deren Rotationsachse senkrecht zur Kameraachse liegt, modelliert. Obwohl die gewählten Beispielbilder relativ einfache Problemfälle behandeln, konnten doch die Vorteile eines autonom agierenden aktiven Objekterkennungssystems demonstriert werden. Besonders hervorzuheben ist die selbständige Suche nach gut unterscheidbaren Objektansichten. Die Autoren betonen ausdrücklich die besondere Eignung des Systems auch für industrielle Anwendungen, da hier eigenständig gelernt und agiert werden kann. Sowohl die

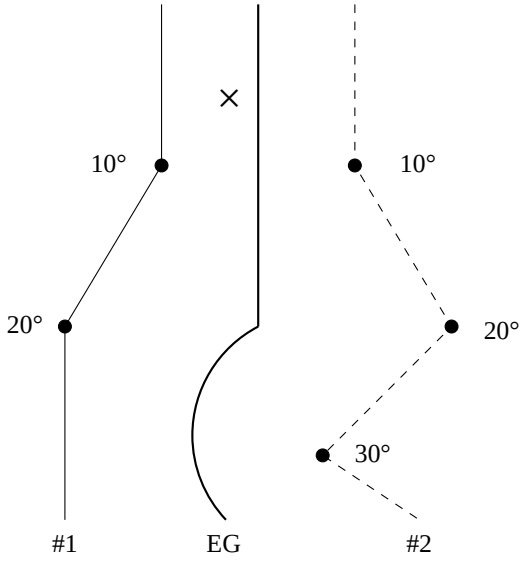


Abbildung 3.31: Merkmalsraum-Trajektorien für zwei Objekte #1 und #2 in einer hypothetischen Zwei-Klassen-Umgebung (Erläuterungen siehe Text; nach [SIPE und CASASENT 1997])

Erkennungsraten und die einhergehende Lageschätzung der Objekte waren überzeugend.

In [GVOZDJAK und LI 1998] wird ebenfalls die Bedeutung eines aktiven Verhaltens für künstliche Sehsysteme hervorgehoben. Zum Finden und Erkennen von Objekten werden verschiedene Objektansichten eingesetzt. Abbildung 3.32 zeigt die Architektur des Erkennungssystems. Zur Berechnung wird eine speziell entworfene Rechnerarchitektur eingesetzt: die *SFU hybrid pyramid vision machine* [ENS und LI 1995].

Das vorgestellte System soll selbständig Fakten über seine Umgebung durch eine aktive Exploration, verbunden mit einer Objekterkennung, sammeln. Es soll in der Lage sein, ein vorher spezifiziertes Objekt im dreidimensionalen Raum wiederzufinden. Zur Lösung dieser Aufgaben werden folgende drei Techniken verwendet:

- hierarchische *top-down*-Suche zur Verringerung des Suchraumes
- Einsatz von Roboter- bzw. Kamerabewegungen zum Finden von Ansichten
- Verwendung einer Mehrebenenauflösung für die Objekte zur Verringerung der zu bearbeitenden Datenmenge und zur Erhöhung der Robustheit

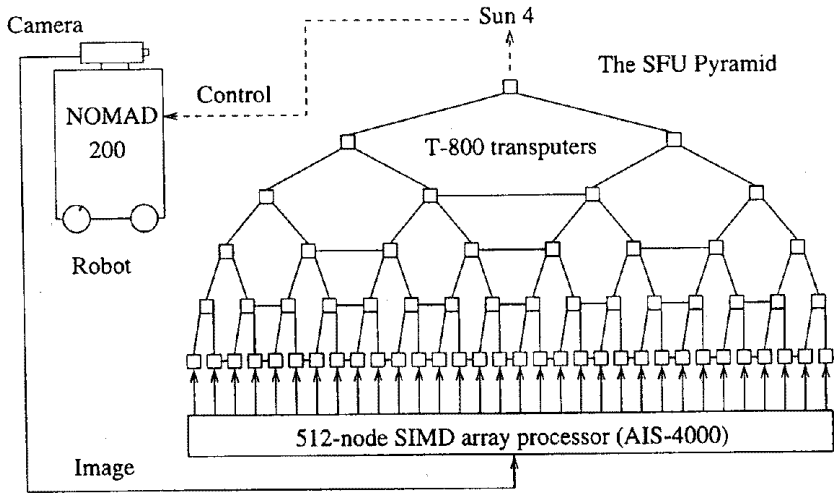


Abbildung 3.32: Architektur des Erkennungssystems (nach [GVOZDJAK und LI 1998])

Die Modelle der Objekte sind durch hierarchische Beschreibungen gegeben, die in einer Baumdarstellung repräsentiert werden. Erweitert wird diese Darstellung noch um Informationen zu den nötigen Kamerabewegungen, um eine bestimmte Ansicht zu erzielen. Jeder Knoten des Baumes repräsentiert Teile des Objekts, wobei hierarchisch höher liegende grobere und niedriger liegende feinere Objektstrukturen angeben. Bewertet sind die Knoten mit sog. Schlüsselaspekten, ihrer relativen Position zum Vorgängerknoten und die nötigen Kamerabewegungen. Mit Schlüsselaspekten werden bestimmte, typische Ansichten eines Objektes bezeichnet, die wesentlich zum Wiederfinden und zur Diskriminierung beitragen können. Abbildung 3.33 zeigt eine beispielhafte Objektrepräsentation.

Anstatt eine Übereinstimmung zwischen präsentierter Szene und der Ansichtendatenbank zu suchen (Suche im Modellraum), werden hier aktive Roboter- bzw. Kamerabewegungen verwendet, um die Schlüsselaspekte des gesuchten Objektes zu finden (Suche im Szenenraum). Für die zielgerichteten Bewegungen werden drei Bewegungsmodi verwendet:

- Rotation der Kamera um ihre Schwenkachse (siehe Abbildung 3.34a)
- Bewegung des Roboters in die Richtung (alternativ auch entgegengesetzt) des Objektes (siehe Abbildung 3.34b)
- Bewegung des Roboters um ein Objekt herum (siehe Abbildung 3.34c)

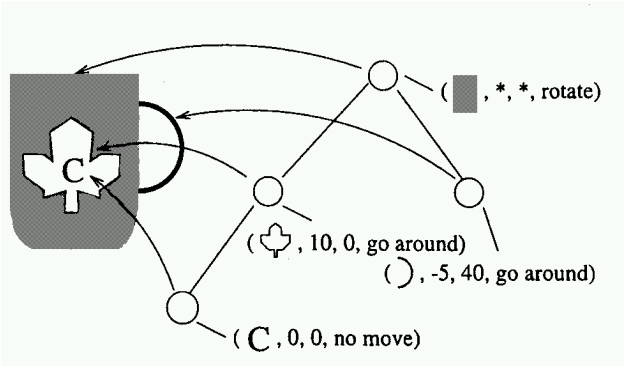


Abbildung 3.33: Repräsentation eines Objekts durch einen Baum (nach [GVOZDJAK und LI 1998])

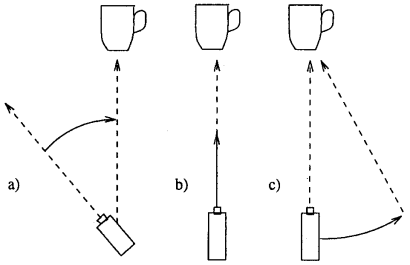


Abbildung 3.34: Verwendete Kamerabewegungen (Erklärungen siehe Text; nach [GVOZDJAK und LI 1998])

Roboter- bzw. Kamerabewegungen werden weiterhin unterstützend innerhalb des Erkennungssystems zur präzisen Erkennung der Objekte eingesetzt. Bildinformationen werden genutzt

- zum Halten des Objektes innerhalb des Bildes während einer Bewegung,
- zur groben Bestimmung der Größe des Objekts und seiner Entfernung,
- zum Beschränken der Suche auf kleinere Gebiete und
- zur Auswahl der passenden Schritte zur effektiven Erkennung.

Abbildung 3.35 gibt einen groben Überblick zur Erkennungsstrategie. Im folgenden werden die wesentlichen Schritte näher erläutert:

- *grobe Suche nach Objekt:* Die Kamera wird solange geschwenkt, bis ein Objekt durch die Repräsentation im Wurzelknoten (hierarchisch gröbste Auflösung, siehe auch Abbildung 3.33) erkannt werden konnte. Zum Matchen wird die verallgemeinerte Hough-Transformation eingesetzt.

- *Einstellen der richtigen Entfernung:* Besitzt das Objekt nicht die erwartete Größe, so bewegt sich der Roboter vor bzw. zurück. Dabei wird ständig das Zielobjekt durch Kamerabewegungen im Bild gehalten. Die ungefähre Größe des Objekts wird auf der Basis von Aufnahmen vor und nach einer Bewegung bestimmt. Kann die erwartete Größe durch Bewegungen nicht erreicht werden, wird die Erkennung abgebrochen und wieder in den Suchmodus übergegangen.

- *Suche nach Teilobjekten:* Entsprechend der Baumrepräsentation des Objektes wird nach Teilstrukturen gesucht. Die dazu evtl. nötigen Kamerabewegungen werden der Modellrepräsentation entnommen. Wieder kommt die verallgemeinerte Hough-Transformation zum Einsatz.

- *Iteriere:* Eine signifikant große Anzahl von Teilstrukturen gemäß der Baumrepräsentation muß wiedererkannt worden sein, bevor das gesamte Objekt als erkannt gilt. Ansonsten wird im Suchmodus fortgefahren.

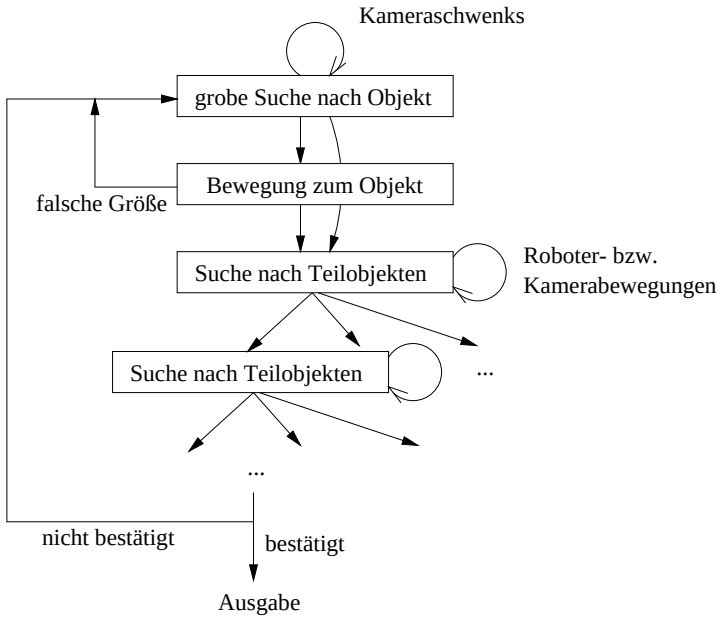


Abbildung 3.35: Ablauf der Erkennung (nach [GVOZDJAK und LI 1998])

Die Funktionsweise des Erkennungssystems wird anhand der Suche und Erkennung von verschiedenen Tassen in einer Büroszene demonstriert. Der vorgestellte Ansatz zeichnet sich durch ausführliche und nachvollziehbare Beschreibungen aus. Es wird die erfolgreiche Einbindung einer mobilen Plattform mit einer Kamera auf einer Schwenk-Neige-Einheit demonstriert und ihre Bedeutung für die Objekterkennung unterstrichen. Es konnte eine Repräsentationsform gefunden werden, die neben Objektinformationen auch die zu deren Erlangung notwendigen Roboter- bzw. Kamerabewegungen enthält. Insgesamt eine sehr überzeugende Arbeit bezüglich der Kopplung der Forschungsrichtungen Robotik und aktives Sehen.

Mit der Frage des Findens der nächsten Kameraposition zur Verbesserung eines Klassifikationsergebnisses wird sich u.a. auch in [LIU et al. 1998] befaßt. Hier wird Umfang und Richtung der Kamerabewegungen aus einem Gütekriterium der Merkmale abgeleitet. Zum Einsatz kommen Oberflächenmerkmale: der Flächeninhalt, die Kompaktheit und die Orientierung einer 3D-Objektoberfläche. Aus diesen Werten wird ein Maß zur Oberflächen-Unterscheidbarkeit entworfen. Dieses gibt das Vertrauen an, daß das Objekt durch eine bestimmte Oberfläche mit ihren Merkmalen eindeutig erkannt werden kann. Modelle werden durch eine objektzentrierte Darstellung ihrer Oberflächen repräsentiert. Zur Klassifizierung wird die Erkennung in ein Labeling-Problem umformuliert und mittels *Markov Random Fields* annähernd gelöst. Die Güte dieser Zuordnung wird innerhalb einer Terminierungsanalyse bestimmt. Es werden die Vertrauenswerte in die jeweiligen Klassifikationsentscheidungen miteinander verglichen. Gefordert wird ein vielfach höheres Vertrauen in die endgültige Objektzuordnung. Im negativen Fall wird eine Blickrichtungssteuerung benutzt, um neue Objektansichten zu gewinnen. Gewählt wird die Kameraposition, von der aus eine bisher nicht gematchte Oberfläche des wahrscheinlichsten Objekts mit maximaler Oberflächen-Unterscheidbarkeit sichtbar sein würde. Somit konnten die Klassifikationen verbessert werden. Für den vorgestellte Ansatz sind perfekt segmentierten Szenen mit einfachen polyedrischen Objekten Voraussetzung. Die gewählten Merkmale sind nur in diesen Fällen mit ausreichender Genauigkeit berechenbar. Somit ist der Anwendungsbereich dieses Verfahrens stark eingeschränkt.

3.7 Wertung der Systeme

Die Vielzahl der in diesem Kapitel vorgestellten Erkennungsverfahren mit ihren zum Teil sehr unterschiedlichen Herangehensweisen bezüglich Repräsentation oder Erkennung sollte die Bandbreite der Forschungs- und Entwicklungsaktivitäten aufzeigen. Im folgenden werden bestehende Schwachstellen aufgezeigt und

darauf aufbauend Kriterien für den Entwurf leistungsfähiger Erkennungssysteme abgeleitet.

Zunächst ist festzustellen, daß auch nach jahrelanger intensiver Forschung auf dem Gebiet der Objekterkennung z.Z. kein System in der Lage ist, auch nur annähernd die Leistungsfähigkeit biologischer Vorbilder zu erreichen. Die stärkere Betonung des Sehens als einen aktiven, intentionalen Prozeß eröffnet zwar neue, vielversprechende Wege, um Teilproblematiken bzw. Vorverarbeitungsebenen effektiver bzw. überhaupt erst lösbar zu gestalten, allerdings bleibt die Lösung der Objekterkennung als allgemein gefaßte Aufgabe (noch) Utopie. Nichtsdestotrotz sind die erreichten Erfolge der beschriebenen Erkennungssysteme bemerkenswert. In der für sie vorgesehenen Umgebung, in ihrem Ausschnitt aus der Wirklichkeit lassen sich erstaunliche Ergebnisse erzielen. Häufig wird die Ausdehnung dieser Anwendungsfelder im Ausblick der Arbeiten angesprochen. Formulierungen wie “ist prinzipiell möglich”, “kann einfach vorgenommen werden” oder “ist für die Zukunft beabsichtigt” sind Ausdruck der Zufriedenheit mit dem Erreichten und des Wissens um die fehlenden Aspekte, deren Lösung allerdings oftmals illusorisch erscheint.

In der Vergangenheit wurde bei der Bewertung von Erkennungssystemen oft die Meinung geäußert, daß mit dem entworfenen System zwar zur Zeit aufgrund einer sowohl rechenzeit- als auch speicherintensiven Strategie nur eine eingeschränkte Funktion gewährleistet werden kann, aber die zu erwartende Steigerung der verfügbaren Rechenleistung und Speicherintegration werde dem abhelfen. Natürlich geht die Leistung von Objekterkennungssystemen einher mit der Leistung der gerade aktuellen Rechnergeneration. Allerdings ist festzustellen, daß allein durch mehr Rechenleistung die Erkennung nicht prinzipiell auf weitere bzw. größere Anwendungsfelder ausgedehnt werden kann. Durch ein einfaches Vervielfachen der genutzten Objektbasen wird man nicht zum Ziel kommen.

Gerade in der Objekterkennung werden die eingesetzten Strategien gerne mit dem biologischen Vorbild Mensch verglichen. Das über Jahrmillionen gewachsene menschliche Sehsystem stellt das ultimative Erkennungssystem dar, dessen Leistungsfähigkeit auch in der nahen Zukunft unerreicht bleiben wird. Sich dem biologischen Vorbild zu nähern, würde aber bedeuten, die entsprechenden neuronalen Verarbeitungszweige von der Retina bis zur Großhirnrinde zu simulieren bzw. zu kopieren. Die Verwendung biologisch motivierter Vorverarbeitungsstufen allein reicht dazu nicht aus. Es fehlt (noch) an dem prinzipiellen Verständnis des Zusammenwirkens von Repräsentation und Erkennung, von Lernen und Erfahrung, von Wahrnehmung und Rückkopplung. Allerdings erscheint die Kombination der Vorteile beider Systeme, biologischer und technischer, ein gangbarer Weg zu sein. Die Technik bietet bestimmte Aspekte, wie z.B. lokale Rechenleistung, Sensoren, Optik u.a., in denen sie der Biologie weit überlegen ist. Eine Besinnung auf die Stärken aktueller Computertechnik, die sich aber an den Aufbau biolo-

gischer Systeme orientiert (biologische Plausibilität), sei es nun des binokulare menschliche Sehsystem, die Verwendung von sechs Beinen anstelle Rädern oder die Benutzung negativer Rückkopplungen in künstlichen neuronalen Systemen, führte in der Vergangenheit häufig zum Erfolg.

Das junge Forschungsfeld der aktiven Sehsysteme kann schon erstaunliche Erfolge vorweisen. Durch die Einbeziehung von Aktorik, die sowohl bildgetrieben (*bottom-up*) als auch aufgabengetrieben (*top-down*) gesteuert wird, lassen sich dynamisch agierende Erkennungssysteme realisieren. Häufig werden hierbei Schwenk-Neige-Einheiten, die auf mobilen Plattformen angebracht sind, eingesetzt. Solche Systeme sind in der Lage, interessierende Objekte selbständig zu finden, zu klassifizieren und u.U. auch selbständig zu lernen. Durch den hohen Rechenbedarf der Erkennungssysteme in Verbindung mit den harten Echtzeitanforderungen an aktive Sehsysteme sind bisher allerdings nur wenig leistungsfähige Erkennungssysteme realisiert. Es zeigt sich aber, daß die bisher realisierten Systeme den Weg weisen, um mit Erkennungsproblemen abseits einschränkender Bedingungen bezüglich Objektähnlichkeit, Beleuchtung u.ä. umzugehen. In diesem Zusammenhang sei aber nochmals auf das Fehlen von akzeptierten Benchmarks zur Evaluierung von aktiven Erkennungssystemen hingewiesen, um ein objektives Vergleichen verschiedener Ansätze bzw. Architekturen zu ermöglichen.

Im folgenden werden konkrete Kritikpunkte zusammengestellt, die für verschiedene Ansätze wiederholt festgestellt wurden. Es sind Beispielarbeiten aufgeführt, wobei zu beachten ist, daß jeweils nur wenige Arbeiten aufgeführt werden können und deren Nennung ihre Stellung als wissenschaftliche Arbeiten keineswegs in Frage stellen soll. Genauere Ausführungen zu den Ansätzen lassen sich den Kapiteln 3.3 bis 3.6 entnehmen.

- *Domänenbeschränkung:* Der häufigste einschränkende Faktor für Objekterkennungssysteme ist deren Einsatzfeld, z.B. eine Laborumgebung, kein störender Hintergrund, polyederförmige Objekte, vorhandene CAD-Modelle u.a. Meist werden Systeme durch die Wahl der eingesetzten Merkmale schon auf ein Anwendungsfeld festgelegt. Prinzipiell muß hier das Ziel sein, so viel wie möglich diskriminierende, höhere Merkmale zu generieren, aber redundante Informationen strikt auszusondern.

  Beispiele: [ALEXANDRE et al. 1991], [CHAO und DHAWAN 1992], [BASAK und PAL 1995], [BELLAIRE 1995], [NELSON 1995], [ANDO 1996], [LOURENS und WÜRTZ 1998]

- *Nutzereingriffe:* Oft muß der Nutzer Teilaufgaben eines Erkennungssystems übernehmen. Hierbei spielen vor allem Segmentierungsaufgaben eine Rolle. In der Entwurfsphase ist diese Herangehensweise sicherlich vertretbar.

Zum stabilen, autonomen Betrieb gehört aber gerade die selbständige Vorverarbeitung. In diesem Zusammenhang ist auch das Konzept des überwachten Lernens zu überprüfen. Für kleinere Objektbasen sicherlich vertretbar, stößt der menschliche Lehrer schnell an seine Grenzen, soll er Objektteile oder Merkmale auswählen, um ein Objekt von vielleicht Hunderten anderer diskriminieren zu können. Hier sollten Erkennungssysteme basierend auf ihrem Modellwissen selbständig handeln können.

Beispiele: [BURGER et al. 1996], [GVOZDJAK und LI 1998], [LIU et al. 1998]

- *Aufwand:* Schnell geraten Systeme an die Grenzen ihrer Leistungsfähigkeit, wenn mehr als nur ein paar Objekte überprüft werden sollen. Aufwendig zu berechnende Merkmale und iterative Erkennungsverfahren bedeuten hohe Rechenlast und damit lange Antwortzeiten. Gängige Wege zur Lösung sind die Unterteilung des Erkennungsprozesses in eine Hypothesen- und Validierungsphase bzw. der Einsatz einer hierarchischen Erkennungsstrategie, wobei aufwendige Prozesse nur bei Bedarf hinzugezogen werden können.

  Beispiele: [KOLLIAS et al. 1991], [COSTA und SHAPIRO 1996], [DICKINSON und METAXAS 1997]

- *Integration:* Viele realisierte Erkennungssysteme würden ihre Leistungsfähigkeit durch eine umfassende Integration in eine aktive Sehumgebung steigern können. Verfahren des aktiven Sehens bieten Möglichkeiten zur Reduzierung der nötigen Vorverarbeitungsschritte durch das Bereitstellen von Methoden z.B. der Aufmerksamkeit oder der stereobasierten Objekt-Hintergrund-Trennung. Auch die oben aufgeführten zu beobachtenden Restriktionen könnten dadurch abgeschwächt werden.

  Beispiele: [ESHERA und FU 1984], [LI und NASRABADI 1989], [BENNAMOUN und BOASHASH 1997]

- *Komplexität:* Bedingt durch das relativ frühe Entwicklungsstadium aktiver Sehsysteme sind häufig Objekterkennungsleistungen in realisierten Systemen nur rudimentär integriert. Meist werden nur sehr einfache, synthetische Grundmuster erkannt. Zur Demonstration der prinzipiellen Funktionsweise einer Objekterkennung in aktiven Sehsystemen ist dies sicher ausreichend. Allerdings muß es langfristig auch möglich sein, komplexere Erkennungsaufgaben zu lösen. Ausgangspunkt hierfür sind die erlangten Erfahrungen für die Erkennung in passiven Sehsystemen.

  Beispiele: [GIEFING et al. 1992a], [HERBIN 1996]

Aufbauend auf die im Kapitel 2.2 dargestellten aktuellen Forschungsaktivitäten im Bereich aktiver Sehsysteme sowie den in diesem Kapitel vorgestellten Objekterkennungssystemen und deren Wertung sollen allgemeine Kriterien für ein Erkennungssystem abgeleitet werden. Diese Anforderungen bildeten die Grundlage für den Entwurf der in den folgenden Kapiteln angegebenen Erkennungssysteme. Dabei soll ein in eine aktive Sehumgebung eingebettetes 3D-Objekterkennungssystem möglichst viele der in Tabelle 3.3 mit absteigender Priorität aufgeführten Kriterien erfüllen, wobei die Priorisierung jeweils in Abhängigkeit vom gewünschten Einsatzfeld zu sehen ist.

Tabelle 3.3: Entwurfskriterien für 3D-Objekterkennungssysteme (wird fortgesetzt)

Kriterium	Erläuterung
Erkennungs-leistung	Hauptfunktion eines Erkennungssystems ist natürlich die stabile, robuste Erkennung von gelernten Objekten.
Invarianzen	Entsprechend den Anforderungen an eine stabile Objekterkennung sollen möglichst viele Transformationen des Objektes in einer Szene (Translation, Rotation in der Bildebene und im Raum, Skalierung) immer noch zu einer korrekten Klassifizierung führen. Eine dementsprechende Vorverarbeitung bzw. Arbeitsumgebung mit angepaßten Repräsentationsformen können Objekttransformationen teilweise aufheben bzw. rückgängig machen und so die eigentliche Objekterkennung stark vereinfachen.
Einsatz-flexibilität	Der Einsatz des Systems sollte nicht auf eine spezifische Domäne eingeschränkt werden. Zwar ließen sich z.B. für industrielle Normteile dedizierte Merkmalsbasen entwerfen, allerdings ist eine Übertragung der gewonnenen Erfahrungen auf andere Problembereiche kaum möglich.
Auswahl Objekt-ansichten	Es sind Verfahren wünschenswert, die aus allen möglichen Objektansichten selbständig typische für den Lernprozeß auswählt, um so den Suchraum klein zu halten. Dementsprechend muß dann das Erkennungssystem über Generalisierungseigenschaften verfügen, um neue, unbekannte Ansichten den ähnlichsten gelernten zuordnen zu können.

Tabelle 3.3: Entwurfskriterien für 3D-Objekterkennungssysteme

Kriterium	Erläuterung
biologische Plausibilität	Die Sehsysteme von Lebewesen besitzen die am besten angepaßten, effektivsten, schnellsten und robustesten Erkennungssysteme. Eine hohe biologische Plausibilität von entworfenen künstlichen Sehsystemen versteht sich somit als Schritt, sich in Verbindung mit den Stärken moderner Rechentechnik auch der Leistungsfähigkeit der natürlichen Systeme anzunähern.
Okklusions-verarbeitung	Die Frage, inwieweit ein Erkennungssystem mit Verdeckungen umgehen kann, ist gerade beim Einsatz in natürlichen Szenen wesentlich. Es muß dabei stabil mit fehlenden Merkmalen bzw. Objektcharakteristika umgegangen werden können.
Parameter-festlegung	Objekterkennungssysteme haben immer parametersensible Komponenten. Allerdings sollte angestrebt werden, die Auswahl und Festlegung der Parameter zumindest domänenspezifisch a priori angeben zu können.
Speicher-effizienz	Trotz der immer höheren Integrationsdichten der Speichermedien sollte angestrebt werden, die Datenmenge pro gelerntes Objekt bzw. pro gelernter Ansicht so gering wie möglich zu halten. In diesem Zusammenhang muß auch das Kriterium der Erweiterbarkeit gesehen werden.
Geschwindigkeit	Gerade bei aktiven Sehsystem, die in einem dynamischen Umfeld agieren, ist der Geschwindigkeitsaspekt nicht vernachlässigbar. Das System sollte in der Lage sein, rechtzeitig auf sich ändernde Umweltfaktoren zu reagieren. Konträr dazu verlangen besonders die notwendigen Verarbeitungsschritte bei Objekterkennungssystemen enorme Rechenleistungen. Dementsprechend müssen solche Systeme u.U. mit dedizierter Hardware bzw. mit parallelen Implementationen aufgebaut werden.

Die in diesem Kapitel aufgeführten Erkennungsansätze lassen eine Reihe von erfolgversprechenden Entwicklungsrichtungen erkennen, um Systeme entsprechend obiger Tabelle zu entwerfen. Im folgenden werden einige Aspekte mit zugehörigen Arbeiten herausgestellt.

- *Merkmalsbasis:* Häufig wird eine orientierungs- und skalierungsspezifische Zerlegung von Grauwertkanten als Merkmalsbasis verwendet. Dadurch ent-

stehen zwar äußerst hochdimensionale Merkmalsräume, die entsprechend hohe Rechenzeitanforderungen mit sich bringen, allerdings konnten damit sehr gute Erkennungsergebnisse erzielt werden.

Beispiele: [RYBAK et al. 1994], [RAO und BALLARD 1995], [SHAMS 1995]

Weiterhin werden erfolgreich geometrische Primitive zur Objektbeschreibung verwendet. Solche Merkmale, wie Ecken, Kreisstrukturen o.ä., lassen sich bei entsprechender Vorverarbeitung gut detektieren und bieten Möglichkeiten zum robusten Aufbau von Objektbeschreibungen.

Beispiele: [BASAK und PAL 1995], [NELSON 1995], [COSTA und SHAPIRO 1996]

- *Repräsentation:* Gerne werden Graphenstrukturen eingesetzt, um das Potential der Verbindung einer effektiven Repräsentation mit optimierten Suchalgorithmen voll auszunutzen. Kann man dem Problem der möglichen kombinatorischen Explosion der Graphensuche begegnen, lassen sich schnelle und robuste Erkennungssysteme aufbauen.

  Beispiele: [BUNKE und MESSMER 1995], [CHUNG 1995], [BURGER et al. 1996]

  In diesem Zusammenhang sind auch die Bemühungen zum Aufbau von strukturellen Objektbeschreibungen zu sehen. Hierbei bieten sich Möglichkeiten zu einer effektiven Generierung von Beschreibungen und damit einhergehend einer direkten Interpretierbarkeit.

  Beispiele: [SHAMS 1995], [WILLIAMSON 1996b], [BENNAMOUN und BOASHASH 1997]

- *Erkennung:* Als Erkennungsalgorithmus werden u.a. Deformationsverfahren eingesetzt, um einen schrittweisen Vergleich mit Modellobjekten durchzuführen. Damit können auch komplexere Erkennungssituationen erfolgreich bearbeitet werden. Die dabei auftretenden hohen Rechenlasten können durch geeignete Startbedingungen beschränkt werden.

  Beispiele: [SCHWARZINGER et al. 1995], [ANDO 1996], [DICKINSON und METAXAS 1997]

Kapitel 4

Aktives Sehen im *N*euronalen *A*ctive *V*ision *S*ystem NAVIS

Die Arbeitsgruppe IMA entwickelt seit 1994 gemeinsam mit der Arbeitsgruppe von Herrn Dr. S. Drüe an der UGH Paderborn das aktive Sehsystem NAVIS (*N*euronales *A*ctive *V*ision *S*ystem), welches in der Lage sein soll, interessierende 2D- und 3D-Objekte im Raum unter Einsatz verschiedener Invarianzleistungen zu lokalisieren und zu erkennen ([DRÜE et al. 1994], [MERTSCHING et al. 1998], [MERTSCHING et al. 1999]). Entwicklungsziele im NAVIS-Projekt sind das Schließen von Forschungslücken gemäß der im vorigen Kapitel aufgeführten Forschungsdesiderate. Dabei wird ein strikt modularer Aufbau angestrebt, um unabhängig verschiedene Aspekte des Entwurfs aktiver Sehsysteme untersuchen zu können. Es werden neuroinformatische Modelle zur visuellen Informationsverarbeitung von Lebewesen erarbeitet, die gestützt auf Erkenntnisse aus der Psychophysik und der Neurophysiologie der visuellen Wahrnehmung für die Entwicklung von technischen Bildanalysesystemen genutzt werden. Aktuelle Forschungsfragen in diesem Zusammenhang sind die Fovealisierung von Objekten, die 2D- und 3D-Objekterkennung, Farbkonstanzuntersuchungen, Kameramodellierungen, das Korrespondenzproblem in der Stereoskopie bei räumlich verteilten Merkmalsrepräsentationen und Bewegungssehen. Hierfür wurden zwei experimentelle Plattformen zur Bildaufnahme aufgebaut, die abhängig von der Aufgabenstellung und den Auswertungsergebnissen geregelt werden können. Im Rahmen des NAVIS-Projektes sind eine Reihe von Arbeiten mit dem Schwerpunkt der Objekterkennung und angrenzender Themen, wie Stereovorverarbeitung und Aufmerksamkeitssteuerung, durchgeführt worden. Nach einer Übersicht über die vorhandenen experimentellen Plattformen folgt eine kurze Vorstellung der durchgeführten Projekte.

4.1 Experimentierumgebungen

Die primäre Experimentierplattform besteht aus einem mit zwei Farbkameras ausgestatteten Stereo-Kamerakopf. Er wurde an der Universität Paderborn unter besonderer Berücksichtigung von umfassenden Kalibrierungsmöglichkeiten und niedrigen Kosten entworfen. Eine detailierte Beschreibung des Designs und der Kalibrierung ist in [DRÜE und TRAPP 1996] zu finden.

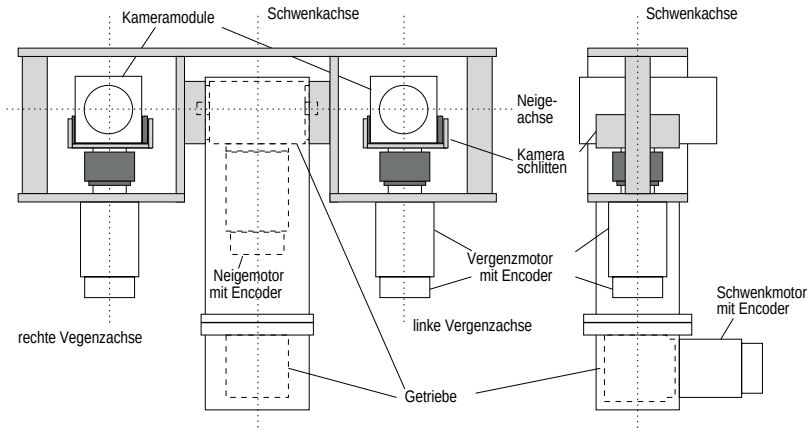
Die Kameramodule können um unabhängige vertikale Vergenzachsen rotieren und sind auf einer Schwenk-Neige-Einheit montiert. Dadurch werden insgesamt vier Freiheitsgrade zur Simulation von Augen- und Halsbewegungen bereitgestellt. Abbildung 4.1a zeigt eine Skizze der Vorder- und Seitenansicht des Kamerakopfes. In Abbildung 4.1b ist der reale Kopf abgebildet.

Jedes Kameramodul besitzt jeweils noch weitere drei Freiheitsgrade in der Befestigung auf der Schwenk-Neige-Einheit. Dadurch bestehen weitergehende Kalibrierungsmöglichkeiten, um die optischen Zentren der Kameras im Schnittpunkt zwischen Vergenz- und Neigeachse zu plazieren. Intern besitzen die Kameramodule nochmal drei Freiheitsgrade zur Kalibrierung. Die Kameras können in ihrem Gehäuse rotiert werden, um die optischen Ebenen in Übereinstimmung zu bringen. Zusätzlich kann der Nick- und Gierwinkel der Kameras verändert werden, damit die Sensorebene senkrecht auf der optischen Achse steht.

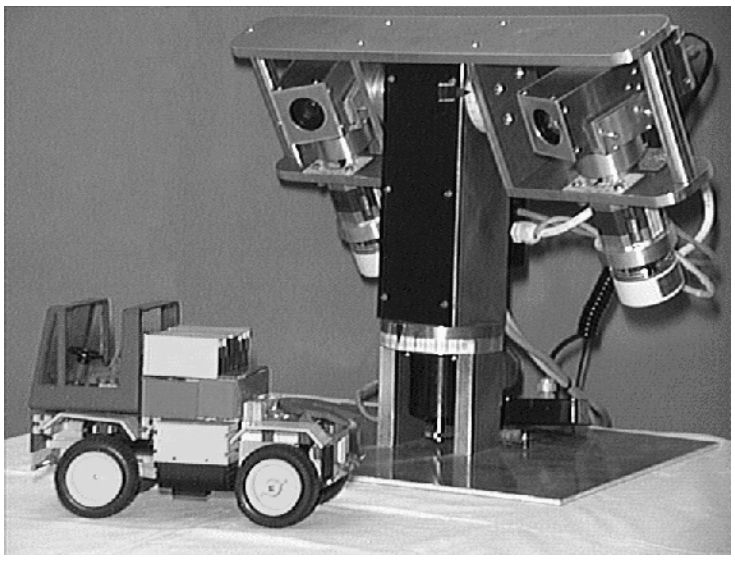
Ein mobiler Roboter Pioneer 1 dient als zweite experimentelle Basisplattform. Das Basis-Fahrzeug besitzt einen differentiellen Antrieb, acht Sonarsensoren, bereits installierte Steuerungsbausteine und eine Greifvorrichtung mit einem Freiheitsgrad. Das Gerät wurde hinsichtlich der Anforderungen an ein aktives Sehsystem mit einer Schwenk-Neige-Einheit und einer darauf montierten Farbkamera erweitert. Sowohl die Bilddaten- als auch die Steuersignalübertragung für das Fahrzeug und die Schwenk-Neige-Einheit wurde durch eine Funkübertragung realisiert. Somit konnte ein komplettes autonomes System aufgebaut werden (siehe auch Abbildung 4.2). Weitere Details über das mobile Fahrzeug und die eingesetzte Softwarearchitektur können z.B. in [GUZZONI et al. 1997] gefunden werden.

4.2 Merkmalsextraktion

In den folgenden Unterkapiteln werden die NAVIS-typischen Vorverarbeitungsmodule vorgestellt (siehe auch [MERTSCHING und BOLLMANN 1997], [BOLLMANN 1999]). Diese kommen zum Teil in den nachfolgend beschriebenen Erkennungssystemen zum Einsatz.



a)



b)

Abbildung 4.1: Der Stereokamerakopf: *a)* Skizze, *b)* Realisierung

Orientierte Kantenelemente

Zur Generierung orientierter Kantenelemente werden Gaborfilter eingesetzt, da diese durch ihr minimales Ortsfrequenz-Bandbreitenprodukt optimale Eigenschaften bezüglich Rechenzeit und Lokalität besitzen. Die Filterkerne werden im Frequenzraum beschrieben durch [TRAPP 1996]:

$$G(k) \;=\; e^{-\frac{1}{2}(k-k_0)^T\,(\mathcal{A}^{-1})^T\,(k-k_0)}. \tag{4.1}$$

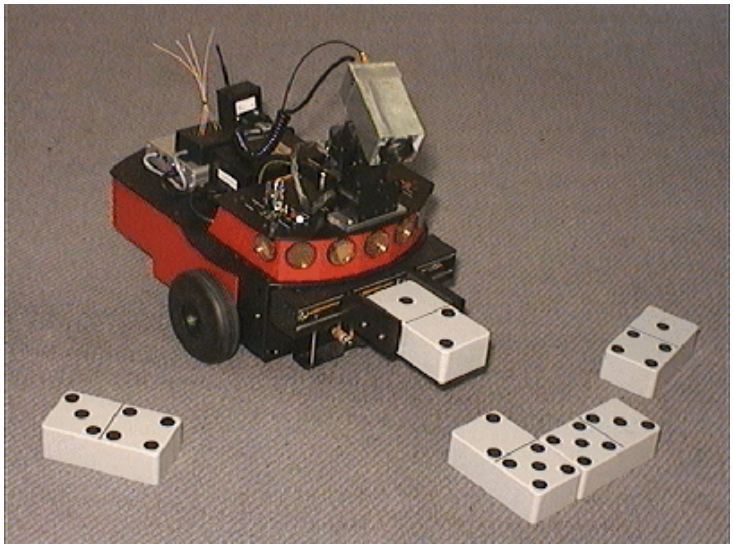


Abbildung 4.2: Das mobile Fahrzeug

Durch den Parameter k_0 wird die Schwerpunktsortsfrequenz angegeben mit $k_0 = \begin{bmatrix} |k_0| \cos\phi \\ |k_0| \sin\phi \end{bmatrix}$ und damit die zu suchenden Kanten gemäß ihrer Orientierung beschrieben. Die Matrix $\mathcal{A}$ setzt sich aus einer Rotationsmatrix R und einer Parametermatrix P zusammen:

$$\mathcal{A} = \mathcal{R}\mathcal{P}\mathcal{R}^T = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} a^{-2} & 0 \\ 0 & b^{-2} \end{bmatrix} \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \quad (4.2)$$

Durch $\lambda = \frac{a}{b}$ wird das Verhältnis der Ausdehnung des rezeptiven Feldes in Längs- und Querrichtung beschrieben. ϕ bezeichnet die Drehung dieses Bereiches. Im NAVIS-Projekt werden im allgemeinen 12 Orientierungen unterschieden.

Flächenschwerpunkte

Aufbauend auf eine Gaborfilterung wird ein *Split-and-Merge*-Verfahren eingesetzt, um homogene Flächen zu detektieren. Dazu werden, basierend auf orientierte Grauwertkanten, hinsichtlich ihrer Grauwertvarianz homogene Regionen zusammengefaßt. Als Merkmalsposition wurde der Schwerpunkt der segmentierten Fläche gewählt. Die Orientierung der Flächen wurde nicht mit einbezogen, um sich hinsichtlich angestrebter Invarianzleistungen des Systems nicht einzuschränken.

Symmetriepunkte

Auf der Basis der oben beschriebenen Gaborfilterung werden symmetrische Bildstrukturen gesucht. Die Aktivierungen der orientierten Kantenelemente in den 12 Merkmalskarten werden ausgehend von einem Bildpunkt in orthogonaler Richtung aufsummiert. Dabei werden mehrere Radien unterschieden. Entsprechend der Anzahl der "passenden" Kantenelemente wird die Güte des Symmetriepunktes bewertet. Abbildung 4.3 zeigt schematisch das angewendete Verfahren.

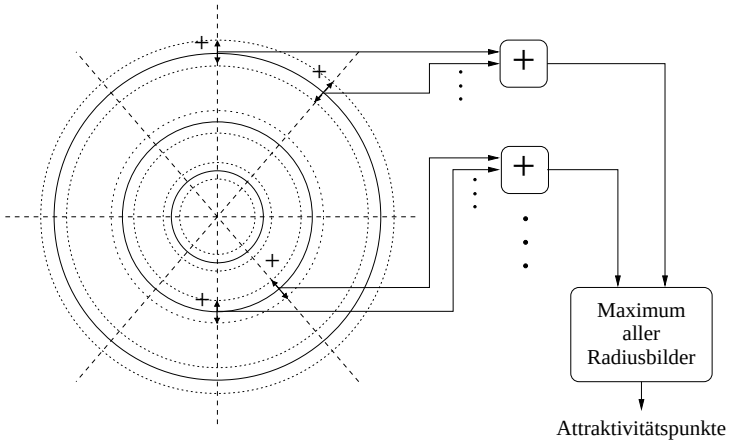


Abbildung 4.3: Berechnung symmetriebasierter Attraktivitätspunkte (nach [MERTSCHING und BOLLMANN 1997])

4.3 2D-Objekterkennung

In [GÖTZE et al. 1996] wird das innerhalb des NAVIS-Projektes eingesetzte 2D-Objekterkennungssystem vorgestellt. Es stellte den Ausgangspunkt für alle folgenden Entwicklungen (siehe Kapitel 5, 6 und 7) dar. Das Ziel war der Entwurf eines rückgekoppelten, aktiven Systems, welches sich durch eine hohe Einsatzflexibilität und biologische Plausibilität auszeichnet (für den kompletten Anforderungskatalog siehe Tabelle 3.3 im Kapitel 3.7).

Das Erkennungssystem basiert auf durch die physiologische und psychologische Forschung am visuellen System von Primaten aufgestellten Modellen zu einer mehrstufigen Objekterkennung von Sensorinformationen (siehe Kapitel 3.1.2). Wie bereits in [RAO und BALLARD 1995] wird dazu explizit die Suche nach Objekten (*Wo* - Verarbeitungszweig) und die Klassifizierung von Objekten (*Was* - Verarbeitungszweig) unterschieden. Dazu werden zwei unterschiedliche Betriebsmodi eingeführt: die Hypothesengenerierungsphase und die Hypo-

thesenvalidierungsphase. Die besondere Bedeutung der selektiven Aufmerksamkeit wird durch Einführung entsprechender Funktionseinheiten (Attraktivitäts-Einheit, Aufmerksamkeits-Einheit) berücksichtigt. Desweiteren wird von einem überwachten Lernen ausgegangen, um dem System mitzuteilen, welche Teile (hier Formen genannt) eines Objekts typisch sind und somit gelernt werden sollen. Grundidee des Ansatzes ist die Annahme, daß sich jedes komplexe Objekt durch eine spezifische Konfiguration charakteristischer Punkte, d.h. höherer Merkmale, auszeichnet (siehe Abbildung 4.4). Zur Bildung einer Hypothese reicht dann die Klassifikation dieser Punktmenge aus. Anschließend erfolgt eine Validierung durch die Suche nach den in der Lernphase festgelegten charakteristischen Formen.

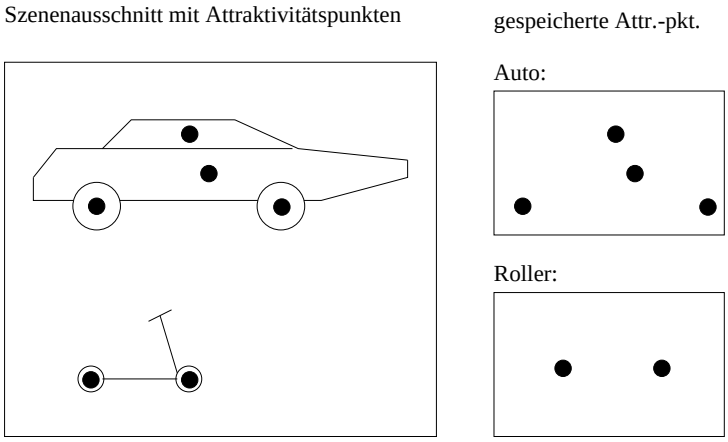


Abbildung 4.4: Konfigurationen charakteristischer Punkte

In Abbildung 4.5 findet sich ein allgemeiner Überblick des Gesamtsystems. Im einzelnen erfüllen die angegebenen Funktionseinheiten die folgenden Aufgaben:

Kamera-Einheit

Hier erfolgt die Steuerung der Schwenk-Neige-Einheit zur Ausrichtung auf durch das Erkennungs-Modul vorgegebene Positionen. Dabei kommt ein einfaches Verfahren der inversen Kinematik zum Einsatz: Jeweils die Pixelwerte eines Fensters um den gewünschten Zielpunkt und um den tatsächlich angefahrenen Punkt werden zeilen- und spaltenweise aufsummiert. Diese Vektoren werden korreliert und ergeben so eine evtl. nötige Verschiebung des nach der Kamerabewegung aufgenommenen Bildes.

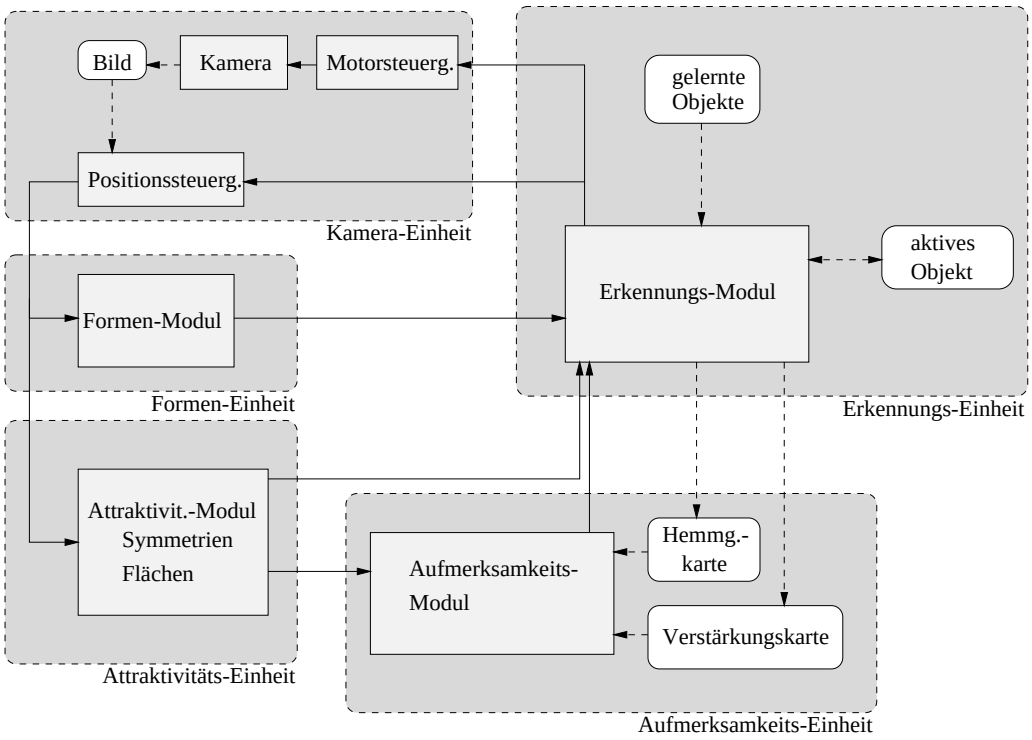


Abbildung 4.5: Das Erkennungssystem nach Götze und Mertsching (nach [GÖTZE et al. 1996])

Formen-Einheit

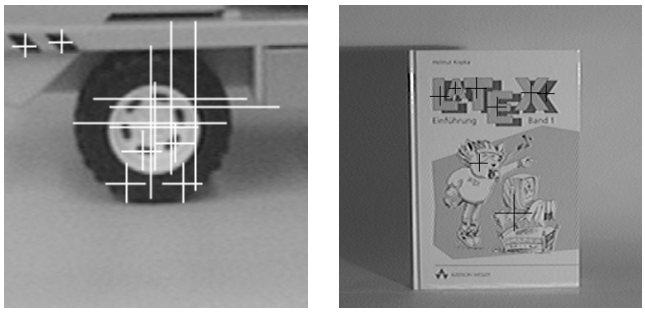
Hier werden während der Validierungsphase an durch die Erkennungseinheit angegebenen vermuteten Objektformenpositionen Kantenbilder generiert. Diese Kantenbilder werden entsprechend dem Verhalten der komplexen Zellen im Sehsystem von Primaten weiterverarbeitet, um eine höhere Ortstoleranz bereitzustellen (siehe Abbildung 4.6). Während der Hypothesengenerierungsphase hat dieses Modul keine Funktion.

Attraktivitäts-Einheit

Durch diese Einheit wird die gerade beobachtete Szene nach aufmerksamkeitswürdigen Punkten, sog. Attraktivitätspunkten, untersucht. Dabei kommen Verfahren zum Einsatz, die im Rahmen des NAVIS-Projektes entstanden sind. Im einzelnen wird die Szene nach auffälligen Symmetrien und homogenen Flächen untersucht (siehe Kapitel 4.2). Abbildung 4.7 zeigt die so generierten Attraktivitätspunkte.

a) b)

Abbildung 4.6: Gelernte Objektform *a)* und präsentierte *b)*



a) b)

Abbildung 4.7: Symmetriebasierte *a)* und flächenbasierte *b)* Attraktivitätspunkte

Aufmerksamkeits-Einheit

Hier wird aus der Menge der Attraktivitätspunkte p_{attr} ein Aufmerksamkeitspunkt p_{auf} ausgewählt (siehe Gleichung 4.3). Dazu werden zunächst die durch die Attraktivitätseinheit berechneten Gütewerte $g(p_{attr})$ mit den Einträgen in der Hemmungskarte $\mathcal{HK}$ und der Verstärkungskarte $\mathcal{VK}$ verrechnet und so die Menge der möglichen Aufmerksamkeitspunkte p^k_{auf} gebildet. Der Punkt mit maximaler Güte bestimmt dann den Aufmerksamkeitspunkt. Durch die Hemmungskarte wird für einen gewissen Zeitraum sichergestellt, daß sich das System nicht wieder auf die gleichen Punkte konzentriert. Auf der anderen Seite sorgt die Verstärkungskarte während der Hypothesenvalidierungsphase dafür, daß vermutete Objektpositionen bessere Chancen auf Aufmerksamkeit erhalten.

$$g(p^k_{auf}) \;=\; g(p_{attr}) \quad \cdot \prod_{p_{HK} \in \mathcal{HK}} (1 - g(p_{HK}) \cdot d_1(p_{attr}, p_{HK})) \cdot \prod_{p_{VK} \in \mathcal{VK}} (1 + g(p_{VK}) \cdot d_2(p_{attr}, p_{VK})) \tag{4.3}$$

$g(x)$	bezeichnet die Güte eines Punktes x in $[0,1]$
$d(x,y)$	bezeichnet ein Abstandsmaß für x und y in $[0,1]$ (1 bei identischen Positionen)

Erkennungs-Einheit

Innerhalb der Erkennungseinheit werden zunächst die berechneten Attraktivitätspunkte auf das Vorhandensein von gelernten Punktekonfigurationen überprüft. Basispunkt ist dabei der Aufmerksamkeitspunkt. Innerhalb der gelernten Konfigurationen wird nach typgleichen Punkten gesucht und die restlichen Punkte entsprechend verschoben. Die sich mit den Szenenpunkten ergebende Überlappung bezeichnet den Hypothesenmatchwert (siehe Abbildung 4.8). Sei $\mathcal{LAP}$ die Menge der gelernten Attraktivitätspunkte und $\mathcal{PAP}$ die Menge der präsentierten, dann bezeichnet das Maximum vom Verhältnis $\frac{match_{hypo}}{norm}$ für alle möglichen Punkte $p \in \mathcal{LAT}$ sowohl die vermutete Position des Objekts in der Szene als dessen vermeindliche Identität. $match_{hypo}$ stellt die distanzgewichtete Summe der matchenden Attraktivitätspunkte dar und mittels $norm$ wird eine Normierung in das Intervall $[0,1]$ gewährleistet:

$$match_{hypo} = \sum_{p_l \in \mathcal{LAP}\setminus\{p\}} g(p_l)\cdot g(p')\cdot d(p_l,p') \quad (4.4)$$

mit $p' \in \mathcal{PAP}$ und $d(p_l,p') \geq d(p_l,p'')\ \forall p'' \in \mathcal{PAP}$

$$norm = \sum_{p_l \in \mathcal{LAP}\setminus\{p\}} g(p_l)\cdot \begin{cases} g(p') & p' \in \mathcal{PAP} \text{ und } 0 < \\ & d(p_l,p') \geq d(p_l,p'')\ \forall p'' \in \mathcal{PAP} \\ g(p_l) & \text{sonst} \end{cases}$$

Nach der erfolgten Hypothesenbildung werden in der Hypothesenvalidierungsphase die in der überwachten Lernphase spezifizierten charakteristischen Objektteile durch den Kamerakopf[1] angefahren und entsprechend gematcht. Eine robuste Verarbeitung von Okklusionen wäre hier durch eine unabhängige Untersuchung der gelernten Formen denkbar. Beim Matchen kommen die verbreiterten Kantenstrukturen nach Abbildung 4.6 zum Einsatz. Es wird die Anzahl der gelernten Bildelemente mit der Anzahl der präsentierten in Beziehung gesetzt:

$$match = \frac{\sum_{Pixel} präsentiert \wedge gelernt}{\sum_{Pixel} gelernt} \quad (4.5)$$

[1] Es wird hier ausschließlich monokular gearbeitet, d.h. es werden nur Schwenk- und Neigebewegungen, aber keine Vergenzbewegungen verwendet.

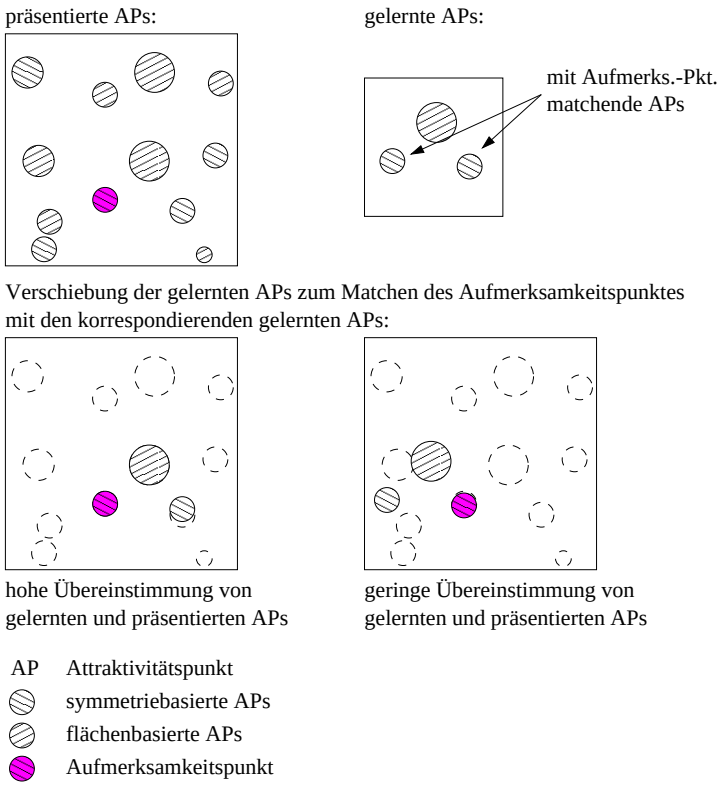


Abbildung 4.8: Generierung von Objekthypothesen (nach [GÖTZE et al. 1996])

Experimente und Ergebnisse

Abbildung 4.9 zeigt einen exemplarischen Erkennungsprozeß beginnend mit der Suche nach interessierenden Objekten bis hin zu deren Klassifikation. Es werden ein Spielzeugauto und ein Buch vor einem stark strukturierten Hintergrund (Tageszeitung) präsentiert. Die Objekte werden auf Grund ihres unterschiedlichen Kontrasts zum Hintergrund gewählt. Die einzelnen Erkennungsschritte sind in Tabelle 4.1 erläutert. Das System war in der Lage, stabil zunächst die zu klassifizierenden Objekte zu finden und die gelernten Objektteile sukzessiv zu matchen.

Probleme bereiten Änderungen der Objektausprägungen durch geringe Transformationen in der Ebene oder im Raum. Hier macht sich die starre Repräsentation der Attraktivitätspunkte und Objektformen negativ bemerkbar. Nachteilig wirkt sich ebenfalls die u.U. sehr große Anzahl von Szenenattraktivitätspunkten aus. Die auf Grund minimaler Invarianzleistungen sehr niedrig gewählten Erkennungsschwellen führen in diesem Fall zu Fehlklassifikationen. Bei den durchgeführten Experimenten war weiterhin zu beobachten, daß fast immer das Matchen der Objektformen die vorher gebildeten Hypothesen bestätigten. Hiermit ergibt sich ei-

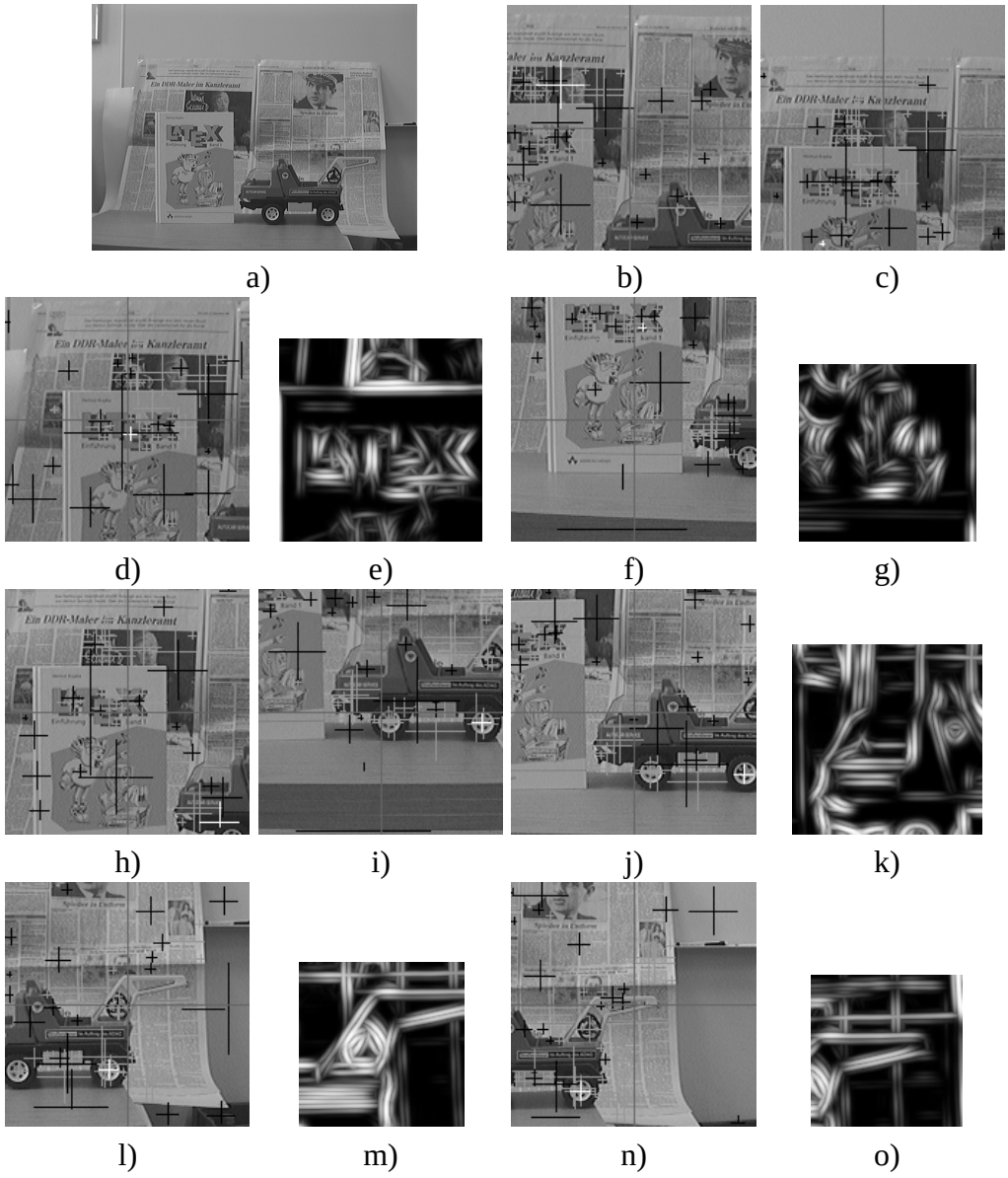


Abbildung 4.9: Erkennungsexperiment für zwei Objekte vor einem stark strukturierten Hintergrund (Erläuterungen siehe Text)

Tabelle 4.1: Erläuterung der Erkennungsschritte aus Bild 4.9

Schritt	Erklärung
a)	Experimentierumgebung mit zwei Objekten vor einer Tageszeitung
b)	Ergebnis der ersten Szenenanalyse durch das Erkennungssystem; Schwarze Kreuze stellen flächenbasierte Attraktivitätspunkte dar, graue Kreuze symmetriebasierte. Der Aufmerksamkeitspunkt ist als weißes Kreuz markiert. Ein Fadenkreuz verdeutlicht den Fokussierungspunkt der Kamera. Für das Szenenbild konnte keine Hypothese generiert werden. Diese Region wird im folgenden durch einen Eintrag in der Hemmungskarte inhibiert.
c)	Nachdem die Kamera sich auf den Aufmerksamkeitspunkt ausgerichtet hat, kann durch die anschließende Analyse die Hypothese für das Objekt “Latex-Buch” abgeleitet werden.
d)	Kamera auf das erste gelernte Objektteil (Titel) ausgerichtet
e)	das Konturbild kann mit dem abgespeicherten gematcht werden
f)	Kamera richtet sich auf das zweite Objektteil (Illustration) aus
g)	die Hypothese “Latex-Buch” kann endgültig durch das Matchen auch des zweites Objektteiles bestätigt werden
h)	das Erkennungssystem sucht nach neuen, noch nicht klassifizierten Objekten; Alle zum Objekt “Latex-Buch” gehörende Attraktivitätspunkte wurden in die Hemmungskarte eingetragen, so daß ein Aufmerksamkeitspunkt außerhalb dieses Objektes generiert wurde.
i)	Nach einer erneuten Fixation kann die Objekthypothese “Abschleppwagen” generiert werden. Dieses Objekt wurde durch die aktive Exploration der Szene überhaupt erst sichtbar.
j)	Kamera richtet sich auf das erste Objektteil aus
k)	das erste Objektteil kann gematcht werden; Ausrichtung auf das zweite
l)	das zweite Objektteil wird fixiert
m)	das zweite Objektteil wird gematcht
n)	Suche nach dem dritten Objektteil
o)	Die Objekthypothese “Abschleppwagen” kann nach dem erfolgreichen Matchen auch des dritten Objektteiles bestätigt werden.

ne interessante Fragestellung als Ansatzpunkt für alternative Systeme: Wenn die Hypothesengenerierung auf der Basis von Attraktivitätspunkten schon so sichere Entscheidungen liefert, warum sollte man dann noch die langwierige und aufwendige Prozedur des Matchens mit Kantenstrukturen durchführen? Dementsprechend ist festzustellen, daß die Verwendung von Punkten mit charakteristischen Eigenschaften (hier: Symmetrie- und Flächenschwerpunkte) ein großes Potential hinsichtlich einer robusten 2D-Objekterkennung in sich birgt.

In Bezug zu den in Tabelle 3.3 aufgestellten Anforderungen an Erkennungssysteme ist festzustellen, daß die Kriterien Erkennungsleistung, Einsatzflexibilität und biologische Plausibilität gut erfüllt wurden; eine Okklusionsverarbeitung ist leicht zu integrieren. Die hohe biologische Relevanz zeigt sich in der Einarbeitung folgender, aus der Neurophysiologie und Psychophysik bekannten Konzepte:

- mehrstufige Sensordatenverarbeitung: Die Trennung der Verarbeitung in einen *Was*- und *Wo*-Zweig, die für visuelle Systeme von Primaten nachgewiesen werden konnte (siehe Kapitel 3.1.2), wird durch die Einführung der Betriebsmodi Hypothesengenerierung und -validierung explizit in die Verarbeitungsstruktur übertragen.

- präattentive Merkmalsextraktion: Vor entsprechenden Kamerabewegungen wird das sichtbare Umfeld des aktuellen Fokussierungspunktes auf spezifische Merkmale untersucht.

- selektive Aufmerksamkeit: Zur Hypothesenvalidierung werden spezifische Szenenregionen sequentiell durch das Kamerasystem angefahren und analysiert.

- Hemmung untersuchter Regionen: Bereits analysierte Szenenbereiche werden durch eine zeitlich abklingende Hemmungskarte zunächst von weiteren Untersuchungen ausgeschlossen.

Defizite sind für die bereitgestellten Invarianzen, Parameterfestlegungen und Geschwindigkeit auszumachen. Es ist aber zu betonen, daß innerhalb der Entwicklung die Geschwindigkeit des Erkennungsvorganges nicht im Vordergrund stand. Vielmehr sollte eine Ausgangsbasis für nachfolgende Entwicklungen auf dem Gebiet der Objekterkennung geschaffen werden. Da das System als 2D-Erkennungssystem konzipiert war, ist keine Auswahl typischer Objektansichten und deren Verarbeitung nötig. Die Objekterkennung nach Götze und Mertsching stellt ein System dar, das sich ausdrücklich durch die Einbindung in eine aktive Sehumgebung definiert. Dadurch ist es in der Lage, selbständig auf Szenenveränderungen zu reagieren bzw. überhaupt zu klassifizierende Objekte zu finden. Weiterhin ist die im Hinblick auf Hintergrundstörungen sehr robuste Erkennung

positiv zu vermerken. Trotz der Verwendung sehr einfacher Verfahren zur inversen Kinematik, resultierend in nur geringen Geschwindigkeitseinbußen bzw. in der Möglichkeit des Einsatzes unkalibrierter Systeme, konnte eine pixelgenaue Anfahrgenauigkeit erreicht werden. Erstmalig ist im NAVIS-Kontext ein Erkennungssystem entstanden, das zwei wesentliche Komponenten aktiver Sehsysteme, Aufmerksamkeitssteuerung und Objekterkennung, in einem rückgekoppelten System mit Möglichkeiten zu Kamerabewegungen und zur Verarbeitung natürlicher Szenen vereinte.

4.4 Stereobildverarbeitung

Das Sehsystem der Primaten kann durch dessen binokularen Sehapparat im Nahbereich die Umwelt dreidimensional wahrnehmen. Dadurch sind präzise Bewegungen auch für kompliziertere Tätigkeiten möglich. Der in der Arbeitsgruppe vorhandene Stereokamerakopf bildet dieses binokulare Sehsystem nach (siehe Abbildung 4.1). Damit wird es möglich, die Szenenumgebung durch Gewinnung von 3D-Daten in ihrer Tiefenausdehnung zu rekonstruieren. Dies kann Basis für eine Objekt-Hintergrund-Trennung sein, wodurch der Suchraum für Objekterkennungsaufgaben erheblich eingeschränkt werden kann.

Die Stereobildverarbeitung stellt dabei aber nur eine Möglichkeit zur Szenenrekonstruktion dar. Zur Anwendung könnten prinzipiell auch folgende monokularen Ansätze kommen [PINZ 1994]:

- *Shape from Shading:* Aus den Helligkeitsverläufen in einem Bild kann auf die Oberflächenneigungen und damit auf die Tiefenstruktur der Szene rückgeschlossen werden. Als Nachteil ergibt sich u.a., daß mehrere Lösungen (Beleuchtungsrichtung und Oberflächenneigung) pro gegebener Szene angegeben werden können. Gerade in natürlichen Szenen mit diffuser Beleuchtung und ungleichmäßigen Grauwertverläufen führt dieses Verfahren oft zu nicht befriedigenden Ergebnissen.

- *Shape from Texture:* Hier wird aus der Veränderung von Objekttexturen auf die Neigung der Objekte rückgeschlossen. Basisannahme ist, daß zusammenhängende Szenenbereiche sich durch ähnliche Textureigenschaften bzw. Grauwertverteilungen auszeichnen. Aus den Unterschieden in der Grauwertstruktur benachbarter Regionen wird auf die Oberflächenneigung geschlossen. Ansätze dieser Art sind meist sehr rechenintensiv und im besonderen Maße von den gegebenen Voraussetzungen abhängig. Besitzen Objekte selbst eine sich ändernde Textur, so wird daraus eine nicht vorhandene Oberflächenneigung abgeleitet. Auf der anderen Seite müssen die

zu untersuchenden Szenen auch ausreichend detektierbare Texturen aufweisen - eine Annahme, die gerade für die üblichen Experimentierumgebungen nicht voll zutrifft.

- *Shape from Motion:* Werden Objekte innerhalb von Bildsequenzen verfolgt, kann deren Bewegungstrajektorie im Raum bestimmt werden. Dabei ist es prinzipiell auch möglich, daß sich die Kamera selbst bewegt und das Objekt starr angeordnet ist. Weiterhin lassen sich durch die Analyse rotierender Objekte 3D-Repräsentationen gewinnen.

- *Shape from Focus:* Hierbei werden die Szenenobjekte derart mehrmals aufgenommen, daß immer unterschiedliche Teilbereiche scharf abgebildet werden. Die einzelnen Tiefenebenen lassen sich dann durch eine Ortsfrequenzanalyse voneinander trennen.

- *Shape from Structured Light:* Die betreffenden Objekte werden mit strukturiertem Licht bestrahlt und aufgenommen. Je nach 3D-Oberflächenanordnung werden dabei die Lichtmuster verzerrt. Aus diesen Veränderungen kann die 3D-Form des Objekts bestimmt werden.

Unter Verwendung des Stereokamerakopfes wird innerhalb des NAVIS-Projektes mittels *Shape from Vergence* die Tiefenstruktur eines Objektes bestimmt. Dabei wird das Kamerasystem so gesteuert, daß beide Kameras das Objekt fixieren. Aus den sich ergebenden Vergenzwinkeln läßt sich auf die Entfernung schließen.

Abbildung 4.10 verdeutlicht die Abbildungsverhältnisse an einem Stereo-Kamerakopf. Beide Kameras fixieren hierbei einen Punkt, so daß dieser in beiden Bildebenen im Zentrum zu finden ist. Die Punkte P_1 bis P_3 werden jeweils im linken und rechten Bild verschoben abgebildet, womit sich folgende Disparitäten ergeben:

$$\begin{aligned} d_1 &= x_{l1} - x_{r1} > d_2 > 0 \\ d_2 &= x_{l2} - x_{r2} > 0 \\ d_3 &= x_{l3} - x_{r3} < 0 \end{aligned} \tag{4.6}$$

Die Analyse der Disparitäten für verschiedene Bildbereiche liefert eine Möglichkeit, die Szene bezüglich ihrer räumlichen Tiefe zu segmentieren, die Größe und die Entfernung von Objekten zu bestimmen. Aus dieser Segmentierung läßt sich eine Objekt-Hintergrund-Trennung ableiten. Im Rahmen der hier vorgestellten Arbeit soll zunächst nur die Tiefensegmentierung als Mittel zur Beschleunigung und zur Stabilitätserhöhung von Objekterkennungssystemen dienen. Die damit verbundene Objekterkennung wird im Kapitel 5 vorgestellt.

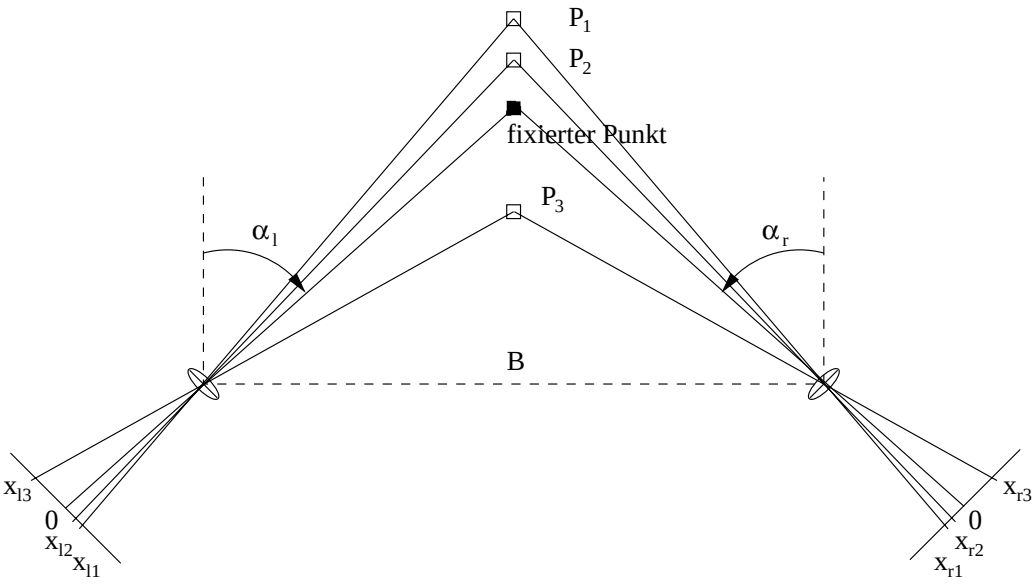


Abbildung 4.10: Abbildungsgeometrie eines Stereokamerakopfes. Die Punkte P_1, P_2 und P_3 werden aufgrund der Geometrie in beiden Bildebenen an verschiedene Orte projiziert, die mit x_{r1} bis x_{r3} und x_{l1} bis x_{l3} gekennzeichnet sind. α_l bzw. α_r stellen die Auslenkung der Kameraachsen bezüglich der Senkrechten der Stereobasis B dar.

Das unten beschriebene Verfahren zur Objekt-Hintergrund-Trennung auf der Basis von Stereodaten baut auf bereits realisierte Algorithmen im NAVIS-Kontext auf. Zunächst werden Kantenelemente $g(x, y, \phi)$[2] durch die Anwendung eines Gabor-Filtersatzes generiert [TRAPP et al. 1995], [TRAPP 1996]. Die lokalen Maxima der Filterantworten werden als Kanten interpretiert. Zur Bestimmung der Korrespondenzen zwischen dem linken und rechten Kamerabild wird ein einfaches Korrelationsverfahren eingesetzt: Ein Zeilenabschnitt des einen Bildes wird mit der gesamten zugehörigen Zeile des anderen Bildes, welche durch das epipolar-*constraint* bestimmt ist, kreuzkorreliert. Die anschließende Anwendung eines *continuity-constraint* erzeugt eindeutige Disparitätswerte $d(x, y)$ pro Bildpunkt. Abbildung 4.11 zeigt die sich ergebende Disparitätskarte für den eingesetzten Algorithmus. Zu erkennen sind die innerhalb der Region des Objektes stark zerrissenen Disparitätsfelder.

Die an den interessierenden Regionen nur dünn besetzten Disparitätsfelder

[2] x, y bezeichnen kartesische Koordinaten und ϕ eine Vorzugsorientierung. Kanten werden nach 12 verschiedenen Orientierungen unterschieden.

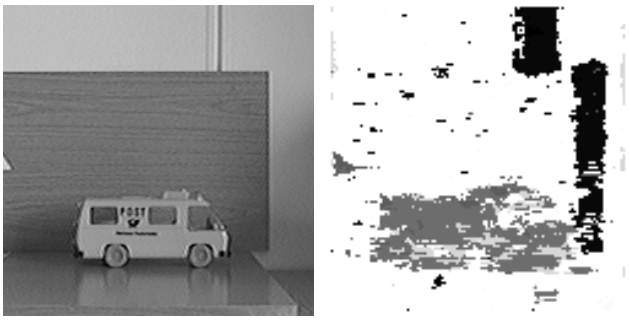


Abbildung 4.11: Szene mit generierter Disparitätskarte

sollen durch einen iterativen Prozeß verdichtet werden, d.h. gesteuert durch heuristische Regeln sollen vorhandene Disparitäten in bisher undefinierte Bereiche expandiert werden [LIEDER et al. 1998]. Basis sind die durch die Gaborfilterung erhaltenen Kantenbilder $g(x, y, \phi)$. Nach einer Schwellwertfilterung (Schwelle t_1) zur Rauschunterdrückung wird die Breite der Konturen verringert (Schwelle t_2). Der in diesem und in den folgenden Algorithmen verwendete Suchbereich ist eine Maske, die in Orientierung und Breite der aktuell zu untersuchenden Kantenorientierung angepaßt wird. Dadurch werden verschieden starke Gaborfilterantworten an Ecken bzw. nicht exakt mit der Filterorientierung übereinstimmenden Kanten ausgeglichen. Abbildung 4.12 verdeutlicht den Algorithmus zur Generierung der Konturelemente $g_c(x, y)$.

Zur Expansion der Disparitätswerte $d(x, y)$ wird ein "Flut"-Algorithmus eingesetzt. Heuristiken steuern dabei die Ausbreitung der Tiefenwerte:

- Konturelemente begrenzen die Expansion der Tiefendaten, da an diesen Positionen die Wahrscheinlichkeit für das Vorhandensein realer Tiefensprünge groß ist.
- Besitzen benachbarte Pixel unterschiedliche Tiefenwerte, so stellen diese ebenfalls Barrieren für die Ausbreitung der Tiefendaten dar.
- Entlang von Kanten sollen sich Disparitätswerte leicht ausbreiten können.

In Abbildung 4.13 wird die gewünschte Funktionsweise grafisch skizziert und Abbildung 4.14 gibt den Algorithmus an. Während der Iterationen wird die Tiefenkarte geglättet, allerdings bleiben Sprünge in den Disparitätswerten erhalten, um bezüglich ihrer Tiefe separierte Objekte nicht zusammenfließen zu lassen. Schrittweise werden die Lücken in der Disparitätskarte aufgefüllt. Der Algorithmus endet mit der expandierten Disparitätskarte $d_e(x, y)$, wenn keine neuen Disparitätswerte generiert werden konnten.

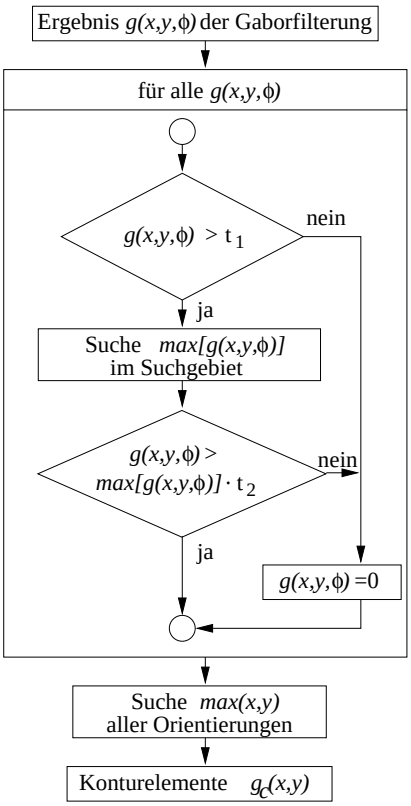
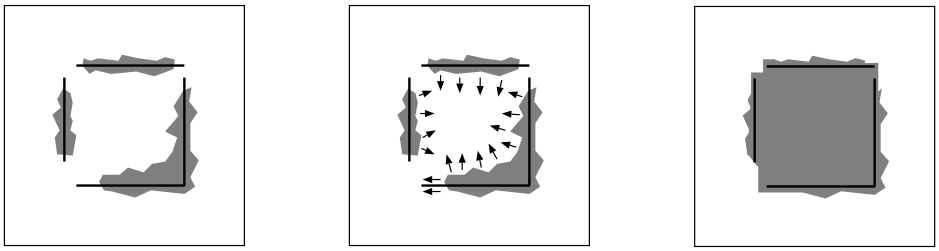


Abbildung 4.12: Algorithmus zur Generierung von Konturelementen (nach [LIEDER et al. 1998])



a) b) c)

Abbildung 4.13: Prinzip des Expansionsverfahrens für Disparitätswerte (nach [LIEDER 1997]): *a)* Szene mit Kanten und initialen Disparitätswerten; *b)* Expansionsrichtungen; *c)* Ergebnis

Die bereits berechneten Konturelemente $g_c(x, y)$ werden mit den noch zu berechnenden Tiefendiskontinuitäten $d_c(x, y)$ zusammengefaßt. Die $d_c(x, y)$ werden

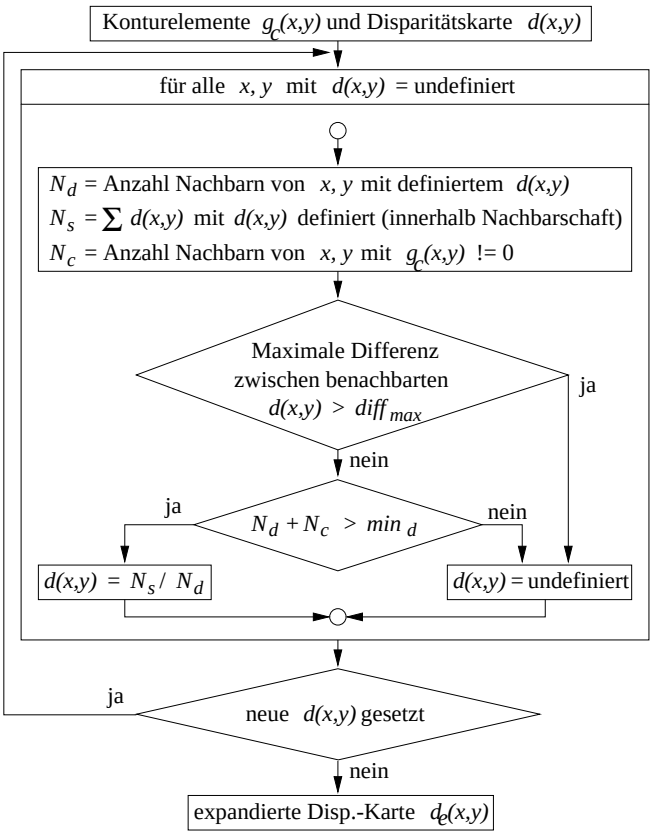


Abbildung 4.14: Algorithmus zur Expansion von Tiefendaten (nach [LIEDER et al. 1998])

aus der expandierten Disparitätskarte $d_e(x, y)$ abgeleitet. Gesucht werden Sprünge in den Disparitätswerten, die auf reale Tiefenübergänge hinweisen. In Abbildung 4.15 ist der Algorithmus dargestellt.

Die Zusammenfassung nah beieinander liegender Konturelemente $g_c(x, y)$ und der Tiefensprünge $d_c(x, y)$ dient zur Begrenzung eines weiteren "Flut"-Algorithmus zur Szenenextraktion [LIEDER et al. 1998]. Es kommen die gleichen Heuristiken wie zur Expansion der Tiefendaten zur Anwendung. Ausgangspunkte für die Extraktion sind die durch Aufmerksamkeitsmechanismen gelieferten Positionen. Pro Bildpunkt werden in einer 8-Nachbarschaft die vorhandenen expandierten Disparitätswerte $d_e(x, y)$ untersucht. Bei kleinen Differenzen wird der Bildpunkt dem gerade extrahierten Szenensegment hinzugefügt; bei großen Differenzen entsprechend nicht.

In Abbildung 4.16 ist eine beispielhafte Expansion von Tiefendaten und eine anschließende Szenensegmentierung angegeben. Die beiden Objekte wurden aus

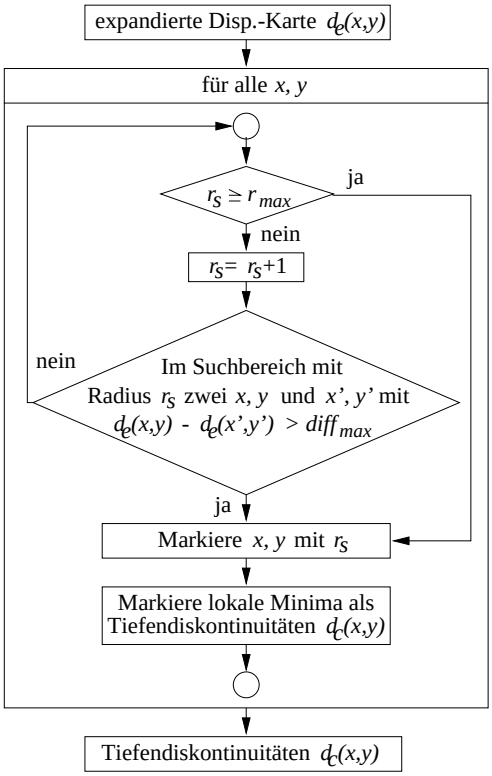


Abbildung 4.15: Algorithmus zur Detektion von Diskontinuitäten (nach [LIEDER et al. 1998])

der Umgebung herausgelöst und können entsprechend weiter analysiert werden. In Abbildung 4.17 ist eine weitere Szene dargestellt, wobei hier das Problem auftritt, daß das Objekt durch seine räumliche Schrägstellung einen weiten Bereich an Disparitätswerten besitzt. Trotzdem konnte es zufriedenstellend aus der Umgebung herausgelöst werden. Durch die Szenensegmentierung kann somit basierend auf Bilddaten eines Stereokamerakopfes eine Objekt-Hintergrund-Trennung als Vorverarbeitungsschritt zur Verringerung des Suchraumes in Objekterkennungssystemen durchgeführt werden.

Als verbesserbar muß die Rechenzeit des Systems gesehen werden. Hierbei fällt vor allem die Anwendung des epipolar-*constraints* und des *continuity-constraints* zur initialen Berechnung der Disparitätsfelder ins Gewicht. Schnellere Verfahren sind in der Literatur durchaus bekannt (z.B. [FUSIELLO et al. 1997]) bzw. dedizierte Hardwarerealisierungen ermöglichen eine Echtzeitberechnung von dichten Disparitätsfeldern (z.B. [KONOLIGE 1997]).

Zusammenfassend ist festzustellen, daß die Stereovorverarbeitung einen ge-

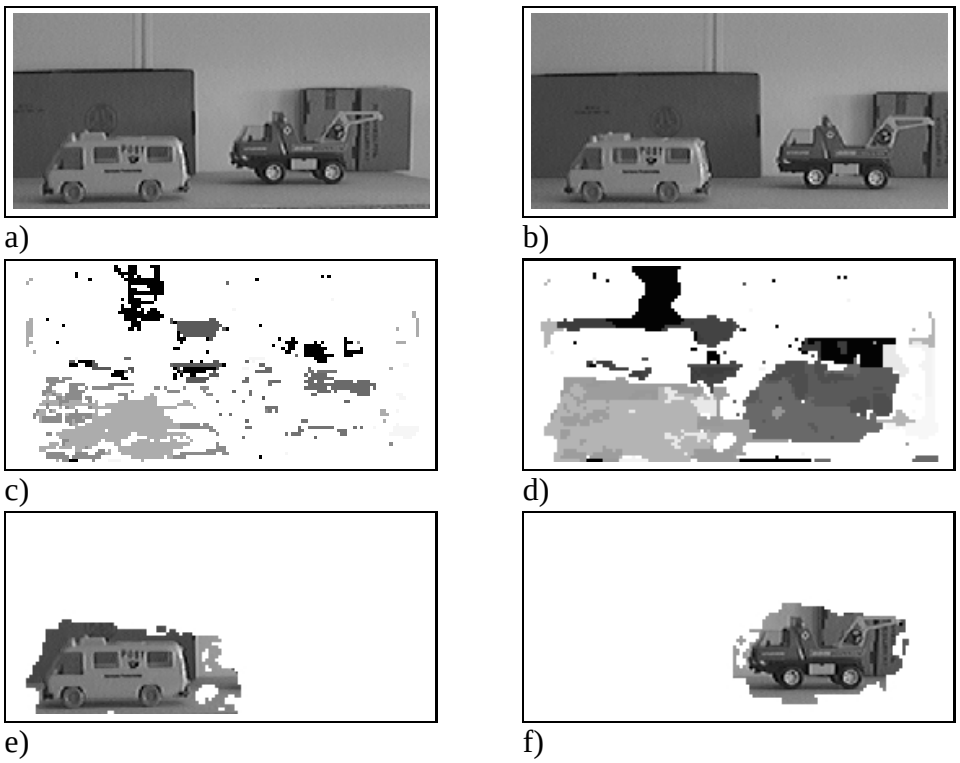


Abbildung 4.16: Demonstration der Szenensegmentierung: *a)*, *b)* Stereobildpaar; *c)* abgeleitete Disparitätskarte; *d)* expandierte Disparitätskarte; *e)*, *f)* extrahierte Szenensegmente (nach [LIEDER et al. 1998])

eigneten Weg darstellt, um eine robuste ansichtenbasierte Objekterkennung zu unterstützen. Durch die Ausblendung von störendem Hintergrund durch eine tiefenbasierte Objekt-Hintergrund-Trennung wird der Einsatz von Erkennungssystemen in natürlichen Umgebungen ermöglicht.

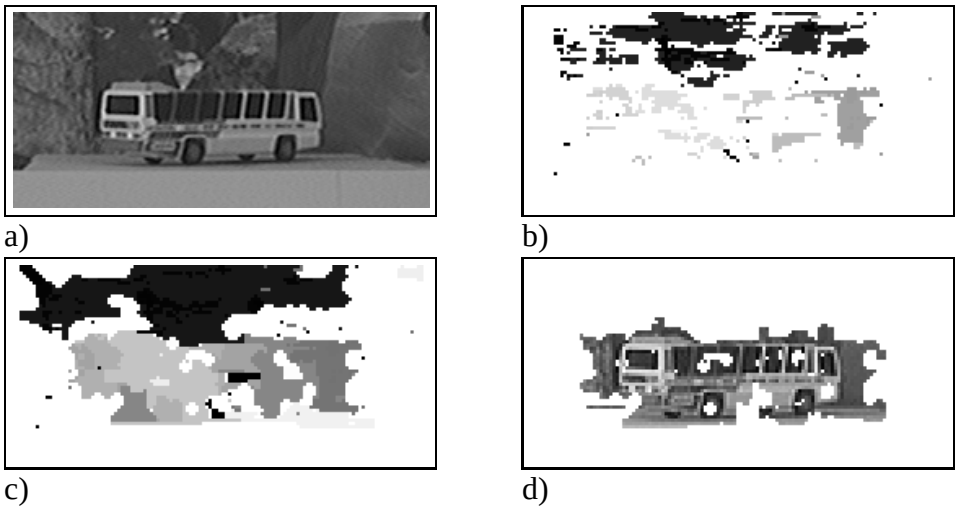


Abbildung 4.17: Demonstration der Szenensegmentierung: *a)* linkes Bild eines Stereobildpaares; *b)* abgeleitete Disparitätskarte; *c)* expandierte Disparitätskarte; *d)* extrahierte Szenensegment (nach [LIEDER et al. 1998])

Kapitel 5

Objekterkennung mit Modellskalierungen

In [LIEDER et al. 1998] wird ein Erkennungssystem vorgestellt, welches durch die Einbeziehung von Tiefeninformationen in den Erkennungsprozeß eine höhere Robustheit erzielt. Zur Demonstration der Leistungsfähigkeit wurde dieser Ansatz in das im Kapitel 4.3 vorgestellte Erkennungssystem integriert. Ziel war damit das Kriterium der Invarianzen (siehe Tabelle 3.3) stärker zu berücksichtigen. Konkret sollte eine starke Größeninvarianz durch eine Modellskalierung erreicht werden. Diese Skalierung wird aus dem Größenverhältnis des Aufmerksamkeitspunktes zum matchenden gelernten Attraktivitätspunkt abgeleitet. Die Bestimmung der idealen Modellskalierung erfolgt durch einen *trial-and-error*-Prozeß. Zur Verringerung des Suchraumes wird die Szene entsprechend zuvor generierter Tiefendaten segmentiert (siehe hierzu Kapitel 4.4).

5.1 Erkennungssystem

In Abbildung 5.1 ist das Blockschaltbild des um eine 3D-Tiefenauswertung erweiterten Erkennungssystems von Götze und Mertsching (siehe auch Abbildung 4.5) zu sehen. Neu hinzugekommen ist ein Stereomodul, welches die Stereobildpaare analysiert und darauf aufbauend eine Modellskalierung und eine Tiefensegmentierung zur Herauslösung des Objekts aus dem Hintergrund ermöglicht. Die dazu nötigen Erweiterungen betreffen alle Module, da auf extrahierten Szenenbildern gearbeitet wird und die Modelle jeweils skaliert werden müssen. Die Generierung der Attraktivitätspunkte erfolgt nicht nur für die segmentierten Bildregionen, sondern für die Originalszenenbilder, da zum einen während der Merkmalsberechnung Artefakte an den Segmentationsgrenzen entstehen könnten und zum anderen ein Abbruch der aktuellen Hypothesenvalidierung beim Auftauchen interessante-

rer (attraktiverer) Bildregionen weiterhin möglich sein soll.

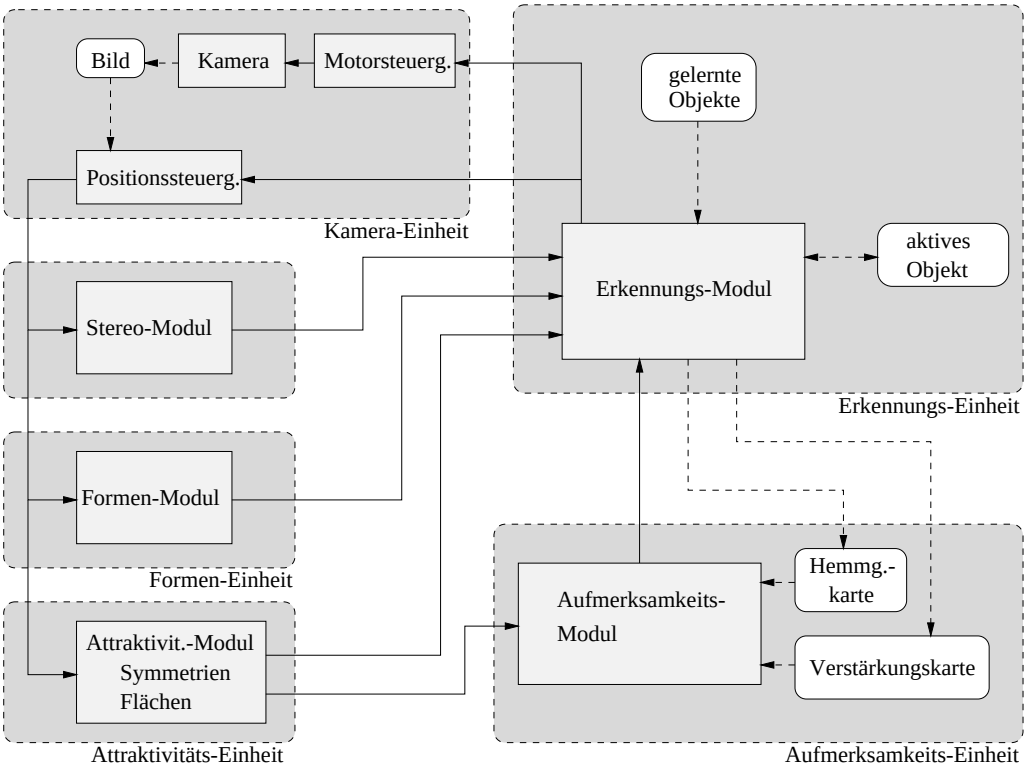


Abbildung 5.1: Blockschaltbild der Objekterkennung mittels Modellskalierung (nach [LIEDER 1997])

Abbildung 5.2 zeigt das neue Ablaufschema des überarbeiteten Erkennungssystems im Detail. Die Skalierung der Modellpunkte und -formen basiert auf einem globalen Skalierungsfaktor. Zur Bestimmung dieses Faktors werden zunächst alle Objektpunkte mit zum Aufmerksamkeitspunkt identischen Typattribut selektiert. Aus den Größenverhältnissen wird der Skalierungsfaktor für alle matchenden Typattribute einzeln abgeleitet. Es wird ein Skalierungsbereich von $\pm 5\%$ um diesen Faktor verwendet. Faktoren kleiner als 0.5 werden hierbei nicht zugelassen, da bei dieser Verkleinerung die zu matchenden Objektformen so stark verdichtet werden, daß keine Objektdiskriminierungen mehr möglich sind. Nach der Skalierung der Modellattraktivitätspunkte wird der Hypothesenmatchwert für alle Skalierungsfaktoren innerhalb des Skalierungsbereiches (bei einer Schrittweite von 0.5%) bestimmt. Der maximale Hypothesenmatchwert ergibt die endgültige Skalierung des Modells. Um die Anzahl der irrtümlichen Punktekorrespondenzen mit dem Hintergrund einzuschränken, wird die untersuchte Szene entsprechend gewonnener 3D-Daten in Objekte und Hintergrund segmentiert. Bei Vorhandensein einer

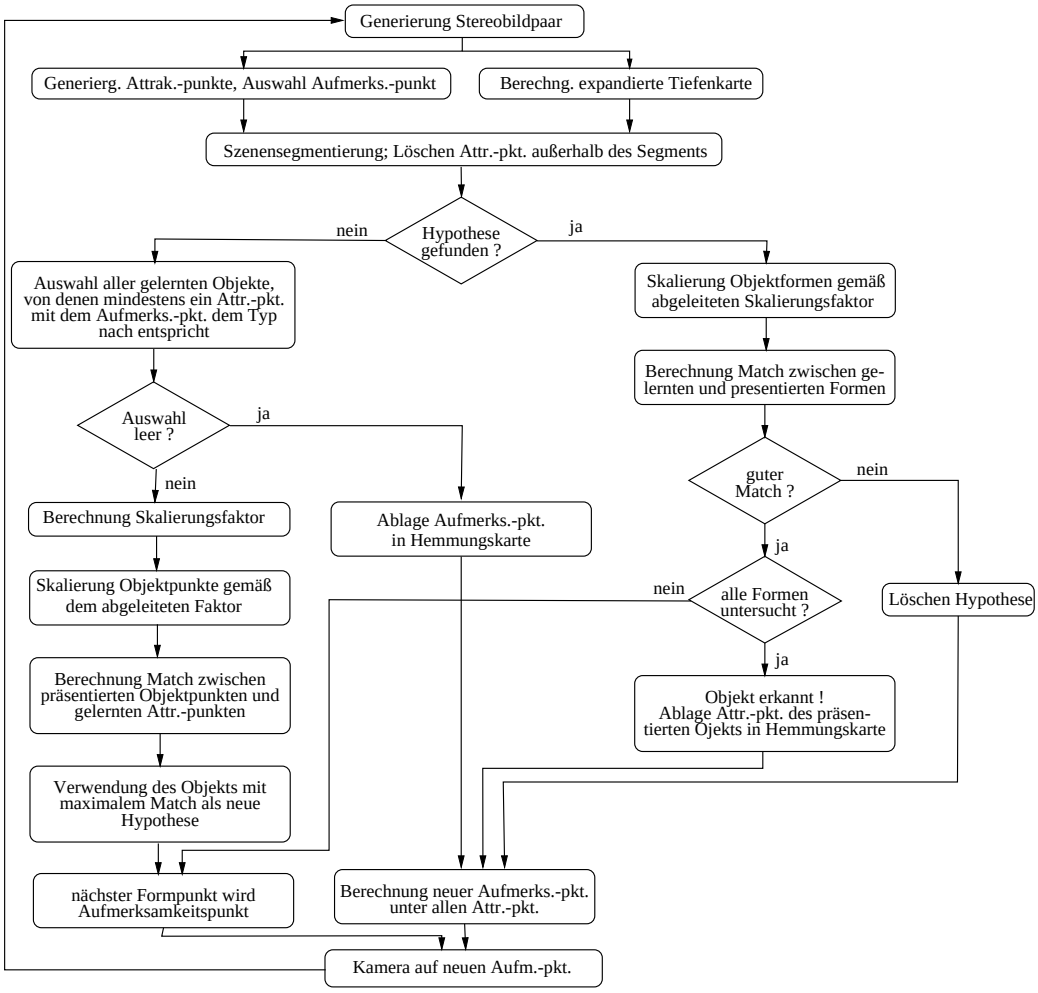


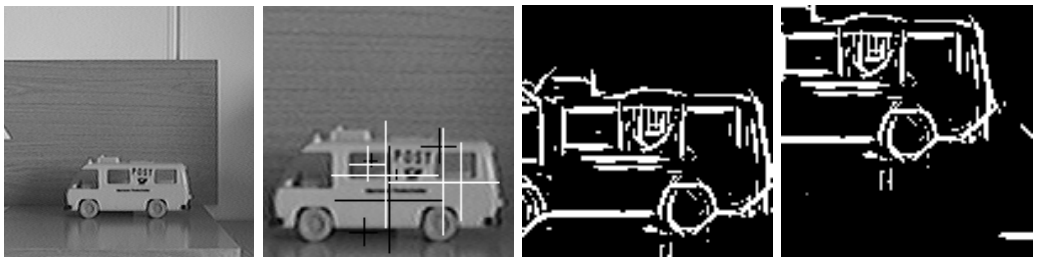
Abbildung 5.2: Ablaufschema der Objekterkennung mittels Modellskalierung (nach [LIEDER 1997])

gültigen Hypothese wird diese durch das Suchen der gelernten Objektformen validiert. Das Fixieren der Objektform sowie die Auswahl des Aufmerksamkeitspunktes entsprechen dem ursprünglichen Erkennungssystem. Bei dem Matchen der einzelnen Objektformen werden jetzt die Modellformen entsprechend des ermittelten Hypothesenskalierungsfaktors skaliert.

Die Szenensegmentierung dient hier ausschließlich zur Verringerung des Suchraums während der Hypothesengenerierung. Alle außerhalb des segmentierten Objekts liegenden Attraktivitätspunkte werden gelöscht. So muß für deutlich weniger Punkte der *trial-and-error*-Prozeß zur Bestimmung der optimalen Hypothesenskalierung durchgeführt werden.

5.2 Experimente und Ergebnisse

Zunächst wird das Erkennungssystem ohne eine Szenensegmentierung getestet. Ziel ist nicht die Validierung der Erkennungsstrategie mittels Objektformen (Aussagen hierzu finden sich im Kapitel 4.3), sondern eine Bewertung der Anwendbarkeit von Modellskalierungen für eine skalierungsinvariante Erkennung. Dementsprechend kann sich auf die Untersuchung einzelner Objekte vor unstrukturiertem Hintergrund beschränkt werden. In Abbildung 5.3 ist ein gelerntes Objekt abgebildet.



a) b) c) d)

Abbildung 5.3: Das Beispielobjekt zur Validierung der Objekterkennung mittels Modellskalierung; *a)* ein Spielzeugauto als Demonstrationsobjekt; *b)* gelernte Attraktivitätspunkte; *c)* und *d)* gelernte Objektformen (nach [LIEDER 1997])

Das Objekt wird dem System mit einer Skalierung von -40% bis +60% präsentiert. Tabelle 5.1 zeigt die ermittelten Matchwerte für die Hypothese und die beiden Objektformen. Die Spalte “ermittelte Skalierung” gibt die bestimmte optimale Modellskalierung innerhalb des durch die Basisskalierung festgelegten Bereiches an.

Zu Erkennen ist eine bis zu 10%ige Abweichung der ermittelten Skalierung von der tatsächlichen. Weiterhin ist festzustellen, daß sich ein maximaler mittlerer Match nicht zwingend bei der besten Übereinstimmung von ermittelter und tatsächlicher Skalierung ergibt. Hauptursache dieser Effekte ist die nicht konstante Ausbildung der Attraktivitätspunkte über den gesamten Skalierungsbereich (siehe Abbildung 5.4 zur Verdeutlichung). Die Ursache für die unterschiedlichen Detektionsergebnisse für Symmetrien liegt z.B. in der intern durchgeführten zweifachen Unterabtastung. Dadurch werden u.U. bei kleinen Strukturen keine Symmetrien mehr detektiert. Bei der Bestimmung der Flächenmerkmale treten ähnliche Effekte auf. Bei einem Teil der präsentierten Objekte war keine Erkennung möglich (Skalierung 0.6, 0.65, 0.85, 0.9, 1.2, 1.25, 1.35-1.6). In diesen Fällen wurden der Erkennungsvorgang nach 20 Iterationen pauschal abgebrochen, da keine erfolgreiche Erkennung zu erwarten ist. Die Klassifikationen für die Skalierungen 0.85,

Tabelle 5.1: Mittlere Matchergebnisse pro Objektskalierung (k.E.=keine Erkennung; nach [LIEDER 1997])

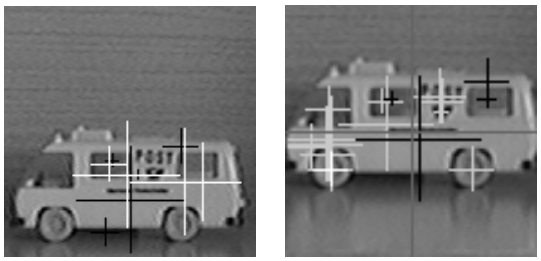
reale Skalierung	ermittelte Skalierung	Matchwerte		
		Hypothese	Form 1	Form 2
0.60	k.E.; keine korrekte Hypothese gebildet			
0.65	k.E.; keine korrekte Hypothese gebildet			
0.70	0.699	0.856	0.851	0.879
0.75	0.699	0.864	0.851	0.877
0.80	0.740	0.831	0.875	0.898
0.85	k.E.; niedriger Formmatch blockiert korrekte Hypothese			
0.90	k.E.; niedriger Formmatch blockiert korrekte Hypothese			
0.95	0.876	0.900	0.921	0.945
1.00	0.949	0.826	0.982	0.982
1.05	0.949	0.777	0.883	0.933
1.10	1.200	0.771	0.869	0.931
1.15	1.215	0.945	0.922	0.964
1.20	k.E.; niedriger Formmatch blockiert korrekte Hypothese			
1.25	k.E.; niedriger Formmatch blockiert korrekte Hypothese			
1.30	1.316	0.763	0.938	0.958
1.35 - 1.60	k.E.; keine korrekte Hypothese gebildet			

0.9, 1.2 und 1.25 konnten aufgrund einer ungünstigen Hypothesenverwaltung nicht erkannt werden. Die Objekte wurden zunächst basierend auf einer falschen Skalierung untersucht. Die Formmatchergebnisse waren entsprechend gering. Da aber diese niedrigen Formmatchergebnisse pro Objekt gespeichert werden, hat auch eine in der Folge ermittelte korrekte Objektskalierung keine Chancen, die alten Formergebnisse zu revidieren.

In einem weiteren Experiment wird das Zusammenspiel von Szenensegmentierung und Objekterkennung getestet. Präsentiert werden zwei gleichgroße Objekte vor einem strukturierten Hintergrund (analog zu Abbildung 4.16a/b). In Tabelle 5.2 sind die Ergebnisse abgebildet. Zu erkennen ist, daß beide Objekte nicht für alle präsentierten Skalierungen erkannt werden konnten[1]. Allerdings ist zu entnehmen, daß eine prinzipielle skalierungsinvariante Erkennung möglich ist.

Zur Verbesserung der Erkennungsergebnisse sind ein Reihe von Modifikationen nötig, die vor allem interne Verwaltungsvorgänge der Hypothesen und Attraktivitätspunkte betrifft:

[1]Objekt Nr. 2 wurde für die realen Skalierungen von 0.9 und 1.2 nicht erkannt.



a) b)

Abbildung 5.4: Generierte Attraktivitätspunkte bei verschiedenen Objektskalierungen (*a)* mit Faktor 1.0; *b)* mit Faktor 1.6; nach [LIEDER 1997])

Tabelle 5.2: Matchergebnisse für das Gesamtsystem bestehend aus Szenensegmentierung und Objekterkennung (nach [LIEDER 1997])

reale Skalierung	Objekt	ermittelte Skalierung	Matchwerte		
			Hypothese	Form 1	Form 2
0.7	Nr. 1	0.827	0.738	0.700	0.732
	Nr. 2	0.649	0.734	0.798	0.782
0.9	Nr. 1	0.949	0.703	0.900	0.882
1.0	Nr. 1	0.954	0.709	0.954	0.935
	Nr. 2	0.949	0.826	0.928	0.915
1.2	Nr. 1	1.049	0.798	0.826	0.818

- Die Formmatchwerte werden global gespeichert, wobei kein Rücksetzen der Werte vorgesehen ist. Wurde nun mit einer ungünstigen Skalierung ein niedriger Formmatch bestimmt, blockiert dieser niedrige Wert alle weiteren Erkennungsversuche mit evtl. passenderen Skalierungsfaktoren. Hier sollte ein geeigneter Rücksetzmechanismus integriert werden, der auch abhängig vom gewählten Skalierungsfaktor sein sollte.

- Die Verwaltung der Hypothesen sieht bisher vor, daß zu einer gegebenen Szene immer nur eine Hypothese, die mit maximalen Hypothesenmatchwert, validiert wird. Alle anderen werden verworfen. War die falsche Hypothese ausgewählt worden, wird nach einem erfolglosen Formmatch die gesamte Objektregion inhibiert, womit eine erneute Hypothesenbildung erst nach dem Abklingen der Inhibitionswerte wieder durchgeführt werden kann. Dadurch wird die Erkennung spürbar verlangsamt bzw. völlig unmöglich gemacht. Sinnvoller wäre hier die Überprüfung mehrerer Hypothesen

pro Szene.

- Die maximale Anzahl von Attraktivitätspunkten ist global festgelegt. Dadurch ergeben sich bei sich ändernden Kontrastverhältnissen instabile Attraktivitätspunkte, da in kontrastreichen Regionen i.a. mehr Punkte detektiert werden als in kontrastarmen.

- Ein Hauptproblem stellt die nicht stabile Detektion der Attraktivitätspunkte über einen größeren Skalierungsbereich dar. Dadurch ergeben sich ändernde Punktekonfigurationen, was wiederum zu Fehlklassifikationen bzw. einem Nichterkennen führt. Abhilfe würde entweder eine Änderung der Detektionsstrategie oder eine Mehrfachrepräsentation für verschiedene Modellskalierungen schaffen.

- Im ähnlichen Zusammenhang wäre eine Überarbeitung des derzeitigen Formmatchverfahrens wünschenswert. Bei der derzeitigen Strategie werden einfach verbreiterte Kanten auf schmale gematcht. Nach einer evtl. Skalierung verschmelzen u.U. die Kanten, so daß keine Diskriminierung mehr möglich ist.

- Die gematchten Objektformen sollten ein Attribut für die zugrundegelegte Skalierung erhalten. Dadurch können sich widersprechende Matchergebnisse durch die Anwendung unterschiedlicher Skalierungen für ein gleiches Objekt ausgeschlossen werden.

Das Erkennungssystem mittels Modellskalierungen kann nur als Studie über die Möglichkeiten skalierungsinvarianter Objekterkennung gewertet werden. Da der Ansatz innerhalb des Erkennungssystems nach Götze und Mertsching realisiert wurde (siehe Kapitel 4.3), sind die prinzipiellen Probleme deses Systems auch hier zu beobachten. Es zeigt sich allerdings, daß mit Hilfe einfacher Modifikationen die Leistungsfähigkeit des bestehenden Systems erhöht werden konnte. Der Einsatz einer Stereovorverarbeitung für Objekterkennungssysteme bietet somit einen Weg, Probleme bedingt durch störende Hintergrundinformationen zu lösen. Eine Skalierung aufgrund von Distanzinformationen ist möglich, sollte aber nicht als alleinige Methode angewendet werden, sondern als Ergänzung zu in das Erkennungssystem eingearbeiteter Skalierungsinvarianz.

Kapitel 6

Objekterkennung mit zeitlichen Relationen

In [MASSAD et al. 1998a] wird eine 3D-Objekterkennung realer Objekte durch die Verwaltung charakteristischer Ansichten vorgestellt. Erstmalig wird im NAVIS-Kontext die ansichtenbasierte Erkennung von 3D-Objekten angestrebt. Dazu ist die selbständige, effektive Herausbildung und Verwaltung von Ansichtenklassen nötig. Ähnliche, in der *view-sphere* benachbarte Ansichten sollten in eine Klasse eingeordnet werden. Eventuell auftretende mehrdeutige Ansichten dürfen nur kurzfristig die Erkennungsleistung beeinträchtigen. Durch die Analyse des zeitlichen Kontexts soll es möglich sein, trotz solcher uneindeutigen Ansichten eine korrekte Objektklassifikation zu ermöglichen. Gemäß dem Anforderungskatalog aus Tabelle 3.3 sollen hier weitergehende Kriterien bezüglich Invarianzen, der selbständigen Auswahl von typischen Objektansichten und biologischer Adäquatheit bezüglich der Objektrepräsentation bearbeitet werden. Das Erkennungssystem zielt weiterhin auf eine im Vergleich zu den bisher vorgestellten Erkennungssystemen (2D-Objekterkennung nach Kapitel 4.3 sowie Erkennung mit Modellskalierungen nach Kapitel 5) höheren Stabilität der Matchergebnisse ab.

6.1 Repräsentationsschema

Es wurde sich intensiv der Fragestellung nach einer adäquaten Objektrepräsentation gewidmet. Vor allem auf Grund psychophysikalischer Erkenntnisse werden betrachtungszentrierte Repräsentationen eingesetzt (siehe Kapitel 3.1.1). Abbildung 6.1 zeigt die verwendete Repräsentationshierarchie. Hauptentwurfsaspekt war eine hohe Flexibilität in den Zuordnungen zwischen Ansichten, Sequenzen und Objekten. Ansichten werden zunächst Ansichten-Übergängen (Sequenzen) zugeordnet und diese wiederum Objekten. Nur aus einer Ansicht bestehende Ob-

jektrepräsentationen werden als sehr kurze Sequenzen aufgefaßt. Prinzipiell können Ansichten auch mehreren Sequenzklassen angehören. Dies ist sinnvoll, wenn ein Objekt z.B. einmal um die Längsachse und einmal um die Querachse rotiert präsentiert wurde. Dabei entstehen für beide Rotationen identische Ansichten. Weiterhin ließen sich die Sequenzen mit Hinweisen über ihre Entstehung (Rotation, Verschiebung etc.) versehen, um später die erkannte Bewegungsform eines Objekts angeben zu können.

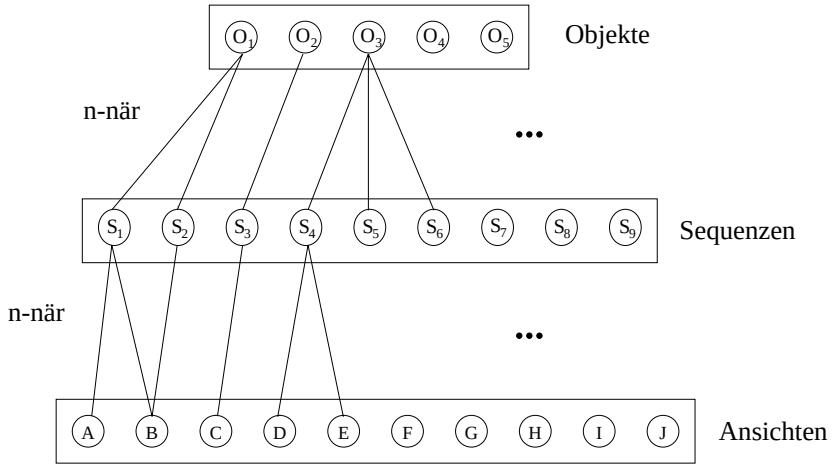


Abbildung 6.1: Repräsentationshierarchie (nach [MASSAD et al. 1998a])

6.2 Erkennungssystem

In Abbildung 6.2 ist das Gesamtsystem dargestellt. Im Vergleich zu dem Erkennungssystem nach Götze und Mertsching werden hier einige vereinfachende Annahmen getroffen, um die Verarbeitungskomplexität zu reduzieren und sich bewußt auf die 3D-Objekterkennung zu beschränken:

- Verwendung von Szenenbildern ohne Hintergrund
- keine Einbindung von Aufmerksamkeitssteuerungen, d.h. das zu untersuchende Objekt befindet sich bereits innerhalb des Bildes
- keine Anbindung an Aktorik

Erster Schritt der Verarbeitungskette ist das Matchen von Ansichten auf Ansichtenklassen durch ein adaptiertes *Dynamic Link Matching*. Die dadurch gewonnenen Klassenzuordnungen werden anschließend “nur noch” verwaltet. Zunächst

werden Einzelansichten zu zeitlichen Ansichtensequenzen kombiniert und darauf aufbauend die Objektklasse bestimmt. In den folgenden Kapiteln werden die einzelnen Verarbeitungsstufen vorgestellt.

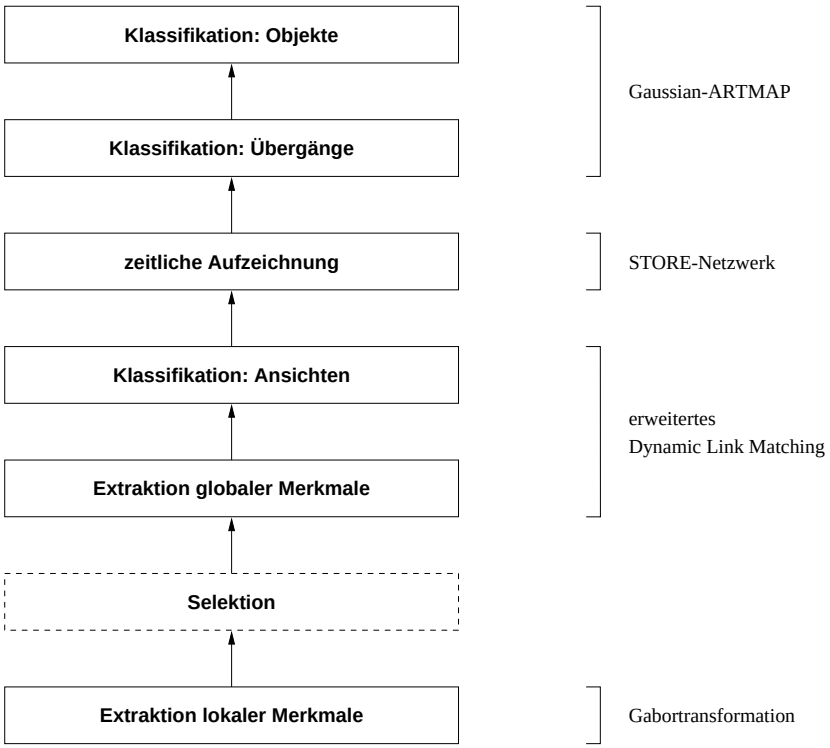


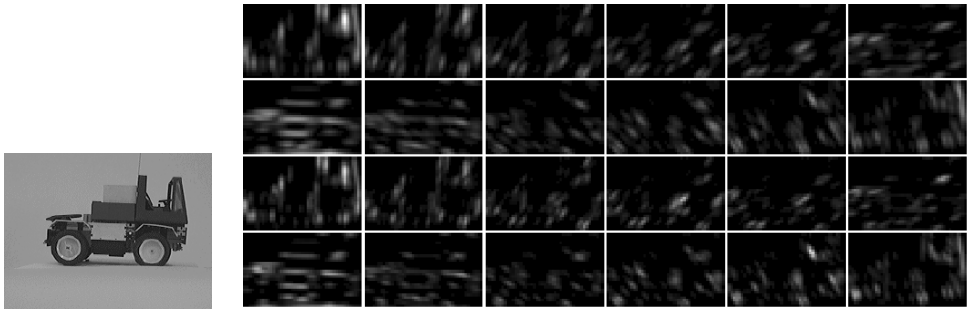
Abbildung 6.2: Das 3D-Erkennungssystem mit einer Analyse zeitlicher Relationen (nach [MASSAD et al. 1998a])

6.2.1 Merkmalsbasis

Als Merkmalsbasis werden gerichtete, orientierte Kanten eingesetzt, wobei die Orientierungen diskretisiert und verschiedene Auflösungsebenen eingesetzt werden. Zur Merkmalsgenerierung werden Gaborfilter eingesetzt (siehe Kapitel 4.2). Die Amplitudenwerte der rücktransformierten Filterantworten bilden dabei die Merkmalsvektoren - *Gaborjets* genannt.

Insgesamt werden 12 Orientierungen in 2 Auflösungsstufen unterschieden, womit sich ein 24-dimensionaler Merkmalsraum ergibt (siehe Abbildung 6.3). Die Nutzung hochdimensionaler Merkmalsräume hatte sich z.B. schon in [RAO und BALLARD 1995] bewährt. Zur Verringerung der Rechenkomplexität

werden die ursprünglich 256×256 Pixel großen Bilder auf 64×64 Pixel unterabgetastet.



a) b)

Abbildung 6.3: Generierte Gaborfilterantworten; *a)* Originalbild; *b)* Ergebnis als Zerlegung in zwei Auflösungsebenen (jeweils obere und untere zwei Zeilen) mit je 12 Orientierungen

6.2.2 Ansichtenklassifikation

Zur Klassifikation der präsentierten Ansichten gegen die gelernten Ansichtenklassen wird eine modifizierte Variante des *Dynamic Link Matchings* eingesetzt ([VON DER MALSBURG 1981], [KONEN et al. 1994]). Hauptmotivation für den Einsatz dieses Verfahrens war dessen Möglichkeit einer invarianten Mustererkennung im Merkmalsraum sowie die guten Generalisierungseigenschaften. Somit wird die Verwendung einer geringen Menge von typischen Ansichten für ein zu repräsentierendes Objekt möglich. Das System ist dann in der Lage, zwischen den gelernten Ansichten zu interpolieren. Grundidee der Klassifikation ist hierbei die Annahme, daß die lokale Transformation eines Merkmalspunktes auch auf die Nachbarn dieses Punktes übertragen werden kann. Basis ist eine definierte Ähnlichkeitsfunktion zwischen gelernten und präsentierten Merkmalsmustern. Zwischen zwei Merkmalen f_a und f_b wird das normierte Skalarprodukt als Ähnlichkeitsfunktion definiert:

$$S(f_b, f_a) := \frac{f_b}{\|f_b\|} \cdot \frac{f_a}{\|f_a\|} \tag{6.1}$$

Es existieren zwei Schichten, in denen jeweils die präsentierten (Zellen a in Schicht X) und gelernten Muster (Zellen b in Schicht Y) kodiert sind. Zwischen diesen Schichten existieren Inter-Layer-Verknüpfungen, die ein Kurzzeitgedächtnis modellieren und als *dynamic links* bezeichnet werden. Zwischen ähnlichen

Mustern bilden sich im Verlauf des Iterationsprozesses stabile Verbindungen aus. Es werden zufällig ausgewählte Regionen der präsentierten Szene, sog. *Blobs*, ausgewählt und in den Modellen passende Regionen aktiviert. Diese Regionen können eine beliebige, meist eine einfache rechteckige Form, besitzen. Innerhalb der Merkmalsschichten existieren ebenfalls Verknüpfungen: lokale exzitatorische und global inhibitorische. Der Algorithmus existiert in zwei Varianten: eine neuronale Formulierung, die durch numerische Lösungen von Differentialgleichungen gelöst wird, sowie eine vereinfachte Lösung, das sog. *Fast Dynamic Link Matching (FDLM)*. Im folgenden wird immer die vereinfachte, schnellere Variante benutzt. Der Algorithmus ist in Abbildung 6.4 dargestellt.

Dynamic Link Matching — Matchen von Gaborjets

Initialisierung der Inter-Layer-Verknüpfungen J_{ba} gemäß der Ähnlichkeitsfunktion $S(f_b, f_a)$: $J_{ba} = \frac{T_{ba}}{\sum_{a' \in X} T_{ba'}}$ und $T_{ba} = S(f_b, f_a)$.
Zufällige Auswahl eines Zentrums $a_c \in X$ und dortiges Plazieren des *Blobs*: $x_a = B(a - a_c)$. Berechnung der resultierenden Aktivierung I der Zellen $b \in Y$: $I_b = \sum_a J_{ba} T_{ba} B(a - a_c)$.
Suche des Maximums $b_c \in Y$ des Potentials V: $V(b_c) = \sum_{b \in Y} B(b - b_c) I_b$. Positionieren des *Blobs* an diese Position: $y_b = B(b - b_c)$.
Neuberechnung der Verknüpfungen zwischen den aktiven *Blobs* und Normalisierung: $J_{ba} = \frac{J_{ba} + \Delta J_{ba}}{\sum_{a' \in X} (J_{ba'} + \Delta J_{ba'})}$ mit $\Delta J_{ba} = \epsilon J_{ba} T_{ba} Y_b X_a$.
Erfüllung eines Stoppkriteriums, z.B. durch eine Analyse der Iterationszahl

Abbildung 6.4: Algorithmus zum Matchen von Gaborjets (nach [KONEN et al. 1994])

Mehrere Abänderungen vom Originalalgorithmus waren nötig, da bei einer Sequenzklassifikation prinzipiell alle präsentierten Objektansichten benachbarter Betrachtungswinkel untereinander sehr ähnlich sind. Trotzdem muß noch eine Diskriminierung zwischen verschiedenen Klassen möglich sein. Dementsprechend waren folgende Veränderungen nötig:

- Das Matchen erfolgt parallel gegenüber allen gelernten Ansichten innerhalb eines Wettbewerbsprozesses. Eine *Winner Takes All*-Strategie wählt die beste Übereinstimmung aus. Nur diese Gewichte werden angepaßt. Der Algorithmus nach Abbildung 6.4 muß nun m Schichten Y_m, wobei m die Anzahl

der Modelle angibt, gleichzeitig berücksichtigen. Nur die Schicht $Y_{m'}$, die die Region mit maximaler Aktivität enthält, formiert einen *Blob* und aktualisiert die Verbindungen zum *Blob* in Schicht X. Damit läßt sich ein Erkennungswert $r^m(t) \in [0,1]$ für ein Modell m zum gegebenen Iterationsschritt t angeben:

$$r^m(t_0) = 1 \qquad \dot{r}^m(t) = \lambda_r r^m \left(F^m - \max_{m'} \left(r^{m'} F^{m'} \right) \right) \tag{6.2}$$

F^m bezeichnet die *Fitness* der Schicht m, berechnet als die Summe der Aktivitäten aller ihrer Zellen. λ_r stellt eine Zeitkonstante dar. Die rechte Gleichung drückt einen Wettbewerb zwischen dem besten gefundenen Modell und allen anderen aus, welche zukünftig in ihrer Aktivität unterdrückt werden, da der letzte Term negativ wird.

- Es werden neue Schwellwerte eingeführt, um nur signifikante Übereinstimmungen zwischen Szene und Modell zu verstärken und damit auch eine Systembeschleunigung zu erreichen. Dazu werden Modelle m ignoriert, bei denen $r^m(t) < \theta_r$ ist. Weiterhin wird vor der Anpassung der Gewichte getestet, ob der zweitbeste Wert $V(b'_c)$ einen signifikanten Abstand zu $V(b_c)$ besitzt. Als Bedingung wird dazu $\frac{V(b_c)}{V(b'_c)} \geq k$ mit z.B. $k = 1.1$ festgelegt.

- Nach einer vordefinierten Anzahl von Iterationen wird ein Aufmerksamkeitsfenster gebildet, welches die Positionen des *Blobs* auf dessen Region beschränkt. Dadurch kann schneller zu einem stabilen Ergebnis gelangt werden. Es wird das Modell m' mit dem besten Erkennungswert benutzt, um auf der Basis seiner Verbindungsmatrix $J_{b^{m'}a}$ das Zentrum $(\bar{a}_x, \bar{a}_y)$ des Aufmerksamkeitsfensters zu berechnen:

$$\bar{a}_{x/y} \quad = \quad \sum_{b^{m'}} \sum_{a_{x/y}} a_{x/y} \cdot J_{b^{m'} a_{x/y}} \tag{6.3}$$

  Höhe und Breite des Fensters werden durch die Standardabweichungen $(2\sigma_{a_x}, 2\sigma_{a_y})$ bestimmt.

- Dem eigentlichen *Dynamic Link Matching* folgt ein Verarbeitungsmodul, welches entscheidet, ob die ausgewählte Modellansicht gut mit dem Szenenbild übereinstimmt. Es kommt ein Matchmaß unter Verwendung der relativen Häufigkeit der gleichzeitigen Aktivität zweier Zellen a und b, der Koaktivität C_{ba}, zum Einsatz. Die Koaktivitäten werden dabei als ein Gitter interpretiert: Positionen mit hohen C_{ba}-Werten stellen feste Stützstellen dar; an Punkten mit geringen Werten wird interpoliert. Bei guter Übereinstimmung wird die Merkmalsrepräsentation der Modellansicht entsprechend angepaßt. Ansonsten wird eine neue Ansichtenklasse angelegt.

In Abbildung 6.5 ist eine exemplarische Sequenz angegeben. Es wird jeweils versucht, eine präsentierte Ansicht den zuvor gelernten zuzuornden. Weiterhin sind die Koaktivitätswerte C_{ba} als Gitter interpretiert abgebildet (dunkle Verbindungen charakterisieren direkte Matchpunkte; helle Regionen schwache Korrespondenzen, bei denen interpoliert wurde). Aus dem Verzerrungsgrad des Gitters wird dann das entscheidende Matchmaß abgeleitet. Zwei geometrische Verfahren sind implementiert: Zum einen die Auswertung der Streuung der Knotenabstände und zum anderen die Analyse der Rechtwinkligkeit der Kreuzungslinien. Beide Verfahren lieferten ähnliche Ergebnisse.

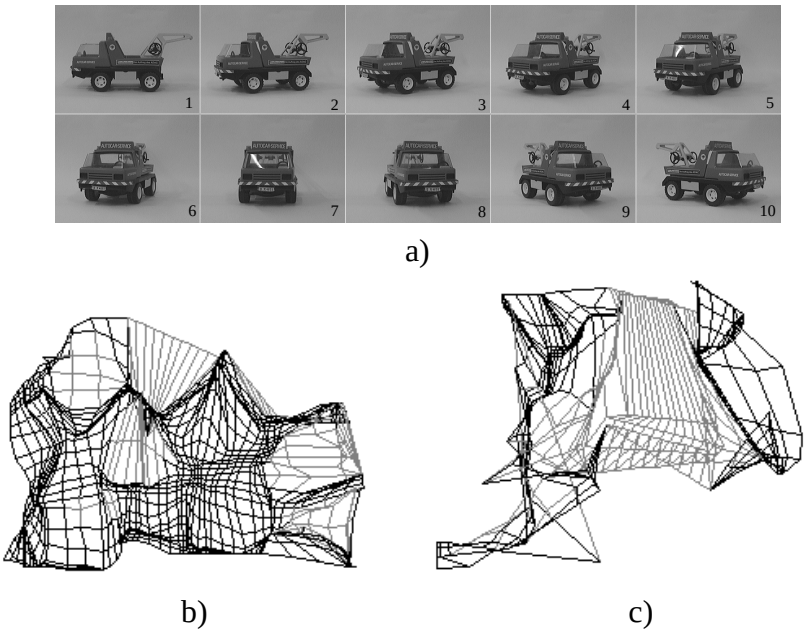


Abbildung 6.5: Ausbildung der Koaktivitätsgitter (nach [MASSAD et al. 1998a]). *a)* Präsentierte Sequenz. *b)* Koaktivitätsgitter beim Matchen von Ansicht 4 auf Ansicht 2. *c)* Gitter bei Präsentation von Ansicht 9. Es konnte kein Match mit den zuvor gelernten Ansichten erzielt werden.

6.2.3 Aufzeichnung zeitlicher Relationen

Bei der zeitlichen Aufzeichnung von Ansichten soll ein Kurzzeitgedächtnis modelliert werden, d.h. Aktivitäten sollen nicht nur das Vorhandensein eines Ereignisses anzeigen, sondern auch in welcher zeitlichen Relation sie zueinander stehen. Eine solche *item-and-order*-Darstellung wird durch die in [BRADSKI et al. 1992] vorgestellte *STORE*-Architektur aus der Klasse der *Su-*

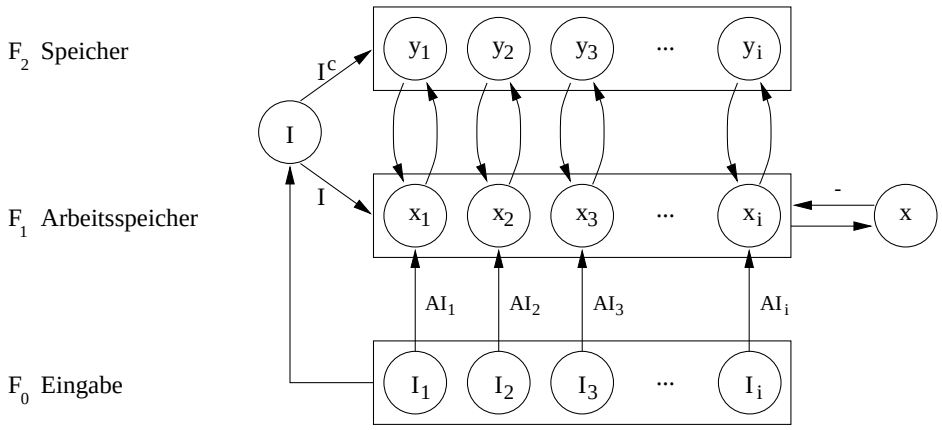


Abbildung 6.6: *STORE*-Architekur (nach [MASSAD et al. 1998a])

stained Temporal Order Recurrent Modelle erzeugt (siehe Abbildung 6.6). Die Architektur ist gekennzeichnet durch zwei gleichgroße, bidirektional verbundene Schichten F_1 und F_2, wobei F_1 weiterhin mit der Schicht F_0 verbunden ist. An die Eingabeschicht F_0 wird das Ergebnis der Ansichtenklassifikation angelegt, d.h. jeweils nur ein Element (bzw. eine Ansicht) besitzt die Aktivität $I_i = 1$. Die Eingaben werden über einen konstanten Faktor A, der den Einfluß der aktuelle Eingabe wichtet, an die Schicht F_1 angekoppelt. Über eine negative Rückkopplung aller Zellen in F_1 über die Zelle x, deren Aktivität der Summe x_i aller Zellen aus F_1 entspricht, findet eine Normierung der Schicht F_1 statt. F_2 fungiert als Speicher für zurückliegende Ereignisse, die erneut an F_1 angelegt werden. Neue Aktivitäten werden erst nach Abklingen der Aktivitäten in I_i übernommen. Die Zelle I sorgt für das wechselseitige Aktivieren der Schichten F_1 und F_2. An die Schicht F_1 kann dann eine Architektur zur Modellierung eines Langzeitgedächtnisses bzw. ein Objektklassifikationsmodul angekoppelt werden. Beispiele für generierte *STORE*-Vektoren können der Erkennungsdemonstration in Abbildung 6.9 entnommen werden. Es ist zu erkennen, wie frühere Klassifikationsergebnisse dem System weiterhin, wenn auch in abgeschwächter Form, zur Verfügung stehen. Damit kann eine anschließende Objektklassifikation in ihrer Entscheidungsfindung auch auf frühere Ergebnisse zurückgreifen.

6.2.4 Objektklassifikation

Zur Klassifikation der Objektsequenzen dient das in [WILLIAMSON 1996a] vorgestellte *Gaussian ARTMAP*-Netzwerk (siehe Abbildung 6.7), welches eine Verknüpfung eines Gaußschen Klassifizierers mit der *Adaptive Resonance Theo-*

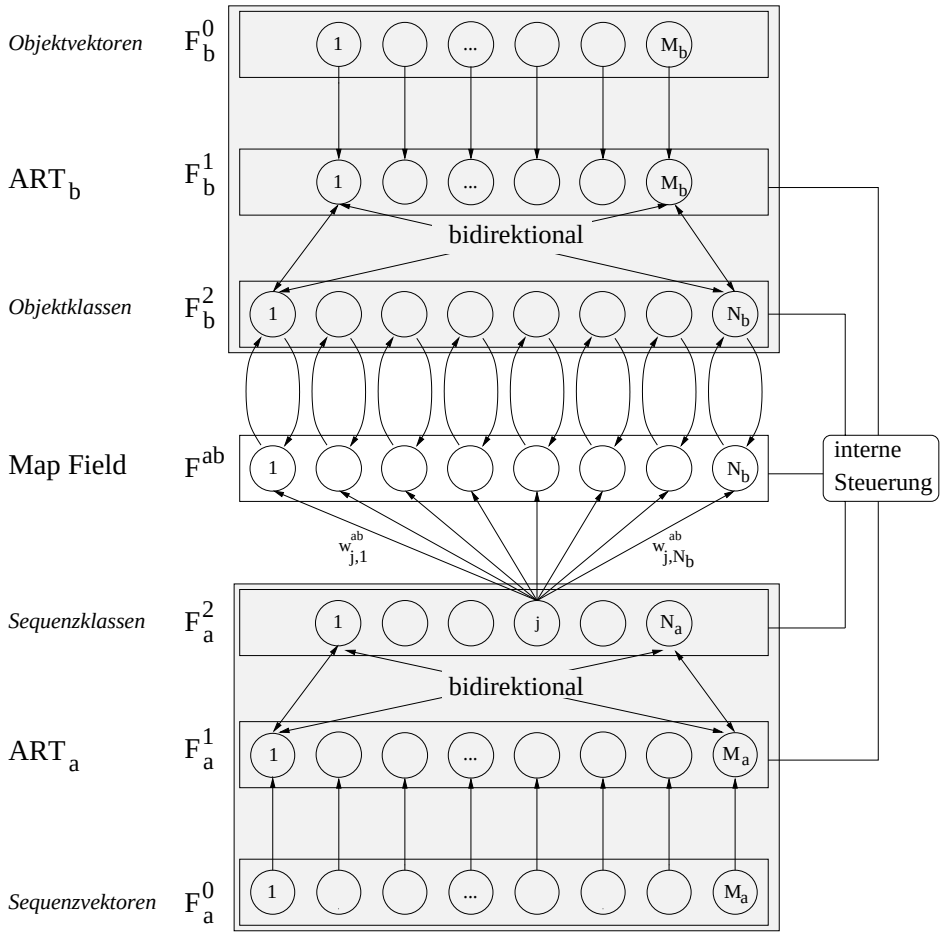


Abbildung 6.7: *ARTMAP*-Architekur (nach [MASSAD et al. 1998a])

ry nach [GROSSBERG 1976] darstellt. Die Architektur stellt eine Abwandlung der Fuzzy-*ARTMAP*-Netze von [CARPENTER et al. 1992] dar. Ziel war die Beseitigung verschiedener Schwachstellen, wie Rauschempfindlichkeit, u.U. ausufernde Klassenanzahl, Reihenfolgeabhängigkeiten u.a. (Näheres findet sich in [SARLE 1995]).

Das Netzwerk besitzt zwei *ART*-Teilnetze und ein *map field*, welches eine Zuordnung einer Klasse J von F_a^2 zu einer Klasse K in F_b^2 herstellt. Das Teilnetz *ART*$_a$ erhält an F_a^0 die vom *STORE*-Netz gelieferten M_a-dimensionalen Sequenzvektoren, d.h. Vektoren, die die zurückliegenden Klassifikationsergebnisse kodieren. Diese Vektoren werden in Sequenzklassen eingeteilt, die durch die Zellen F_a^2 repräsentiert werden. An *ART*$_b$ werden Objektmerkmale zu Objektklassen in F_b^2 klassifiziert. Da Objekte hier direkt als Objektklassen angegeben werden, redu-

ziert sich das Netz *ART*$_b$ zur Schicht F_b^2. Innerhalb dieses Netzwerkes werden die Klassen als Gaußverteilung modelliert. Jede Klasse i wird durch M-dimensionale Vektoren beschrieben, wobei die Mittelwerte μ_i und Standardabweichungen σ_i zusammen mit einem Zähler n_i für die Anzahl der präsentierten Muster gespeichert werden. Innerhalb der Lernphase wird zunächst mittels einer *choice*-Funktion eine Klasse in F_a^2 ausgewählt, die möglichst gut zum angelegten Eingangsmuster paßt. Anschließend wird durch eine *match*-Funktion überprüft, ob durch die Gewinnerzelle im *map field* eine Zelle aktiviert wird, die mit der gewünschten Objektklasse korrespondiert. Die dabei zugelassenen Abweichungen werden durch einen *vigilance*-Parameter bestimmt. Bei Übereinstimmung tritt der Resonanzfall ein und es werden die Gewichte der aktiven Zellen angepaßt. Bei keiner Übereinstimmung setzt ein *match tracking* ein, das die aktive Zelle in F_a^2 zurücksetzt, um so besser passende bzw. bisher untrainierte Klassen zum Matchen zuzulassen und diese Verbindungen entsprechend zu lernen. In der Klassifikationsphase wird nur das Teilnetz *ART*$_a$ benutzt. Eingaben werden an F_a^0 angelegt und im *map field* können die resultierenden Objektklassen abgelesen werden.

6.3 Experimente und Ergebnisse

Zunächst wird die reine Erkennungsleistung des *Dynamic Link Matching* untersucht. Zur Evaluierung werden 4 Objekte (siehe Abbildung 6.8) jeweils um ihre Hochachse rotiert und im ca. $15°$ Grad Abstand Ansichten generiert. Es stehen pro Objekt 25 Ansichten zur Verfügung (Abbildungen siehe Anhang). Davon werden die Ansichten mit ungeradem Index (insgesamt 48) als Modellansichten verwendet. Die Ansichten mit geradem Index (insgesamt 52) fungieren als zu klassifizierende Szenenbilder. In Tabelle 6.1 ist eine Gesamtübersicht über alle Klassifikationsergebnisse aufgeführt. Insgesamt konnten 50 Objekte richtig klassifiziert werden. 44 Ansichten ergaben die ähnlichste Ansicht aus der Menge der Modellansichten. Dies stellt in Bezug auf die Ähnlichkeit der einzelnen Modellansichten ein sehr gutes Ergebnis dar.

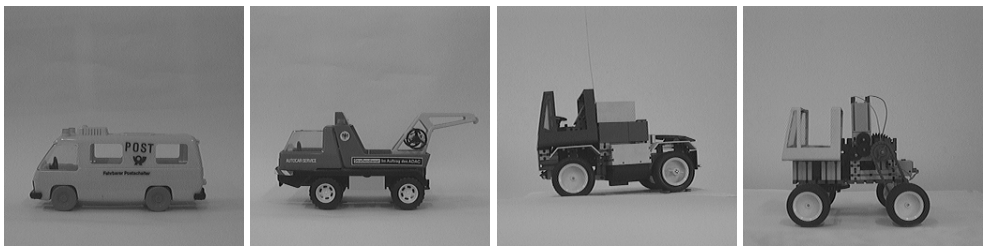


Abbildung 6.8: Die vier Modellobjekte (Index $0 \ldots 3$)

Tabelle 6.1: Modell- bzw. Ansichtenzuordnung für die 52 Szenenbilder. Die ersten beiden Spalten bezeichnen jeweils das Szenenobjekt S mit einer Ansichtennummer SA, die folgenden zwei Spalten das klassifizierte Objekt M mit der bestimmten Ansicht MA.

S	SA	M	MA
0	0	0	23
0	2	0	3
0	4	0	3
0	6	0	5
0	8	0	9
0	10	0	11
0	12	0	11
0	14	0	11
0	16	0	3
0	18	0	19
0	20	0	21
0	22	0	23
0	24	0	23
1	0	1	23
1	2	1	3
1	4	1	3
1	6	3	5
1	8	1	9
1	10	1	11
1	12	1	13
1	14	1	13
1	16	1	15
1	18	2	19
1	20	1	21
1	22	1	23
1	24	1	23

S	SA	M	MA
2	0	2	1
2	2	2	1
2	4	2	3
2	6	2	5
2	8	2	9
2	10	2	9
2	12	2	9
2	14	2	9
2	16	2	15
2	18	2	5
2	20	2	1
2	22	2	1
2	24	2	1
3	0	3	1
3	2	3	1
3	4	3	23
3	6	3	5
3	8	3	9
3	10	3	9
3	12	3	13
3	14	3	15
3	16	3	15
3	18	3	5
3	20	3	19
3	22	3	23
3	24	3	23

In Abbildung 6.9 ist ein beispielhafter Erkennungsvorgang einer Ansichtensequenz abgebildet. Hier findet sowohl die Zuordnung der ähnlichsten Ansicht mittels *Dynamic Link Matching* als auch deren zeitliche Auswertung statt. In der linken Spalte sind die gegebenen Ansichtenbilder zu sehen. Zu beachten sind die beiden nicht eindeutigen Ansichten in der dritten und vierten Zeile. Die zweite Spalte zeigt das Klassifikationsergebnis auf der Basis des *Dynamic Link Matching*. Zu sehen sind jeweils die Vertreter der matchenden Ansichtenklasse. Die

dritte Spalte zeigt die Entwicklung der zeitlichen Aufzeichnung der Klassifikationsergebnisse. Gezeigt werden die Aktivitäten (Ordinate) der Komponenten (Abzisse) des *STORE*-Netzes. Die Komponenten korrespondieren zu den im unteren Teil angegebenen Objektzuordnungen. In der letzten Spalte ist das resultierende Klassifikationsergebnis durch Auswertung der zeitlichen Abfolge angegeben. Durch die Ausnutzung des temporalen Kontexts konnten die ursprünglichen Fehlzuweisungen der Ansichten drei und vier korrigiert werden.

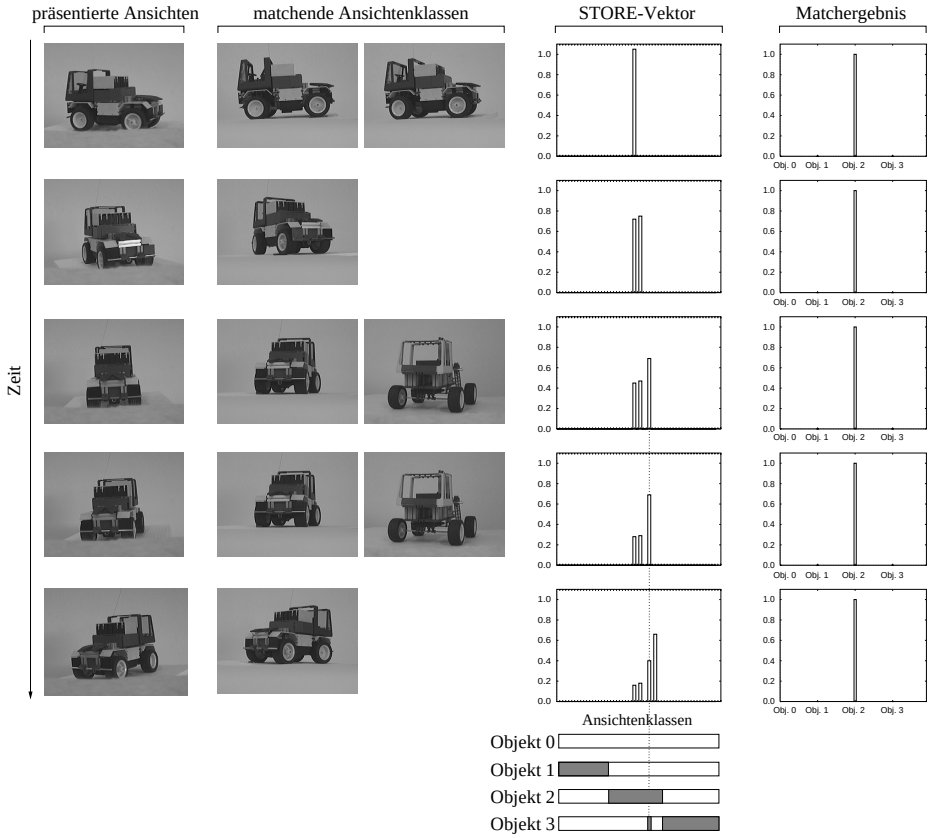


Abbildung 6.9: Erkennung einer Beispielsequenz (Erläuterungen siehe Text; nach [MASSAD et al. 1998b])

Durch das vorgestellte Erkennungssystem konnte erstmalig im NAVIS-Projekt eine ansichtenbasierte 3D-Objekterkennung realisiert werden. Dabei wurden selbständig zu lernende Ansichten ausgewählt und im Merkmalsraum repräsentiert. Die durch mehrdeutige Ansichten hervorgerufenen Zuordnungsfehler konnten durch eine Analyse des zeitlichen Kontexts korrigiert werden. Weiterhin konnte eine im Vergleich zu den bisher vorgestellten Erkennungssystemen (siehe Kapitel 4.3 und 5) höhere Erkennungsstabilität erreicht werden:

- Die zuvor kritisierte starre Repräsentation von Attraktivitätspunkten und Objekten wurde überwunden. Stattdessen werden Objektklassen durch ihre Repräsentanten im Merkmalsraum beschrieben.

- Es wurden stärkere Invarianzen integriert. Eine Translationsinvarianz wurde durch einen Aufmerksamkeitsmechanismus bereitgestellt. Dabei wird während des *Dynamic Link*-Matchens ein Bearbeitungsfenster basierend auf Positionen von aktivierten Zellen abgeleitet. Alle weiteren Iterationen finden dann nur noch innerhalb dieses Fensters statt. Dadurch wurde ebenfalls der Rechenaufwand erheblich reduziert. Kleinere Drehungen werden ebenfalls toleriert. Ausschlaggebend hierzu ist die gewählte Winkelauflösung während der Vorverarbeitung mittels Gaborfilter. Eine Skalierungsinvarianz ist ebenfalls gegeben, da zur Ableitung des Matchmaßes relative geometrische Merkmale aus dem Koaktivitätsgitter verwendet werden.

- Uneindeutige Objektrepräsentationen können durch eine Sequenzanalyse aufgelöst werden.

Hinsichtlich der gewählten Strategie zur Ableitung des Aufmerksamkeitsfensters sind weitere Verbesserungen denkbar. Die rein rechteckig festgelegte Grundform wird vielen Objekten nicht gerecht, was zum Mitlernen von Hintergrundinformationen führen kann. Ein ideales Aufmerksamkeitsfenster würde sich an das zu untersuchende Objekt anpassen.

Die als Entwurfsziel zunächst nicht vorgesehene Anbindung an ein aktives Sehsystem sollte bei Weiterentwicklungen unbedingt genutzt werden. Neben den bereits ausführlich diskutierten Vorteilen wären auch Möglichkeiten zum selbständigen Erfassen von Objektsequenzen nutzbar. In diesem Zusammenhang scheint auch eine enge Kopplung mit Trackingverfahren sinnvoll. Der enorme Rechenaufwand zum Matchen der 24-dimensionalen Merkmalsvektoren ist weiterhin problematisch. Im Testbetrieb war dazu eine parallele Implementation auf einer Vielzahl leistungsfähiger Workstations notwendig, um zu akzeptablen Antwortzeiten zu kommen. Somit erscheint das Arbeiten auf größeren, komplexeren Bildern mit stark strukturierten Hintergrund (kein Aufmerksamkeitsfenster) nicht effektiv durchführbar. Weiterhin muß festgestellt werden, daß die Merkmalsrepräsentation nur auf den *Gaborjets* basiert, wodurch sich das Verfahren auf Objekte, die gut auf diese Merkmalsextraktion ansprechen, einschränkt.

Beim Vergleich mit dem in Tabelle 3.3 aufgestellten Anforderungskatalog an 3D-Objekterkennungssysteme ist festzustellen, daß hier bezüglich Erkennungsleistung, Invarianzen, Einsatzflexibilität, einer selbständigen Auswahl von Objektansichten und biologischer Plausibilität für die Objektrepräsentation gute Ergebnisse erzielt werden konnten. Vor allem Probleme im Zusammenhang mit der Parameterfestlegung, der Speichereffizienz, den erzielbaren Antwortzeiten und oben er-

wähnter Ankopplung an Aktorik sind nicht vollständig befriedigend gelöst. Der Umgang mit Okklusionen wurde in diesem Stadium der Entwicklung nicht weiter untersucht. Als besonders positiv ist die Möglichkeit der zeitlichen Organisation von Objektansichten und deren Nutzung zur Validierung von Klassifikationsentscheidungen zu sehen. Das Potential des Ansatzes scheint durch diese Korrektur unwahrscheinlicher Ansichtenklassifikationen noch nicht ausgeschöpft. Möglich ware u.a. eine Kopplung mit Trackingmechanismen zur Vorhersage von Bewegungen sowie das selbständige Aufstellen typischer Bewegungsprofile.

Kapitel 7

Objekterkennung mit Strukturbeschreibungen

Im Kapitel 3 wurden eine Reihe von Objekterkennungssystemen vorgestellt. Eine Wertung der einzelnen Ansätze fand ebenfalls statt. Ein Anforderungskatalog für ein in einer aktiven Sehumgebung agierendes Objekterkennungssystem wurde in diesem Zusammenhang ebenfalls aufgestellt (siehe Tabelle 3.3). Ausgehend von den geschilderten Problemen vor allem hinsichtlich des getriebenen Aufwandes und den damit erzielbaren Ergebnissen soll nach Wegen gesucht werden, unter Beibehaltung der Ergebnisstabilität der Erkennung mit zeitlichen Relationen (siehe Kapitel 6) Repräsentationen deutlich geringerer Komplexität einzusetzen. Ziel ist das dortige Ansichtenerkennungsmodul, das *Dynamic Link Matching*, durch ein alternatives zu ersetzen. Entsprechend dem Anforderungskatalog soll zusätzlich der Punkt Speichereffizienz befriedigend gelöst werden. Im folgenden werden entsprechende Lösungen vorgeschlagen, die schließlich zu einem implementierten prototypischen System führen [SCHMALZ und MERTSCHING 1999].

Im Kapitel 4.3 wurde ein z.Z. innerhalb des NAVIS-Projektes eingesetztes Verfahren dargestellt und analysiert. Es wurde festgestellt, daß die Benutzung von wenigen stabilen sog. Attraktivitätspunkten vielversprechend erscheint. Es konnten gute Ergebnisse innerhalb einer stark strukturierten Szene erzielt werden. Probleme entstanden durch die starre Anordnung der Punkte, wodurch nur geringe Erkennungsinvarianzen bereitgestellt werden konnten. Im Kapitel 5 wurde ein System zu einer skalierungsinvarianteren Erkennung vorgestellt. Das Grundprinzip der Erkennung blieb aber gleich. Dementgegen steht das im Kapitel 6 vorgestellte System. Durch die Verwendung hochdimensionaler Merkmalsvektoren wird versucht, eine möglichst einmalige Beschreibung für ein zu lernendes Objekt zu erhalten. Durch die enorme Rechenkomplexität mußten die Bilder entsprechend bis zur Größe 64×64 unterabgetastet werden. Bei dieser enormen Verkleinerung geht allerdings das Indivuduelle eines Objekts schon wieder verloren, so daß auch

keine sehr guten Erkennungsleistungen hinsichtlich der Ansichtenklassifikation erreicht werden konnten. Die entsprechenden Probleme wurden dann durch Auswertung des zeitlichen Kontexts kompensiert.

Beide Systeme sollen hier beim Aufbau eines neuen Erkennungssystems einfließen. Zum einen soll weiterhin das Diskriminierungspotential von charakteristischen Punkten, den Attraktivitätspunkten, genutzt werden. Allerdings dürfen diese nicht starr angeordnet werden. Vielmehr müssen geeignete Verfahren zur Adaption der Punktekonfigurationen an die zu erkennenden Objekte bzw. die zu untersuchenden Szenen gefunden werden. Gerade diese dynamische, iterative Anpassung von gelernten Konfigurationen an präsentierte Objekte erwies sich im Erkennungssystem mit zeitlichen Relationen als tauglich. Als Nahziel kann somit formuliert werden, daß ausgehend von "attraktiven" Punkten innerhalb der Szene ein Repräsentationsschema aufgebaut werden soll, das Möglichkeiten zu einer dynamischen Anpassung an andere Objekte bzw. Szenen bietet. Die Selektion von Objektpunkten als attraktiv im Sinne z.B. einer Aufmerksamkeitssteuerung und deren elastische Verkopplung untereinander könnte als Kompromiß zwischen der zu starren Attraktivitätsrepräsentation nach Götze und Mertsching und der hochdimensionalen, bereits beschriebenen 3D-Objekterkennung fungieren. Ist eine entsprechende Repräsentation gefunden, stellt sich die Frage nach angepaßten Matchstrategien. Diese müssen explizit auf die dynamischen Eigenschaften des Repräsentationsschemas aufsetzen. In den folgenden Kapiteln werden zunächst die verwendeten Merkmale als Basis einer Repräsentation vorgestellt, die Repräsentation selbst und schließlich mögliche Matchstrategien.

7.1 Repräsentationsschema

Das Repräsentationsschema ist eminent wichtig für die Leistungsfähigkeit eines Objekterkennungssystems. Mehrere konträre Anforderungen müssen berücksichtigt werden. Zum einen sollen möglichst einfach zu berechnende Informationen zur Diskriminierung eines bestimmten Objekts bereitgestellt werden, allerdings verlangt die Forderung nach Invarianzen die Beschränkung auf stabile, hinsichtlich Transformationen robuste Merkmale mit entsprechend höherem Berechnungsaufwand.

Die angestrebte Verwendung von attraktiven Punkten zur Charakterisierung von Objekten zielt auf eine lokale Repräsentationsform ab, d.h. Objekte werden durch Anordnungen von solchen Punkten beschrieben und nicht durch globale Beschreibungen. Zur effektiven und flexiblen Verwaltung dieser Punkte bieten sich Graphenstrukturen an. Diese werden in der Literatur vielfach innerhalb von Erkennungssystemen verwendet (z.B. [BUNKE und MESSMER 1995], [CHUNG 1995], [BURGER et al. 1996]; weitere siehe Kapitel 3.3). Hervorgeho-

ben wird die besondere Eignung von Graphen für Repräsentationsaufgaben. Die Erkennungsaufgabe stellt sich als das Herauslösen bzw. Segmentieren von Objekten aus der Szene, dem Aufbau von Graphstrukturen und dem Vergleich mit der Menge der Modellgraphen. Das Objekterkennungsproblem wird als Graphmatchproblem umformuliert. Durch die Anwendung von Graphen als lokale Repräsentationsform ergibt sich eine symbolische Beschreibung des strukturellen Aufbaus eines Objektes, d.h. eine Beschreibung von Objektmerkmalen bzw. Objektteilen und deren lokale räumliche Beziehung zueinander, wodurch sich direkt eine invariante Objektbeschreibung bezüglich Größe und Verschiebung ableitet. Problematischer erscheint hierbei die Bereitstellung von Rotationsinvarianz. Diese muß durch geeignete Mechanismen bereitgestellt werden (z.B. in [BENNAMOUN und BOASHASH 1997] erläutert). Durch die Favorisierung eines ansichtenbasierten Erkennungssystems kann das Fehlen von Rotationsinvarianz durch eine höhere Anzahl zu berücksichtigender Ansichten ausgeglichen werden.

Das wachsende internationale Interesse an Strukturbeschreibungen zur Objektrepräsentation spiegelt sich in der zunehmenden Anzahl von Veröffentlichungen wider. Hierbei finden sich häufig die Begriffe *structural description*, *structural representation* oder *structural matching* (z.B. in [AMIN et al. 1998], [PELILLO und HANCOCK 1997], [PONCE et al. 1996], [HLAVÁČ und ŠÁRA 1995]). Strukturbeschreibungen von Objekten (eingesetzt z.B. auch in [CHAO und DHAWAN 1992], [SHAMS 1995], [WILLIAMSON 1996b], [BENNAMOUN und BOASHASH 1997] und den Arbeiten von Shapiro et al.) sind eine direkte Anwendung der Vorteile graphenorientierter Repräsentationsformen.

Als endlicher Graph G wird i.a. das geordnete Paar $G = (V, E)$ disjunkter Mengen mit $E \subseteq [V]^2$ verstanden ([DIESTEL 1996], [BODENDIEK und LANG 1995]). Mit V werden die Ecken oder Knoten des Graphen G bezeichnet; mit E seine Kanten. Innerhalb dieser Arbeit besteht das Ziel u.a. darin, zwei gegebene Graphen in Übereinstimmung zu bringen, ihre Isomorphieeigenschaften zu bestimmen (nähere Ausführungen dazu finden sich im Kapitel 3.3.2). Innerhalb dieses Kapitels wird häufig vom *Match* zweier Graphen gesprochen. Der Begriff *Match* wird hier anders gebraucht als in der klassischen Graphentheorie. Dort bezeichnet der Match bzw. die Paarung eines Graphen die Menge $M \subseteq E$, wobei je zwei verschiedene Kanten $e, e' \in M$ unabhängig sind, d.h. wenn sie in G nicht benachbart sind bzw. keine gemeinsame Ecke in G besitzen [BODENDIEK und LANG 1995]. Hier wird sich mit dem Begriff *Match* auf die übereinstimmenden Eigenschaften zweier Knoten zweier Graphen bezogen, d.h. zwei Knoten $V_i \in G_1$ und $V_j \in G_2$ matchen, wenn für eine gegebene Strukturähnlichkeitsfunktion S gilt: $S(V_i, V_j) > \epsilon$.

Der große Vorteil graphenbasierter Erkennungssysteme ist, daß dort Wissen über die Objekte explizit genutzt wird im Gegensatz zur impliziten Nutzung z.B.

bei künstlichen neuronalen Netzen. Dadurch ist die Wartbarkeit und Erweiterbarkeit derartiger Systeme unproblematischer. Ein weiterer Vorteil expliziter Repräsentationen ist deren Modifizierbarkeit während des Matchprozesses. Es ist möglich, Modell- bzw. Szenengraphen sich dynamisch an eine Vorgabe anpassen zu lassen (eingesetzt z.B. in [CHUNG 1995]). Hier sind Begriffe wie *aktive Konturen* üblich.

Bisher nicht vollständig gelöst ist das Problem der starken Abhängigkeit von der Güte der bereitgestellten Merkmale zum Aufbau von Objektgraphen. Häufig war deshalb eine Beschränkung auf technische oder sehr primitive Objekte, zum Teil ohne störenden Hintergrund, notwendig. Oft wird somit die eigentliche Schwierigkeit des Matchens eines unbekannten Objekts verlagert zu der möglichst optimalen Segmentierung einer Szene oder der Generierung idealer Objektkanten. Allerdings ist ja gerade die Verwendung von dynamischen Graphenstrukturen geeignet, um die unvermeidbaren Fehler während der Segmentierung bzw. Kantengenerierung zu kompensieren.

7.1.1 Merkmalsbasis

Die Merkmalsbasis zum Aufbau von Objektgraphen soll inhärente Objekteigenschaften beschreiben und die Struktur eines Objekts widerspiegeln. Die Verwendung von geometrischen Strukturen bietet sich auf Grund ihrer positiven Eingenschaften bei verschiedenen Beleuchtungsbedingungen, Betrachtungswinkeln und Objekttransformationen an und wurde in der Literatur zahlreich vorgeschlagen (z.B. [BASAK und PAL 1995], [COSTA und SHAPIRO 1996], [BENNAMOUN und BOASHASH 1997]). Diese geometrischen Primitive sind zum einen einfach berechenbar und bieten durch ihre Lokalität und räumliche Relation gute Diskriminierungseigenschaften.

Verwendung von Aufmerkamkeitspunkten aus dem NAVIS-Kontext

Zunächst werden die im NAVIS-Projekt bereits realiserten Verfahren zur Extraktion von eminenten Bildelementen (Flächenschwerpunkte, Symmetriepunkte) zum Aufbau einer Merkmalsbasis verwendet (siehe Kapitel 4.2).

Die in [MERTSCHING und BOLLMANN 1997] weiterhin vorgestellte Farbanalyse der Szene wird hier aus Komplexitätsgründen nicht eingesetzt. Es muß aber betont werden, daß auf spezifische Farbeigenschaften basierende Punkte zur Robustheit eines Erkennungsystems beitragen könnten. Entsprechende Untersuchungen und Experimente sind im Rahmen der Weiterentwicklung des Systems zu berücksichtigen.

Probleme ergeben sich bei den angegebenen Verfahren vor allem durch die sehr komplexe Parametrisierung. Wird versucht, für eine große Domänenanzahl

gültige Einstellungen zu finden, entstehen sehr viele Aufmerksamkeitspunkte in den Szenen, wodurch sich eine wenig diskrimierende Repräsenation ergibt. Experimente unter Verwendung der Flächenschwerpunkte ergaben oft eine wenig spezifische Repräsentation der Szene, die nur durch eine entsprechende Parameteranpassung wieder verbessert werden konnte. Abbildung 7.1 zeigt ein exemplarisches Objekt mit gekennzeichneten Merkmalspunkten. Die Parametrisierung wurde hier so gewählt, daß möglichst wenige Attraktivitätspunkte entstehen. Schwarze Kreuze markieren Flächenschwerpunkte, weiße Kreuze Symmetriepunkte.

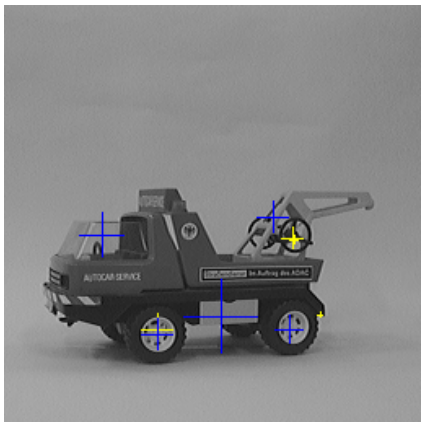


Abbildung 7.1: Beispielobjekt überlagert mit markierten Attraktivitätspunkten

Verwendung von primären Strukturmerkmalen

Alternativ zu den im vorigen Kapitel dargestellten Merkmalen werden weitere herangezogen. Im Kapitel 3.3 wurden eine Reihe geometrischer Verfahren vorgestellt. Es zeigte sich, daß geometrische Primitive als Merkmale eingesetzt werden, da diese höhere Diskriminierungseigenschaften als Grauwertmerkmale besitzen. Hier kommen zum Einsatz:

- besonders intensive ausgeprägte Kanten,
- Ecken mit Begrenzungen bezüglich des eingeschlossenen Winkels und
- Mittelpunkte von symmetrische Strukturen.

Für die ersten beiden Merkmale werden die Verfahren nach [LIEDER 1996] eingesetzt. Es werden Regionen gesucht, in denen mehrere Geradenstücke radial um das Zentrum angeordnet sind. Eine Ecke wird definiert als ein Punkt,

an dem mindestens zwei orthogonal zueinander stehende Geradenstücke lokalisiert sind. Die untersuchten Geradenabschnitte müssen sich im betrachteten Punkt nicht berühren, um innerhalb der Konturen Lücken zuzulassen (siehe auch Abbildung 7.2). Diese Verarbeitung ist seinerseits durch die FACADE-Theorie von Grosberg motiviert ([GROSSBERG 1980, GROSSBERG 1990], [NEUMANN und STIEHL 1992]). Dort sollen dynamisch geschlossene Konturen erzeugt werden. Für jeden Bildpunkt wird ein Entscheidungsmaß zur Zugehörigkeit zu einem längeren Kantenstück oder einer Ecke durch eine Summation der Aktivitäten innerhalb der Detektormasken angegeben. Als Vorverarbeitung zur Bestimmung der Kanten kommt die NAVIS-typische Orientierungszerlegung zum Einsatz (siehe auch Kapitel 4.2). Es entstehen zwei Ergebnisbilder entsprechend den detektierten Kanten und Ecken. Anschließend finden noch diverse Maximum- und Schwellwertoperationen statt, um zum einen unempfindlicher gegenüber Rauschen zu sein und zum anderen nicht entlang von schrägen Kantenstücken Ecken zu detektieren.

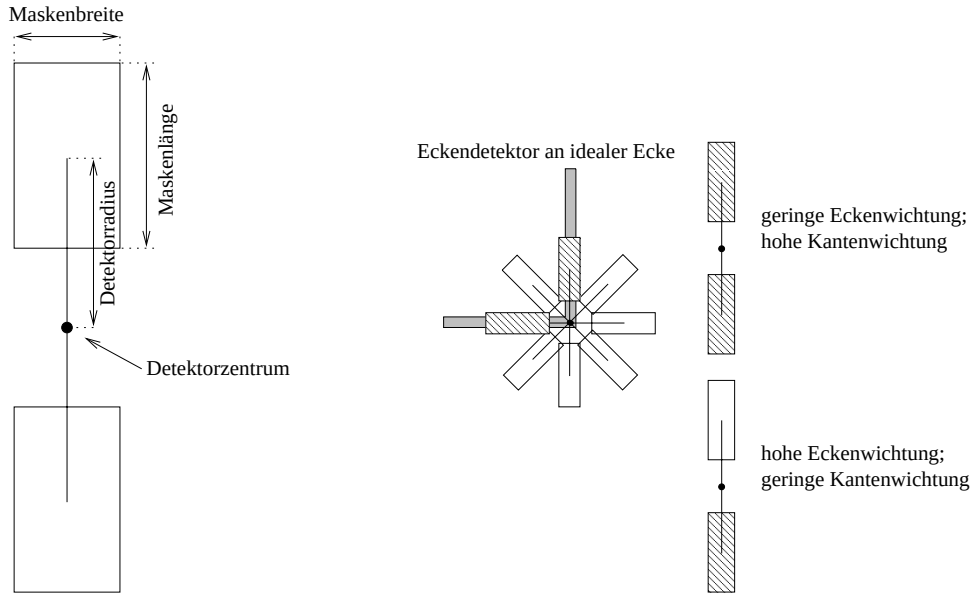
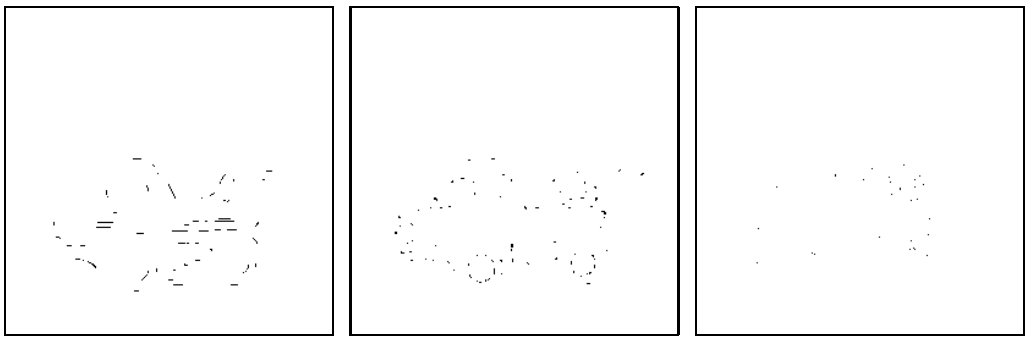


Abbildung 7.2: *a)* Aufbau eines Detektors für eine Orientierung; *b)* Funktionsprinzip der Summation, die schraffierten Flächen geben Detektorflächen mit passenden Grauwertkanten an (nach [LIEDER 1996])

Zur Extraktion von symmetrischen Strukturen wird wieder das Verfahren aus dem vorigen Kapitel eingesetzt. Die Abbildung 7.3 zeigt die generierten Merkmale für das in Abbildung 7.1 gezeigte Objekt.



a) b) c)

Abbildung 7.3: Merkmalskarten für das Beispielobjekt aus Abbildung 7.1. *a)* eminente Kanten; *b)* Ecken; *c)* Symmetriepunkte

Die gewonnenen Merkmale können zum Aufbau der Repräsentation priorisiert werden. Folgenden Gewichtsfaktoren werden verwendet ($\sum w_i \neq 1$ wegen Erweiterungsfähigkeit):

1. Ecken ($w_e = 0.35$),
2. Kreisstrukturen ($w_{kr} = 0.35$) und
3. Kanten ($w_{ka} = 0.1$)

Die Merkmale werden wegen der Handhabbarkeit im folgenden teilweise als Bitvektoren repräsentiert. Entsprechend obiger Priorisierung ergibt sich folgende Bitzuordnung: Bit 0 kodiert Kanten, Bit 2 Kreisstrukturen und Bit 3 Ecken (Bit 1 ist für Erweiterungen noch frei). Aus der Priorisierung kann pro Bildpunkt ein *interest value* abgeleitet werden. Anhand dieses Wertes werden die dynamischen Eigenschaften der Repräsentation (siehe folgendes Kapitel) gesteuert.

7.1.2 selbstorganisierende Repräsentation

Nach der Festlegung der zu verwendenden Merkmalsbasis und der Favorisierung graphenverwandter Repräsentationsverfahren stellt sich die Frage nach der Generierung dieser Graphenstrukturen. Die Merkmale sind durch kartesische Bildkoordinaten gegeben. Dementsprechend sind folgende Verfahren denkbar:

- Alle Punkte werden als Knoten im Objektgraphen betrachtet. Diese äußert triviale Vorgehensweise scheint zunächst sinnvoll. Allerdings wird spätestens bei der Suche nach sinnvollen Kantenbeziehungen zwischen den Knoten bewußt, daß diese Art *brute force*-Generierung starke Mängel aufweist.

Weiterhin ist mit der Merkmalsberechnung keine Beurteilung der räumlichen Verteilung der Merkmale verbunden, d.h. unter Umständen entstehen in bestimmten Bildregionen Merkmals-Wolken aus dicht beieinander liegenden Merkmalspositionen (diese Wolken werden durch den Symmetrieoperator häufig innerhalb kreisartiger Strukturen generiert). In diesem Falle jedem Merkmalspunkt einen Graphenknoten zuzuordnen erscheint äußert fragwürdig. Solche Objektgraphen würden gerade nicht die typische Struktur des Objektes widerspiegeln, sondern nur spezifische Merkmalsausprägungen. Abbildung 7.4 verdeutlicht die Probleme.

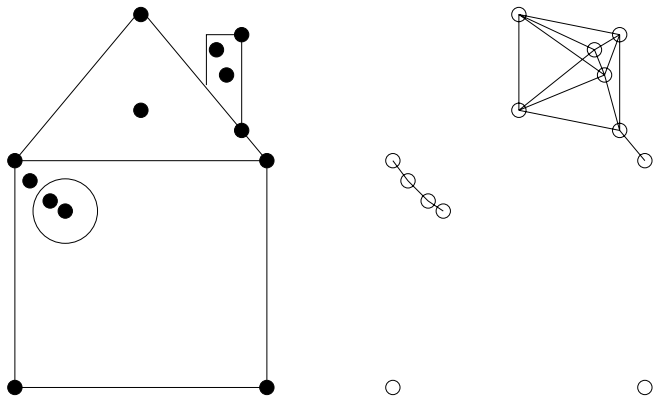


Abbildung 7.4: Beispielobjekt mit überlagerten Merkmalspunkten und sich ergebenden Objektgraphen bei direkter Übernahme aller Punkte in die Graphendarstellung. Zur Generierung der Graphenkanten werden bis zu einer bestimmten Entfernung alle benachbarten Knoten verkettet.

- Ein selbstorganisierendes Verfahren versucht, die gegebene Menge an Merkmalspunkten durch eine sinnvolle Menge an Graphenknoten bzw. -kanten zu approximieren. Wesentlicher Aspekt hierbei ist, daß nicht jeder Merkmalspunkt in die endgültige Repräsentation einfließen muß. Merkmalswolken sollten nur durch wenige Knoten repräsentiert werden, isolierte Merkmalspunkte sollten erhalten bleiben, da diese evtl. eine typische Objekteigenschaft darstellen und zur Diskriminierung bezüglich ähnlicher Objekte hilfreich sein können. In Abbildung 7.5 wird eine solche angestrebte Objektrepräsentation angeben. Obwohl der Objektgraph zunächst als verzerrtes Originalobjekt erscheint, hat die Darstellung doch deutliche Vorteile gegenüber der oben angegebenen Repräsentation. Der generierte Graph kann symbolisch weiterverarbeitet werden und die Knoten mit ihren räumlichen Relationen spiegeln die Strukur des Objekts wider.

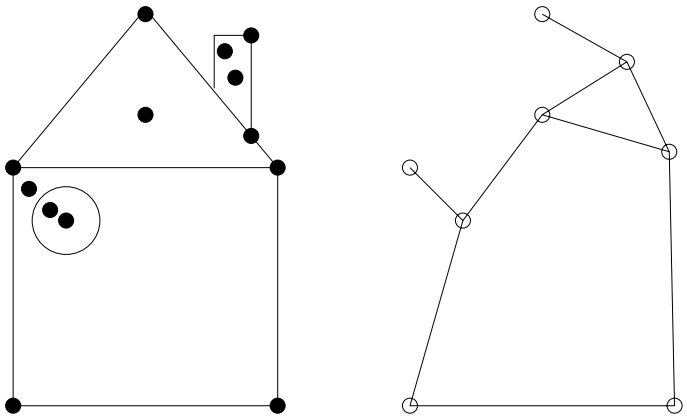


Abbildung 7.5: Beispielobjekt mit überlagerten Merkmalspunkten und erwünschter Objektgraph zur Verdeutlichung der grundlegenden Struktur des Objekts

Als Basis für die Generierung der Objektgraphen durch selbstorganisierende Verfahren wird der *Neural Gas*-Ansatz von Fritzke verwendet ([FRITZKE 1993], [FRITZKE 1995a], [FRITZKE 1995b], [FRITZKE 1995c], [FRITZKE 1996]), welches seinerseits auf Ideen von Martinetz und Schulten zurückgeht [MARTINETZ und SCHULTEN 1991]. Eingesetzt werden solche Verfahren zur Datenvisualisierung, Mustererkennung und Funktionsapproximierung. Grundidee ist ein inkrementeller Aufbau einer Netzwerkstruktur gemäß lokalen statistischen Maßen (*competetive hebbian learning*). Derartige wachsende oder adaptive Netzwerkstrukturen haben keine vordefinierte Topologie. Es findet vielmehr ein Prozeß der Anpassung des Netzwerkes an die Verteilung der gegebenen Daten statt. Lernalgorithmen dieser Art werden in der Literatur auch als *constructive learning algorithms* bezeichnet ([YEUNG 1993], [BERTHOLD und DIAMOND 1995], [BERTHOLD 1996], [YANG et al. 1997], [BERTHOLD und DIAMOND 1998], [PAREKH et al. 2000]). Die eingesetzten Knoten werden in Anlehnung an künstliche neuronale Netzwerke als Neuronen bezeichnet. Die sonst so entscheidende Übertragungsfunktion von künstliche Neuronen ist hier zunächst ohne Bedeutung, da hier nur die Herausbildung der Netzwerkstruktur von Interesse ist. Weiterhin sind die Verbindungen zwischen den Neuronen nicht gewichtet. Damit grenzt sich der *Neural Gas*-Ansatz deutlich gegenüber herkömmlichen *Multilayer Perceptron*-Ansätzen ab. Ausgangspunkt ist eine zufällige Verteilung von Neuronen. Im Gegensatz zu dem Ansatz von Martinetz und Schulten muß bei Fritzke keine a priori Festlegung der Anzahl der Neuronen durchgeführt werden. Die initiale Anzahl von Neuronen, deren Nachbarschaftsbeziehungen und die Dimensionalität kann sich entsprechend der zu approximierenden Merkmalsverteilung $P(\xi)$ ändern. Diese Art

der Aufteilung des Datenraums ist vergleichbar mit der *Voronoi*-Zerlegung. Hierbei wird eine Anzahl Zentren in R^n bestimmt und die Daten nach ihrer minimalen euklidischen Distanz zu diesen Zentren zugeordnet. Es entstehen sog. Parzellen, die sich bezüglich ihrer Größe an die zu approximierende Funktion adaptieren. Die Zentren für Datenraum-Zerlegung werden häufig durch Verfahren zur Vektorquantisierung bestimmt. Ein Problem ist hierbei, daß diese Verfahren oft nur die Häufigkeit der Eingabedaten auswerten und weniger die Verteilung der innerhalb der Parzellen vorhandenen Daten. Dadurch können Regionen mit vielen Zentren entstehen, obwohl die dort vorhandenen Daten einfache, lineare Zusammenhänge darstellen. Ein weiteres Problem ist, daß sowohl die *Voronoi*-Zerlegung als auch die Vektorquantisierung keine Nachbarschaftsbeziehungen herstellen bzw. auswerten. Dieser entscheidende Vorteil führt zur Favorisierung des *Neural Gas*-Ansatzes. Folgende Parameter steuern die Entfaltung eines *Neural Gas*-Netzes:

- a_{max}: das maximale "Alter" von neuronalen Verbindungen. Jeder neuronalen Verbindung c wird ein Attribut "Alter" $a(c)$ zugeordnet, das die Signifikanz dieser Verbindung während des Iterationsprozesses repräsentiert. Redundante Verbindungen besitzen dabei ein hohes "Alter" $a(c)$ und werden gelöscht. Diese Strategie führt zu wenigeren und robusteren Verbindungen.

- E_b, E_n: Verschiebungsfaktoren $\in$ [0,1]. Neurone werden von ihren alten Positionen l_{alt} zu den neuen l_{neu} in Abhängigkeit der euklidischen Distanz d zwischen l_{alt} und der zufällig generierten Position l_{zuf} verschoben: $l_{neu} = l_{alt} + E \cdot d(l_{alt}, l_{zuf})$ mit $E \in \{E_b, E_n\}$ und $E_b > E_n$.

- α, β: Faktoren, um die sich die Fehlerwerte der Neuronen verringern.

Somit ergibt sich der Algorithmus von Abbildung 7.6 für die Ausbildung eines *Neural Gas*-Netzes nach [FRITZKE 1995b].

Der ursprüngliche Algorithmus mußte grundsätzlich überarbeitet werden, um das in Bezug auf die angestrebte Objektrepräsentation gewünschte Verhalten zu erreichen:

- Innerhalb von Regionen mit wenigen Merkmalen sollen auch wenige Neuronen erzeugt werden. Obiger Algorithmus erzeugt gerade dort dichte Neuronenwolken.

- Im Rahmen eines Objekterkennungssystems reicht es nicht aus, nur ein zufälliges Bildsignal ξ zu erzeugen. Wichtig ist ebenso, daß auch Merkmale an der entsprechenden Position vorhanden sind. Dazu wird zunächst ein rezeptives Feld f_{rez} pro Neuron eingeführt, um deren Lokalität hervorzuheben. Weiterhin wird ein *interest value* aus der Wichtung der im rezeptiven Feld f_{rez} vorhandenen Merkmale abgeleitet.

Netzgenerierung — Originalalgorithmus nach [FRITZKE 1995b]

Beginne mit zwei Neuronen a und b an zufälligen Positionen w_a und w_b in $\mathbf{R}^n$

Generiere ein Eingangssignal ξ gemäß $P(\xi)$

Suche das räumlich nächste Neuron s_1 und das zweitnächste Neuron s_2

Inkrementiere das Alter aller Verbindungen $a(k)$, die von s_1 ausgehen

Addiere die quadrierte Distanz zwischen dem Eingangssignal ξ und dem nächsten Neuron s_1 zu einer Fehlervariable von s_1: $\Delta\text{error}(s_1) = \|w_{s_1} - \xi\|^2$.

Bewege s_1 und dessen direkte Nachbarn n (bezüglich der Topologie) in Richtung ξ in Abhängigkeit der Distanz zu ξ: $\Delta w_{s_1} = E_b(\xi - w_{s_1})$ und $\Delta w_n = E_n(\xi - w_n)$.

Sind s_1 und s_2 über eine Verbindung k miteinander verbunden?

ja:

- $a(k) = 0$

nein:

- Erzeuge eine Verbindung zwischen s_1 und s_2

Lösche alle Verbindungen, deren Alter größer als a_{max} ist. Sind danach Neuronen ohne ausgehende Verbindungen vorhanden, lösche diese Neuronen ebenfalls.

Entspricht die Anzahl der generierten Eingangssignale ξ einem ganzzahligen Vielfachen eines Parameters λ?

ja:

- Füge ein neues Neuron ein:
  - Suche das Neuron q mit dem maximalen Fehlerwert error(q)
  - Füge ein neues Neuron r genau zwischen dem Neuron q und dessen Nachbarn f mit dem größten Fehlerwert ein: $w_r = 0.5(w_q + w_f)$.
  - Füge eine neue Verbindung zwischen dem neuen Neuron r und den Neuronen q und f ein. Lösche die ursprüngliche Verbindung von q und f.
  - Verringere die Fehlerwerte von q und f um den Faktor α. Initialisiere den Fehlerwert von r mit dem neuberechneten Fehlerwert von q.

nein:

Verringere alle Fehlerwerte um den Faktor β

Erfüllung eines Stoppkriteriums

Abbildung 7.6: Originaler Algorithmus zur Generierung von *Neural Gas*-Netzen nach [FRITZKE 1995b]

- Um Neuronenballungen zu verhindern, sollte der *interest value* i einer bestimmten Bildposition bzw. eines Neurons in Abhängigkeit vom Abstand zu Nachbarneuronen verringert werden (laterale Hemmung).
- Die Neuronen n müssen typbezogen sein, d.h. sie müssen Informationen über die in ihrem rezeptiven Feld f_{rez} vorhandenen Merkmale liefern (Merkmalsattribut $m(n)$).
- Bei der Bewegung der Neuronen und beim Einfügen neuer Neuronen sollen gewisse Priorisierungen hinsichtlich der lokalen Merkmale möglich sein.
- Das Einfügen von neuen Neuronen sollte nicht nur von der Iterationsanzahl abhängig sein. Bedeutender ist ein globales Fehlermaß err_{gl}, welches sich durch Summation aus den einzelnen lokalen Fehlerwerten ergibt, d.h. das Fehlermaß wird abgeleitet von den Merkmalswerten $i(p)$, wenn p nicht durch ein rezeptives Feld eines Neurons abgedeckt ist.
- Im Laufe der Bewegungen der Neuronen können diese rezeptive Felder besitzen, die keine Merkmale enthalten ($i(n) = 0$). Nach Erfüllung des Abbruchkriteriums sollten diese entfernt werden, da sie keine Informationen über die Objektstruktur liefern. Dabei sind die Verbindungen zu benachbarten Neuronen besonders zu berücksichtigen.
- Neue Neuronen wurden bei Fritzke immer zwischen zwei bestehenden eingefügt. Diese Festlegung ist wenig sinnvoll bei nah beieinander liegenden Neuronen, da dann an dieser Stelle eine Neuronenballung auftreten würde.
- Das Einfügen eines neuen Neurons sollte gezielt an einer Stelle p mit $i(p) > 0$ geschehen, die noch nicht durch rezeptive Felder abgedeckt ist. Neue Neurone werden jetzt jeweils sofort bei Übersteigen einer globalen Fehlerschranke $err_{gl_{max}}$ eingefügt. Damit werden zum einen Neuronenballungen vermieden, zum anderen ist sofort eine ideale Position gefunden. In diesem Zusammenhang sind die Parameter α und β nicht mehr notwendig.

Die sich ergebenden Endpositionen der Neurone werden als Knoten und die neuronalen Verbindungen als Kanten der Modellgraphen interpretiert. Die neuronalen Verbindungen legen damit die räumlichen Beziehungen zwischen den Neuronen fest. Neurone n werden an Regionen mit einem hohen *interest value* $i(n)$ $\in$ [0,1] plaziert. i spiegelt die Existenz von Merkmalen innerhalb des rezeptiven Feldes f_{rez} eines Neurons wider. Eingesetzt werden hier die im Kapitel 7.1.1 vorgestellten Merkmale (Ecken, Kreise und Grauwertkanten), wobei natürlich prinzipiell auch andere Merkmale mit hoher Lokalität eingesetzt werden können. Somit bedeckt jedes Neuron eine Region der Merkmalskarten gemäß ihrer rezeptiven Felder f_{rez}. Entsprechend werden folgende neue Parameter eingeführt:

- f_{rez}: die Größe des rezeptiven Feldes.

- $err_{gl_{max}}$: der maximal tolerierte globale Fehler bevor ein neues Neuron eingefügt wird. Der globale Fehler err_{gl} wird von Merkmalen abgeleitet, die noch nicht durch die rezeptiven Felder von Neuronen abgedeckt sind. Somit wird err_{gl} durch den *interest value* $i(r)$ einer Region r erhöht, wenn das Zentrum von r nicht innerhalb eines beliebigen rezeptiven Feldes liegt.

Damit ergibt sich der in Abbildung 7.7 dargestellte abgewandelte Algorithmus zur Generierung eines *Neural Gas*-Netzes.

Zur Festlegung eines Abbruchkriteriums wurden eine Reihe von Verfahren getestet. Das Kriterium soll gewährleisten, daß der Algorithmus möglichst umfassend die gegebenen Merkmalspunkte zur Neuronengenerierung ausnutzt. Folgende Abbruchmöglichkeiten wurden evaluiert:

- Untersuchung des zeitlichen Verhaltens des Fehlerwertes err_{gl}.

- Untersuchung des zeitlichen Verhaltens der Bewegung aller Neuronen.

- Mittlere Entfernung $\overline{d}(p_{s_1}, p)$ des ausgewählten Merkmalspunktes p zum nächstgelegenen Neuron s_1.

- Das gesamte Eingangsbild wird nach Merkmalspunkten abgesucht und die prozentuale Abdeckung durch die rezeptiven Felder der generierten Neuronen bestimmt.

Alle angeführten Verfahren lieferten gute Ergebnisse. Allerdings wurde für die ersten drei eine Neigung zum verfrühten Abbruch festgestellt. Offensichtlich besitzen die zeitlichen Aufzeichnungen lokale Minima. Die mittlere Entfernung $\overline{d}$ ist qualitativ vergleichbar mit der Analyse des Fehlerwertes err_{gl} und neigt damit ebenso zu lokalen Minima. Letztgenanntes Verfahren erwies sich als äußerst stabil, wobei Befürchtungen hinsichtlich einer großen Laufzeit durch das Absuchen der Merkmalsbilder sich nicht bewahrheiteten. Im Vergleich zur Gesamtlaufzeit der Generierung des *Neural Gas*-Netzes ist der Aufwand für das Abbruchkriterium vernachlässigbar. Der prozentuale Schwellwert zur Neuronenabdeckung wurde empirisch auf 10% festgelegt.

Der resultierende Algorithmus hat zum Teil Ähnlichkeiten zu der Repräsentationsstrategie von [SHAMS 1995] (Beschreibung siehe Kapitel 3.3.4). In folgenden wesentlichen Punkten wird jedoch hier anders verfahren:

- Hier wird die relative Lage der Neuronen zueinander nicht explizit gespeichert. In [SHAMS 1995] wurde die Position der Nachbarneurone durch

Netzgenerierung — Modifizierter Algorithmus

Beginne mit zwei Neuronen an zufälligen Positionen und generiere eine Verbindung k zwischen ihnen mit $a(k) = 0$. Setze $err_{gl} = 0$.

Suche nach einer Region r mit der Zentrumsposition p und $i(r) \neq 0$

Suche nach den zwei euklidisch nächsten Neuronen s_1 an p_{s_1} und s_2 an p_{s_2} zu p

ja: $p \notin f_{rez}(s_1)$ — nein

Inkrementiere err_{gl} um $i(r)$

Inkrementiere das Alter $a(l)$ von allen Verbindungen l, die von s_1 ausgehen, um 1

ja: Sind s_1 und s_2 durch eine Verbindung m verbunden — nein

Setze $a(m) = 0$

Generiere eine Verbindung zwischen s_1 und s_2

ja: $i(r) > i(f_{rez}(s_1))$ — nein

ja: $i(f_{rez}(s_1)) = 0$ — nein

Verschiebe s_1 mit dem Faktor E_n nach p

Verschiebe s_1 mit dem Fakt. E_b nach p

Verschiebe s_1 mit dem Fakt. E_n nach p

ja: $err_{gl} > err_{gl_{max}}$ und $p \notin f_{rez}(s_1)$ — nein

Verschiebe verbundene Nachbarn von s_1 mit dem Faktor E_n nach p

Füge ein neues Neuron an p ein und verbinde es mit s_1

Setze $err_{gl} = 0$

Neuberechnung von $i(f_{rez}(t))$ für alle Neurone t

Lösche alle Verbindungen n mit $a(n) > a_{max}$ und lösche entstehende unverbundene Neurone

Erfüllung eines Stoppkriteriums

Lösche alle Neurone u mit $i(f_{rez}(u)) = 0$ und verbinde die Nachbarn von u neu

Berechne die invarianten Momente $c_1^{undef.} = \eta_{20} - \eta_{02}$ gemäß [HU 1962] für jedes rezeptive Feld aller Neurone ($s(x, y)$ sind die Grauwerte des Modellbildes; x_{cp}, y_{cp} sind die Schwerpunkte):

$$\eta_{\mathrm{pq}} = \frac{\mu_{\mathrm{pq}}}{\mu_{00}^{\frac{p+q}{2}+1}} \qquad \text{mit} \qquad \mu_{\mathrm{pq}} = \sum_x \sum_y (x - x_{\mathrm{cp}})^p (y - y_{\mathrm{cp}})^q s(x, y)$$

Abbildung 7.7: Modifizierter Algorithmus zur Generierung von *Neural Gas*-Netzen

einen Vektor δ_{ij} explizit angegeben. Die Notation erscheint allerdings unnötig, da jedes Neuron seine eigenen Positionsmerkmale besitzt und somit der Verschiebungsvektor einfach bestimmt werden kann. Wesentlich ist die qualitative Lage der Neuronen zueinander, deren Ausprägungen direkt durch das Neuronennetzwerk gegeben ist.

- Jedes Neuron kann hier mehrere im rezeptiven Feld vorhandenen Merkmale kodieren. Der Organisationsaufwand, jedem Merkmal ein eigenes Neuron zuzuordnen, erscheint ebenfalls unnötig.

- In [SHAMS 1995] wird zur stabilen Ortskodierung der Neurone deren rezeptives Feld verändert. Hier werden die *interest values* i dazu benutzt. Die Analyse von i erlaubt weitreichendere Möglichkeiten der Steuerung des dynamischen Verhaltens des Netzes (siehe Algorithmus).

7.2 Matchstrategien

In der Matchphase müssen neue, zunächst unbekannte Objekte bezüglich den Einträgen in einer Modelldatenbank verglichen werden. Folgende Fragestellungen sind relevant:

1. Welche Eingabedaten sollen dem Algorithmus zur Verfügung stehen? Einerseits kann ebenso wie für das Modell ein generiertes *Neural Gas*-Netz verwendet werden. Allerdings stellt dieses Netz bereits eine verarbeitete Variante der für die zu untersuchende Szene zur Verfügung stehenden Informationen dar. Somit ist zu prüfen, ob die direkte Bearbeitung der Merkmalsbilder stabilere Ergebnisse bringt.

2. Ist eine Unterteilung in eine Hypothesen- und Validierungsphase nötig? Bei der Untersuchung komplexer Objekte mit sehr vielen Merkmalspunkten bzw. sehr großen *Neural Gas*-Netzen könnten Bedingungen auftreten, die zu einer kombinatorischen Explosion führen. Ist diese Gefahr gegeben, sollten einfache Algorithmen niedriger Ordnung eine Grobklassifizierung durchführen, die anschließend validiert werden muß.

Eine Hypothesenbildung zur Aufwandsreduktion innerhalb von graphenbasierten Matchverfahren ist in der Literatur vielfach zu finden (z.B. [GMÜR und BUNKE 1989], [BELLAIRE 1995], [KAWAGUSHI und SETOGUCHI 1994], [RAO und BALLARD 1995], [SCHWARZINGER et al. 1995], [GÖTZE et al. 1996]). Bei dem Entwurf eines Matchverfahrens muß berücksichtigt werden, daß die Generierung des

Neural Gas-Netzes zufallsbehaftet ist, d.h. die konkrete Position und Anzahl von Neuronen kann auch bei der wiederholten Präsentation der gleichen Szene unterschiedlich sein; ein konkretes Netz ist nicht 100%ig reproduzierbar. Somit sind fehlertolerante Systeme zu entwerfen, falls ein generiertes *Neural-Gas*-Netz der Szene als Matchgrundlage dienen soll. Evtl. müssen dann die Netze bezüglich ihrer Topologie verändert werden (wie z.B. in [ANDO 1996]), um aussagekräftige Ergebnisse zu erzielen.

Das im Kapitel 6 verwendete *Dynamic Link Matching* wurde zunächst auch in die Untersuchungen mit einbezogen. Das Verfahren lieferte beim Matchen von Merkmalsbildern gegeneinander gute Resultate. Die etwas andere Aufgabenstellung hier (zumindest die Objektmodelle sind als *Neural Gas*-Netz gegeben) würde aber eine intensive Überarbeitung nötig machen. Prinzipiell wäre denkbar, die Bewegungen des sog. *Blobs* nur entlang von Kanten im Modellnetz zuzulassen. Allerdings ergibt sich dann ein Algorithmus, der einem Graphensuchverfahren sehr ähnlich wird (siehe folgendes Kapitel 7.2.3).

Ein Matchverfahren unter Anwendung der im Kapitel 3.3.1 angedeuteten Verfahren zur Formenanalyse scheint zunächst vielversprechend, da sich die *Neural Gas*-Netze in geometrische Grundformen, wie z.B. Dreiecke, zerlegen lassen. Eine Bestimmung der Shape-Distanz könnte dann Aussagen über die Ähnlichkeit zweier Netze liefern. Problematisch ist aber, daß die ursprünglich zur Analyse von Landmarken entwickelten Verfahren sowohl von der gleichen Anzahl von Marken für Modell und Szene ausgehen, als auch von einer gleicher Numerierung der Marken. Diese Voraussetzungen lassen sich aber bei den sehr variabel auftretenden *Neural Gas*-Netzen nicht erfüllen. Ein weiteres Problem liegt darin, daß die Verbindungen innerhalb der Netze nicht berücksichtigt werden würden. Aber gerade diese Informationen sind zur Beschreibung der Objekte wesentlich.

Wesentlich ist die Beachtung des Umfeldes für das zu entwerfende Verfahren: eine aktive Sehumgebung. Dadurch können bestimmte Problemstellungen vereinfacht behandelt werden. Die bereits im Zuge der Merkmalsgenerierung angesprochene Aufmerksamkeitssteuerung liefert die Gewähr, daß sich im zu bearbeitenden Szenenausschnitt auch wirklich zu untersuchende Objekte befinden. Somit kann auf eine Einbeziehung des Grenzfalles, daß das System auch ohne präsentierte Objekte zuverlässig konvergieren soll, zunächst verzichtet werden. Weiterhin sind die Möglichkeiten der Stereobildverarbeitung zu berücksichtigen. Durch die binokulare Fixation auf ein Objekt läßt sich die Tiefenstruktur der Szene ermitteln und dadurch eine Objekt-Hintergrund-Trennung durchführen (für weitere Einzelheiten siehe Kapitel 4.4). Gerade dadurch vereinfacht sich das Erkennungssystem beträchtlich, da als Voraussetzung angenommen werden kann, daß die generierten Bilddaten weitgehend vom störenden Hintergrund befreit sind.

Im folgenden werden die implementierten und untersuchten Verfahren vorgestellt. Alle Algorithmen sind durch bereits erfolgreich realisierte Erkennungsver-

fahren (siehe Kapitel 3) motiviert.

7.2.1 Hash-Verfahren

Durch die Verwendung von Hash-Tabellen soll die Übereinstimmung zweier *Neural Gas*-Netze bestimmt werden. Für eine gegebene Szene muß ein solches Netz zunächst generiert werden. Ein direktes Arbeiten mit den Merkmalsbildern ist nicht möglich, da die Berechnung der Hash-Werte für Modell und Objekt auf der gleichen Netz-Basis erfolgen muß. Bei Verwendung von Hash-Verfahren für Erkennungsaufgaben ist eine Unterteilung in eine *offline*- und *online*-Phase üblich (siehe auch Kapitel 3.3.3). Während der *offline*-Phase werden die Tabellen vorbereitet, aus denen im *online*-Betrieb die Matchergebnisse ausgelesen werden. Die zu generierende Hashtabelle H soll unter Verwendung einer spezifischen Hashfunktion f_h innerhalb der *offline*-Phase aufgebaut werden:

1. Generierung der *Neural Gas*-Netze für die Modelle.

2. Aufbau einer Hashtabelle pro Modell. Dazu werden die innerhalb der rezeptiven Felder der Neuronen vorhandenen Merkmale m, d.h. Kanten, Ecken und Kreisstrukturen als gesetzte bzw. gelöschte Bits im Merkmalsvektor kodiert[1]. Benutzt werden jeweils Neuronenpaare a, b, die direkt verbunden sind. Der Hashindex h_i wird aus den bitkodierten Merkmalswerten der beiden Neuronen abgeleitet: $h_i = f_h(m(a), m(b))$. Weiterhin werden für das Neuronenpaar noch folgende zusätzliche Merkmale berechnet:

   - normierter Winkel zwischen den Neuronen und
   - normierte relative (zur Bildgröße) Länge der Verbindung zwischen den Neuronen.

Die Hashfunktion f_h wurde als Kompromiß zwischen einem erhöhten Speicheraufwand bei feinerer Auflösung (breit streuende Funktion) und geringem Speicherbedarf bei evtl. nötigem sequentiellen Durchsuchen der Hasheinträge (gering streuende Funktion) festgelegt. Eingesetzt wurde die Summe der Merkmalswerte der Neuronen ($f_h = m(a) + m(b)$).

Die *online*-Phase gliedert sich in folgende Teilschritte:

1. Generierung des *Neural Gas*-Netzes für die zu untersuchende Szene.

[1] Hier fand keine Priorisierung statt, da nur das Vorhandensein der Merkmale kodiert werden soll.

2. Zerlegung des Netzes in Subnetzwerke mit 2 Neuronen, die direkt verbunden sind und Berechnung des Hashindexes h_i mit der im *online*-Teil verwendeten Hashfunktion f_h.

3. Sequentielle Suche in der Hashtabelle unter $H(h_i)$ nach übereinstimmenden Merkmalswerten. Dazu wird zunächst ein Neuronenpaar mit passenden Merkmalen gesucht. Anschließend wird überprüft, ob mindestens je ein benachbartes Neuron mindestens die Merkmale von benachbarten Modellneuronen besitzt (siehe Abbildung 7.8 zur Erläuterung). Der Matchzähler für das so gefundene Modell wird inkrementiert. Die Untersuchung nach übereinstimmenden Merkmalen und entsprechenden Nachbarn wird für alle Einträge unter $H(h_i)$ durchgeführt, um ein *one-to-many-mapping* zu ermöglichen.

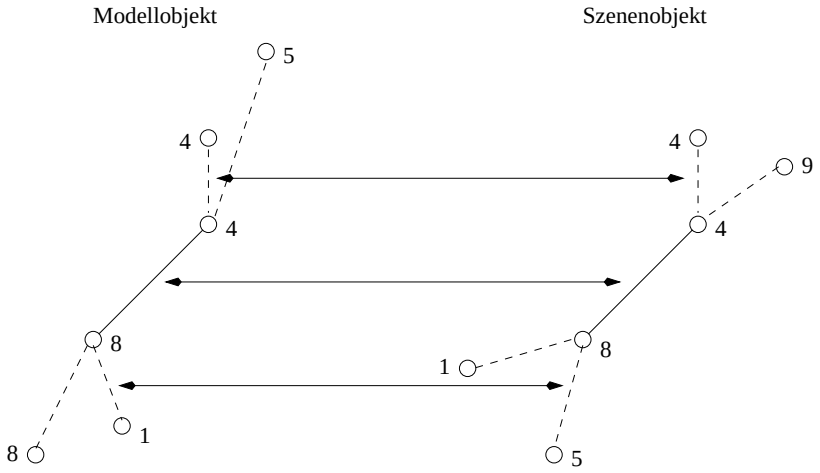


Abbildung 7.8: Die betrachteten Neuronenpaare vom Modell- und Szenenobjekt werden nach übereinstimmenden Merkmalen getestet (hier: 8, 4). Jeweils angrenzende Nachbarneurone werden nach prinzipieller Merkmalsübereinstimmung getestet, d.h. es können in der Szene auch mehr Merkmale detektiert werden, solange das Modellmerkmal auch vorhanden ist (hier: Match von 1 auf 5 und 4 auf 4).

4. Bestimmung des Winkels zwischen dem Neuronenpaar und die Länge der Verbindung. Die Differenzen zu den Werten für das matchende Modell werden gespeichert. Diese berechneten zusätzlichen Merkmale pro Neuronenpaar haben zunächst auf den Aufbau der Hashtabelle keinen Einfluß. Anhand dieser Merkmale sollen die Möglichkeiten einer Schätzung einer Objekttransformation (Rotation und Skalierung), angeregt durch [COSTA und SHAPIRO 1996], untersucht werden.

5. Die aus dem *one-to-many-mapping* resultierenden eventuell mehrdeutigen Matchergebnisse werden mit dem Ziel, deren Anzahl zu verringern, näher analysiert. Folgende Kriterien werden verwendet:

   - Die Anzahl der auf einen Hashtabellen-Eintrag referenzierenden Objekt-Modell-Matches (*many-to-one mapping*) wird herangezogen, um prinzipielle Mehrdeutigkeiten zu finden.
   - Eine Analyse der Entfernungsverteilung der *Neural Gas*-Netz-Verbindungen für das Szenenobjekt liefert ein topologisches Entfernungsmaß[2] als Kriterium zur Überprüfung der räumlichen Distanz von matchenden Elementen. Matcht ein Objektelement auf mehrere weit voneinander entfernte Modellelemente, so konnte dieses Modellelement nicht eindeutig zugeordnet werden und die entsprechenden Matcheinträge können gelöscht werden.

6. Für die resultierenden Matcheinträge werden die berechneten Merkmalsdifferenzen (Winkel und Länge) in Histogramme eingetragen. Die häufigsten Winkel- bzw. Längendifferenzen werden zu einer evtl. Objekttransformationsschätzung verwendet.

Das dargestellte Verfahren wurde implementiert und getestet. In Kapitel 7.3.2 sind Aussagen über die erzielten Ergebnisse zu finden. Dort werden auch die Resultate der Transformationsschätzung präsentiert.

7.2.2 Relaxations-Verfahren

Die Entwicklung eines Matchalgorithmus aus der Klasse der Relaxations-Verfahren war zum einen durch die bekannten *simulated annealing*-Optimierungsverfahren und zum anderen durch die innerhalb ihrer Anwendungsdomäne recht erfolgreichen Hopfield-Ansätze motiviert (siehe Kapitel 3.4). Besonderheit des Ansatzes ist das ausschließliche Arbeiten auf einer Matchmatrix, die die zuvor generierten *Neural Gas*-Netze zueinander in Beziehung setzt. Die Zeilen der Matrix entsprechen der Anzahl der Modellneuronen, die Spalten entsprechend der Anzahl der Szenenneuronen. Jeder Eintrag wird durch ein Matchkriterium für ein Modellneuron n_{Model} und ein Szenenneuron n_{Szene} bestimmt, wobei die Anzahl unterschiedlicher Bitstellen der bitkodierten Merkmale untersucht wird:

[2] Z.Z. wird die zweitminimalste Distanz unter allen Netzverbindungen verwendet.

$$match \quad = \quad NEG\ (m(n_{Model})\ XOR\ m(n_{Szene})) \tag{7.1}$$

m	bitkodierter Merkmalsvektor
XOR	bitweise exklusive ODER-Verknüpfung
NEG	bitweise Invertierung
$match$	Matchwert in $[0, 1]$

Grundidee von Relaxations-Matchverfahren ist eine minimale stochastische Veränderung der Matchmatrix. Wesentlich für den Entwurf eines solchen Verfahrens ist daher die Festlegung eines Fehlermaßes, welches in Abhängigkeit von der herrschenden Systemtemperatur T_{System} zur Entscheidung über die Annahme oder Ablehnung einer Veränderung herangezogen wird. Folgende normierte Größen gehen gleichberechtigt in das globale Fehlermaß $error_{gl}$ ein:

- die Anzahl nicht exakter Matches zwischen Modell- und Szenenneuronen, d.h. Matchbeziehungen mit $match < match_{min}$ bzw. $match > (1 - match_{min})$[3]
- die Anzahl nicht übereinstimmender Merkmalswerte m bei Matchbeziehungen mit $match > match_{min}$
- die Anzahl > 1 matchender Elemente pro Modellneuron
- werden mehrere Szenenneuronen einem Modellneuron zugeordnet, so sollte der Abstand der Szenenneurone untereinander klein sein, da dann die Wahrscheinlichkeit hoch ist, daß sich mehrere Szenenneurone an einer Merkmalsposition gebildet haben
- zwei matchende Modellneurone a bzw. Szenenneurone b sollten einen ähnlichen Abstand zum Schwerpunkt des *Neural Gas*-Netzes COG_N haben $d(a, COG_N) \approx d(b, COG_N)$; damit können Fehlzuordnungen quantitativ erfaßt werden

Folgender Algorithmus wurde implementiert:

1. Generierung des *Neural Gas*-Netzes für die Szene.
2. Berechnung einer initialen Matchmatrix zwischen Modell- und Szenennetz.

[3] $match_{min}$ ist ein globaler Matchschwellwert

3. Stochastische minimale Veränderung der Matchmatrix, d.h. daß eine neue Matchbeziehung zwischen Modellnetz und Szenennetz generiert werden kann bzw. eine bereits vorhandene kann gelöscht werden.

4. Berechnung eines neuen globalen Fehlerwertes gemäß den oben angegebenen Kriterien.

5. Ablehnung eines sich ergebenden zu großen Fehlerwertes in Abhängigkeit von der globalen Temperatur. Die Ablehnungswahrscheinlichkeit ergibt sich aus der Änderung des globalen Fehlerwertes $error_{gl}$ und der Systemtemperatur T_{System}:

$$P_{Ablehnung} = 1.0 - e^{-\frac{(error_{gl(neu)} - error_{gl(alt)}) \cdot F}{T_{System}}} \tag{7.2}$$

   Auf Grund der nur sehr geringen Fehlerwertänderungen ist ein Skalierungsfaktor F (hier gesetzt auf 10) notwendig.

6. Absenkung der Systemtemperatur und weiter bei 3, wenn $T_{System} > 0$.

Das entstandene Verfahren entspricht einem typischen *simulated annealing*-Optimierungsansatz unter Verwendung problemspezifischer Kriterien zur Bestimmung des initialen Matches und eines globalen Fehlermaßes. Wie auch in [AZENCOTT und YOUNES 1997] angedeutet, bieten Optimierungsansätze zur Objekterkennung großen Raum für Modifikationen und Experimente. Hierbei ist besonders die Festlegung des Fehlermaßes kritisch. Auch geringe Modifikationen wirken sich stark aus. Die Frage nach der Art und dem Umfang der Modifikationen ist ebenfalls schwer zu handhaben. Aussagen über die erzielbaren Ergebnisse finden sich im Kapitel 7.3.2.

7.2.3 Graphensuch-Verfahren

Schon in [ULLMANN 1976] wird ein effektiver Subgraph-Isomorphie-Algorithmus vorgestellt. Ullmanns Ziel ist die Bereitstellung eines möglichst allgemeinen Verfahrens für eine Vielzahl von Anwendungen. Der Einsatz innerhalb eines zu entwickelnden Objekterkennungssystems auf der Basis von Objektbeschreibungen durch *Neural Gas*-Netze erscheint allerdings ungünstig, da die vorhandenen Informationen über die räumlichen und strukturellen Beziehungen der Neuronen nicht berücksichtigt werden. Hier soll ein fehlertolerantes Graphmatchverfahren entwickelt werden, daß allerdings nicht auf der Basis von *edit*-Kosten arbeitet. Vielmehr sollen strukturelle Unterschiede zwischen

Modell- und Szenenobjekt durch die Verwendung unscharfer Matchbedingungen toleriert werden. Somit entfällt die problematische Suche nach der günstigsten *edit*-Operation.

Der Algorithmus ist zweistufig aufgebaut. Zunächst wird eine Menge *primärer* Matches gebildet. Dazu werden einfache auf die Merkmalswerte der Neurone bezogene Matchbedingungen verwendet. Anschließend werden durch ein Suchverfahren die angrenzenden Neurone auf strukturelle Gleichheit getestet. Im einzelnen werden folgende Schritte ausgeführt:

1. Generierung des *Neural Gas*-Netzes für die Szene.

2. Bestimmung der Menge der pro Modellneuron matchenden[4] Szenenneurone (*primärer Match*).

3. Suche nach nicht verbundenen Subgraphen innerhalb des *Neural Gas*-Netzes des Objektmodells. Jede Neuronenverbindung erhält ein Attribut über seine Zuordnung zu einem Subgraphen. Das resultierende Matchmaß wird pro Subgraph bestimmt. Dadurch kann dieses Maß besser interpretiert werden. Eine Aussage über das Vorhandensein mehrerer Objekte in der Szene ist möglich.

4. In einem rekursiven Tiefe-Zuerst-Suchprozeß werden pro Modellneuron und mit diesem matchende[5] Szenenneurone geprüft, ob sie strukturell, d.h. bezüglich ihrer Merkmalswerte und ihrer räumlichen Lage zueinander, ähnlich sind (siehe Abbildung 7.9 zur Verdeutlichung). Bestimmt wird die Anzahl der gefundenen matchenden Szenennetzverbindungen relativ zur Gesamtanzahl der Modellnetzverbindungen. Für die räumliche Lage der Neurone zueinander wird ein Suchrahmen verwendet, der Ausdruck für die maximal zugelassenen Abweichungen ist (siehe Abbildung 7.10).

Die Verwendung eines Suchrahmens stellt eine effektive Variante zur Begrenzung des Suchraums des Matchproblems dar. Allerdings wird dadurch auch eine globale Rotationsinvarianz unmöglich. Wie auch in der Literatur nicht unüblich (z.B. [LIN et al. 1991], [BENNAMOUN und BOASHASH 1997]) und durch die Anwendung eines ansichtenbasierten Erkennungssystems unterstützt, müssen dann die Modelle in mehreren Rotationsansichten vorliegen. Im Kapitel 7.3.2 werden die Ergebnisse dokumentiert.

[4] d.h. die Merkmalswerte beider Neuronen werden auf Gleichheit getestet

[5] bestimmt durch den *primären* Match

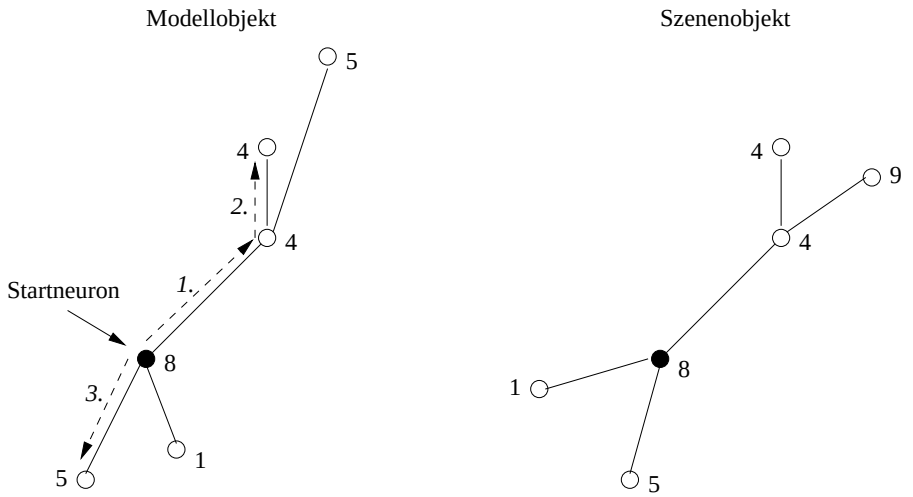


Abbildung 7.9: Tiefe-Zuerst-Suche zur Bestimmung der maximalen Subgraphisomorphie. Ausgehend vom Modell-Startneuron werden benachbarte Modellneuronen hinsichtlich ihrer matchenden Szenenneurone getestet (gestrichelte Linien). Es müssen die Merkmalswerte und die räumlichen Beziehungen übereinstimmen.

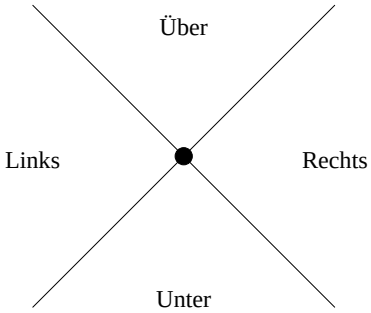


Abbildung 7.10: Bezüglich des gerade untersuchten Neurons (gekennzeichnet durch einen schwarzen Punkt) verwendeter Suchrahmen, d.h. lag im Modellnetz ein Nachbarneuron im Bereich *Über*, werden im Szenennetz alle Neuronen innerhalb dieses Bereichs untersucht.

7.2.4 Deformations-Verfahren

Die bisher beschriebenen Matchansätze versuchten, zwei zuvor generierte *Neural Gas*-Netze miteinander in Beziehung zu setzen. Vorteil dieser Vorgehensweise ist die dadurch reduzierte zu bearbeitende Informationsmenge für einen effektiven Matchalgorithmus. Als Nachteil ergibt sich dagegen, daß die Informationsmenge zur Diskriminierung ähnlicher Objekte nicht mehr ausreichend sein kann. Inner-

halb des hier dargestellten Deformationsverfahrens wird deshalb den in der Szene generierten Merkmalen größere Bedeutung beigemessen. Der Algorithmus arbeitet nur noch mit einem *Neural Gas*-Netz - dem des Objektmodells. Das Szenenobjekt wird nur durch Merkmalspunkte repräsentiert. Grundidee ist nun, das Modellnetz an die vorhandenen Merkmalsausprägungen anzupassen und zu deformieren (siehe z.B. auch [SCHWARZINGER et al. 1995], [ANDO 1996], [NOLL 1996], [DICKINSON und METAXAS 1997]). Der Umfang der nötigen Deformation des Modellnetzes ist Ausdruck für das Maß der Übereinstimmung von Modell- und Szenenobjekt. Ähnlich wie bei dem vorgestellten Relaxationsverfahren bestimmen hier die Wahl der auszuführenden Deformationen des Modell-Netzes, die Beurteilung der durchgeführten Deformationen und der Entwurf eines Abbruchkriteriums die wesentlichen Entwurfsaspekte. In [DICKINSON und METAXAS 1997] werden eine Reihe von Limitierungen für Deformationsverfahren im allgemeinen und für das vorgestellte Erkennungssystem (siehe Kapitel 3.3.1) im speziellen herausgestellt. Im folgenden wird zu den einzelnen Kritikpunkten Stellung genommen:

- Voraussetzung ist eine gute Segmentierung
  Das Problem hierbei ist die störende Hintergrundinformation, die je nach deren Beschaffenheit den Deformationsprozeß fehlleitet. Auch im hier vorgestellten Ansatz wird eine gute Segmentierung bzw. fehlender Hintergrund vorausgesetzt. Hier kann das Objekterkennungssystem von der Einbindung in ein aktives Sehsystem (siehe z.B. Kapitel 4.4) profitieren.

- Probleme mit der initialen Positionen und Rotationen
  Bei Einführung beliebig vieler Freiheitsgrade während des Deformationsprozesses (Translation, Rotation, Skalierung) wird durch die Wahl der Anfangsposition des zu deformierenden Modells stark das Ergebnis beeinflußt. Hier wird jedoch von einer ansichtenbasierten Erkennung ausgegangen, d.h. ein bestimmtes *Neural Gas*-Modell muß bei passender Szene nur gering deformiert werden. Unterscheiden sich Modell und Szene stark, z.B. weil das Szenenobjekt rotiert wurde, würde ein benachbartes Modell geringere Deformationswerte aufweisen. Allerdings verlagert sich jetzt das Problem hin zur Wahl einer ausreichenden Menge an Modellen für die typischen Ansichten eines Objekts. Eine Rotationsinvarianz wird hier, wie auch bereits im vorigen Kapitel dargelegt, durch die Repräsentation rotierter Modelle realisiert.

- Annahme, daß äußere Konturen von Regionen Objektkonturen sind
  Diese Annahme spielt vor allem im Erkennungssystem von Dickinson et al. ein Rolle, da dort der Deformationsprozeß auf eine Anpassung von Modell und Szene an der äußeren Kontur aufbaut. Hier werden die Deformationen

durch die Beziehung von geometrischen Merkmalspunkten an beliebigen Objektpositionen gesteuert.

- Dekomposition von Objekten in Objektteile
  Auch diese Voraussetzung ist nur für das System von Dickinson et al. wesentlich. Hier ist keine Zerlegung von Objekten nötig. Für komplexere Objekte ist weiterhin auch äußerst fehleranfällig.

Die Deformationen des Modell-Netzes werden durch die gezielte Verschiebung von Neuronen erreicht, wobei die Bewegungsrichtung durch die Suche nach matchenden Merkmalsausprägungen determiniert ist. Jeweils das räumlich nächste Neuron mit matchenden Merkmalen wird verschoben, wenn es noch keine passenden Merkmalsposition gefunden hat. Die über neuronalen Verbindungen mit diesem Neuron verbundenen Nachbarn werden ebenfalls verschoben, allerdings in geringerem Umfang. Innerhalb von auf Deformationen basierenden Erkennungssystemen ist die Einführung von Mechanismen zur Erhaltung der Integrität wesentlich. In der Literatur werden den Modellen dazu oft physikalische Eigenschaften wie Masse, Steifheit oder Dämpfung zugeordnet [DICKINSON und METAXAS 1997]. Diese Integritätserhaltung wird hier dadurch gewährleistet, daß jeweils das gesamte Modell und nicht nur ein matchendes Neuron verschoben wird. Zur Beurteilung der im Verlaufe des Matchprozesses ausgeführten Deformationen wird eine Analyse der Winkelbeziehungen zwischen den Modell-Neuronen verwendet. Basis ist eine Winkelmatrix, die die Winkel α mit $0 \leq \alpha < 2\pi$ zwischen allen Neuronenverbindungen angibt, wodurch sich ein invariantes Matchmaß ergibt. Am Ende des Deformationsprozesses ergeben sich dann Winkeldifferenzen, die entsprechend einem matchenden Szenenobjekt unterschiedlich groß ausfallen und ein globales Deformationsmaß D ergeben. Zusammen mit einer Analyse der Differenzen von Grauwertmomenten für die rezeptiven Felder der Neurone wird daraus ein Matchmaß abgeleitet. Als Abbruchkriterium wird ähnlich wie bei der Generierung der *Neural Gas*-Netze überprüft, ob alle Neuronen des Modell-Netzes an matchenden Szenenpositionen lokalisiert sind. Der Deformationsprozeß wird gestoppt, wenn 85% der Modellneurone sich an matchenden Szenenregionen befinden. Dabei wird getestet, ob die Neuronen wenigstens einen Teil der vorhandenen Szenenmerkmale repräsentieren, d.h. die während der Netzgenerierung in den rezeptiven Feldern vorhandenen Merkmale konnten für die Szenenbilder in den evtl. verschobenen rezeptiven Feldern wiedergefunden werden. Weiterhin wird ein pauschaler Abbruch initiiert, wenn bei komplexen Szenen keine Tendenz zu einer vollständigen Neuronenabdeckung der Szene zu erkennen ist, d.h. das Netzwerk bedeckt nicht genügend Merkmalspositionen, welches zu oszillierenden Neuronen führt. Die resultierenden Neurone v ohne Matchpartner gehen durch eine Pauschale D_{straf} negativ in das Matchmaß ein. Folgende Parameter steuern den Matchprozeß:

- D_{straf}: der Deformationswert wird für nicht matchende Neurone v um den Wert D_{straf} erhöht, d.h. diese Neurone befinden sich nicht an matchenden Merkmalsregionen gemäß den repräsentierten Modellmerkmalen.

- F_b, F_n: Verschiebungsfaktoren für die Neurone (analog zu E_b und E_n innerhalb der Netzgenerierung)

Der gesamte Deformationsalgorithmus ist in Abbildung 7.11 dargestellt. Er wird für jedes Modellnetz ausgeführt.

Matchen von *Neural Gas*-Netzen — Netzdeformation

Plaziere das undeformierte Modellnetz an eine zufällige Szenenposition p_{Start}
Berechne die Winkelmatrix A_{Start} für das Modellnetzwerk
Suche nach einer zufälligen Szenenregion r mit der Zentrumsposition p_{zuf} und $i(r) \neq 0$
Deformation I: Suche nach dem nächsten Neuron s_j zu p_{zuf} gemäß der euklidischen Distanz $d(s_j, p_{zuf})$ und matchenden Szenenmerkmalen ($i(f_{rez}(s_j)) \leq i(r)$). Verschiebe s_j nach p_{zuf}: $p_{s_j} = p_{s_j} + F \cdot d(s_j, p_{zuf})$. F ist abhängig vom Zustand des Neurons: ist es bereits an einer matchenden Szenenposition plaziert, so ist $F = F_n$, ansonsten $F = F_b$
Deformation II: Verschiebe alle anderen Neurone s_i $(i \neq j)$ mit dem Faktor F_n nach p_{zuf}
Erfüllung eines Stoppkriteriums
Berechne die Winkelmatrix A_{Ende} für das resultierende deformierte Netzwerk und berechne das Deformationsmaß D: $$D = \left(\sum_{\forall Zeilen, Spalten} \lvert A_{Ende} - A_{Start} \rvert \right) + v \cdot D_{straf}$$
Berechne die Momente $c_1^{def.}$ für die Grauwerte der rezeptiven Felder aller Neurone.

Abbildung 7.11: Algorithmus zur Deformation von *Neural Gas*-Netzen

Das Matchmaß wird zum einen vom Deformationswert D und zum anderen von den Differenzen der Grauwertmomente C der durch die rezeptiven Felder determinierten Grauwertbildregionen abgeleitet. Die Momente werden jeweils für das undeformierte Netzwerk auf den Modellbildern ($c_1^{undef.}$) und für die deformierten auf den Szenenbildern ($c_1^{def.}$) berechnet. Der Matchwert $match$ wird für eine Szenenobjektansicht in Beziehung zu allen Modellnetzen i berechnet:

$$match_i = \beta \left(\frac{D_{min} - D_i}{D_{max} - D_{min}} + 1 \right) + \gamma \left(\frac{C_{min} - C_i}{C_{max} - C_{min}} + 1 \right) \quad (7.3)$$

mit
$D_{min} = \arg(\min_{\forall i}(D_i))$,
$C_{min} = \arg(\min_{\forall i}(|c_{1_i}^{def.} - c_{1_i}^{undef.}|))$,
$D_{max} = \arg(\max_{\forall i}(D_i))$,
$C_{max} = \arg(\max_{\forall i}(|c_{1_i}^{def.} - c_{1_i}^{undef.}|))$,
$C_i = |c_{1_i}^{def.} - c_{1_i}^{undef.}|$,
β, γ = Gewichte mit $\beta + \gamma = 1$ und
D_i, C_i = Deformationswert bzw. Differenz der Grauwertmomente für das Modellnetz i.

Weiterhin werden die zwei besten Matchwerte hinsichtlich einer minimalen Distanz von 1% verglichen, um aussagekräftige Ergebnisse zu erhalten. Sollte die Distanz kleiner sein, so wird zwischen den beiden Modellen auf der Grundlage der geringeren Iterationszahl während des Deformationsprozesses entschieden. Die erzielten Ergebnisse sind im folgenden Kapitel angegeben.

7.3 Experimente und Ergebnisse

Die entworfenen Verfahren zur Objektrepräsentation und -erkennung werden im folgenden anhand einer konkreten Erkennungsaufgabe getestet. Ein Erkennungssystem kann aus Aufwandsgründen nur im Rahmen von wenigen Szenarien getestet werden. Umso entscheidender hinsichtlich der Aussagekraft der gewonnenen Ergebnisse ist die Wahl des Anwendungsfeldes. Hier wurde zunächst eine analoge Aufgabenstellung zum im Kapitel 6 beschriebenen Erkennungssystem unter Nutzung von zeitlichen Relationen gewählt. Konkret sollen die Ergebnisse des *Dynamic Link Matching* direkt verglichen werden. Man vergegenwärtige sich die äußerst schwere Erkennungsaufgabe: verschiedene Ansichten von Objekten (hier Spielzeugautos) sollen so genau diskriminiert werden können, daß die im Abstand von 15° aufgenommenen Ansichten richtig zugeordnet werden. Weiterhin sei erneut auf die äußerst primitiven Merkmale hingewiesen: Kanten, Ecken, Kreise. Zusammen mit der Kodierung deren räumlicher Relationen bilden sie die Grundlage des Erkennungssystems. Eine Analyse des zeitlichen Kontexts ist hier zunächst nicht vorgesehen, sollte aber prinzipiell Anwendung finden, um uneindeutige Ansichten sicher zuzuordnen.

Als Beispielobjekte kamen wieder die in Abbildung 6.8 dargestellten vier Spielzeugautos zum Einsatz. Jedes Modellobjekt wird einmal vollständig um die

eigene Hochachse in ca. 15° Schritten rotiert. Im Anhang sind alle verwendeten Ansichten der Objekte zu sehen. Alle Ansichten mit ungeradem Index werden als Szenenbilder und alle Ansichten mit geradem Index als Modellbilder betrachtet. Durch die geringe Schrittweite der Objektrotationen ergibt sich die schwierige Erkennungsaufgabe, da sowohl das korrekte Objekt als auch die korrekte Ansicht ermittelt werden soll.

7.3.1 Modellnetzgenerierung

Als Modelle fungieren jeweils die Ansichten mit ungeradem Index. Es ergeben sich 12 Modellansichten pro Objekt. Diese Ansichten werden zum Aufbau einer Datenbasis aus *Neural Gas*-Netzen verwendet. In Abbildung 7.12 sind die generierten Modellnetze für die Modellansichten des in Abbildung A.3 dargestellten Objekts zu sehen.

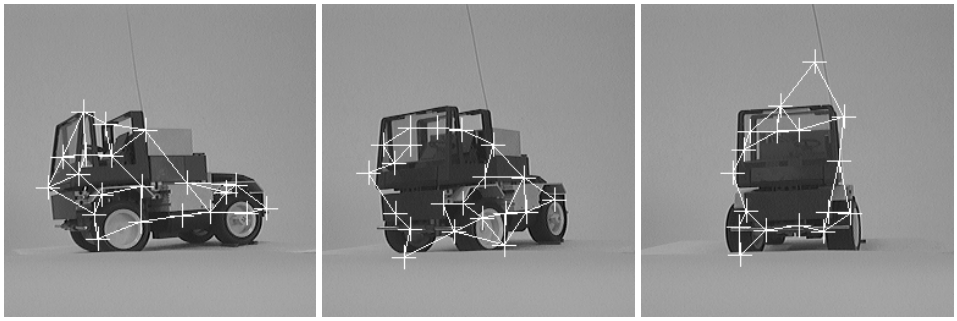


Abbildung 7.12: Die Modellansichten des Objekts 2 (Index 1,3,5) mit überlagerten *Neural Gas*-Netzen

Aus einer Reihe von Experimenten konnten Grundregeln für die notwendige Parametrisierung abgeleitet werden. Für die einzelnen Parameter ist folgendes festzustellen:

- f_{rez}: Durch die Größe des rezeptiven Feldes wird vor allem die gewünschte Dichte der resultierenden Neuronenanordnungen beeinflußt. Damit wird festgelegt, wie stark sich benachbarte Neuronen gegenseitig hemmen und wie schnell der globale Fehlerwert bis zum Einfügen eines neuen Neurons steigt.

- $err_{gl_{max}}$: Der maximale Fehlerwert bis zum Einfügen eines neuen Neurons steuert die relative Anzahl entstehender Neurone.

- a_{max}: Das maximale Alter von Neuronenverbindungen steuert die relative Dichte der generierten Verbindungen und auch inwieweit Verbindungen zwischen entfernteren Neuronen bestehen bleiben.

- E_b, E_n: Diese Faktoren beeinflussen die Dynamik der Neurone. Sie steuern die Geschwindigkeit der Neuronen auf ihrem Weg zu passenden Merkmalspositionen. Die Werte sollten nicht zu groß gewählt werden, da sonst die Neuronen stark oszillieren.

Im folgenden wird die Ausbildung eines *Neural Gas*-Netzes näher demonstriert. Dazu wird beispielhaft die Ansicht 1 des Objektes 2 verwendet. Dazu werden die in Tabelle 7.1 aufgeführten Parameterwerte verwendet. Die Herausbildung der *Neural Gas*-Repräsentation ist in Abbildung 7.13 dargestellt. Gekennzeichnet sind jeweils die Position der Neurone und ihre Verbindungsstruktur. Zu erkennen ist die schrittweise Ausbildung der Netzstruktur, wobei die Neurone an Positionen mit gut detektierbaren Merkmalen verharren. Vergegenwärtigt man sich die zugrundeliegenden Merkmalsrepräsentationen (siehe Abbildung 7.1 im Kapitel 7.1.1), wird durch das resultierende *Neural Gas*-Netz die Struktur des Objekts bezüglich dessen Merkmale und deren räumliche Relationen zueinander widergespiegelt.

Tabelle 7.1: Parameterwerte zur Modellnetzgenerierung

Parameter	Wert
f_{rez}	13
$error_{gl_{max}}$	50
a_{max}	10

Parameter	Wert
E_n	0.03
E_b	0.3

Die Netzgenerierung lieferte gute Ergebnisse, wobei hier die Bewertung der erfolgten Neuronenabdeckung durch subjektive “Betrachtung” erfolgte. Dementsprechend zeigten sich im Verlaufe umfangreicher Versuche eine Reihe von Mängeln:

- *Reproduzierbarkeit:* Da der *Neural Gas*-Algorithmus auf Zufallsprozessen aufbaut, kann nicht gewährleistet werden, daß verschiedene Netzgenerierungen für eine Modellansicht auch identische Ergebnisse liefern. Zwar werden i.a. die wesentlichen Objektstrukturen durch Neurone “abgedeckt”, aber deren Zahl ist nicht konstant über mehrere Testläufe.

- *Parameterabhängigkeit:* Das Ergebnis der Netzgenerierung ist im hohen Maße von der Parametrisierung abhängig. Zwar konnten Regeln für deren

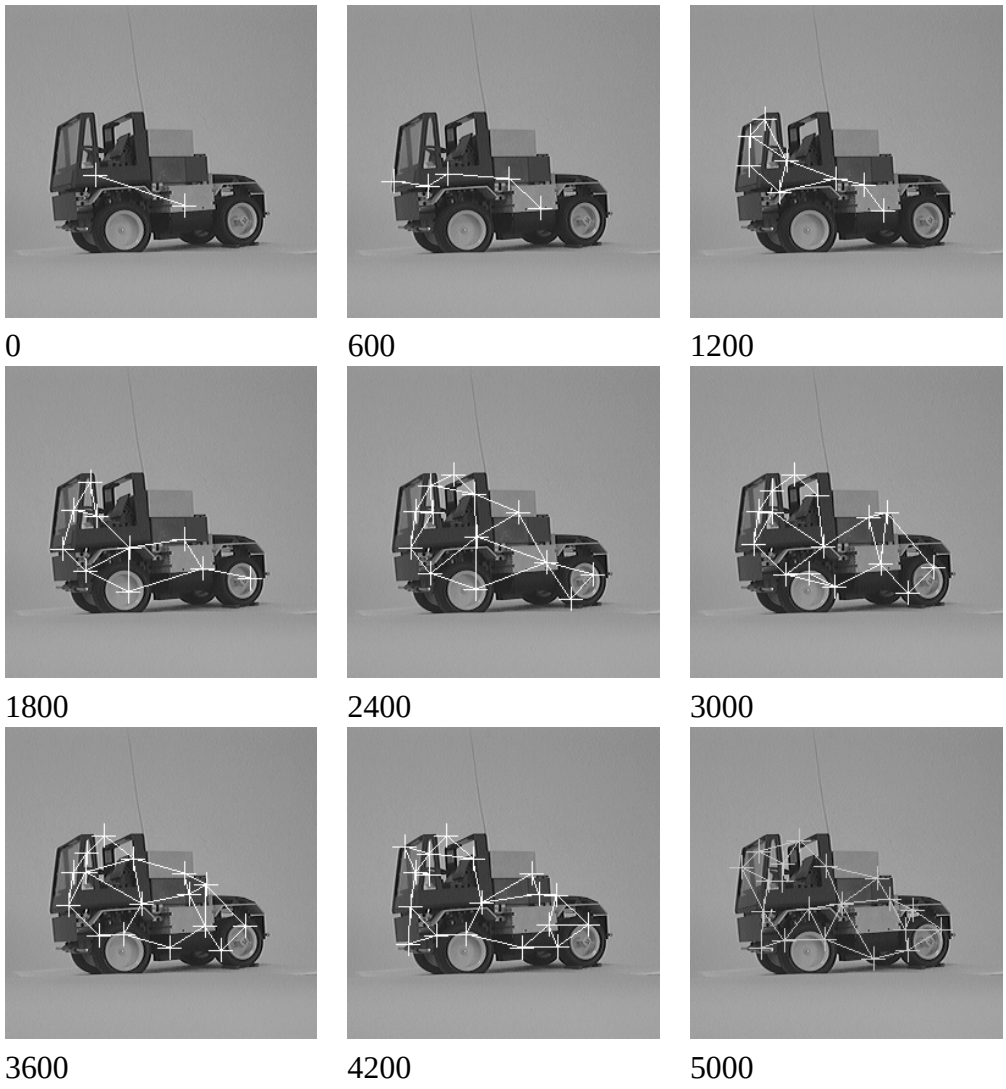


Abbildung 7.13: Ausbildung des *Neural Gas*-Netzes für das Bild Nr. 1 des Objekts 2 mit Angabe der Iterationsnummer

Festlegung aufgestellt werden, allerdings zeigt sich eine direkte Abhängigkeit von der Anwendungsdomäne. Die globale Anzahl generierter Neurone ist für eine konkrete Parameterfestlegung in etwa konstant, d.h. für Anwendungsfelder mit einfachen, unstrukturierten Objekten, die mit weniger Neuronen repräsentiert werden könnten, muß die Parametrisierung entsprechend angepaßt werden.

- *Stabilität des Ergebnisses:* Bei komplexen Objekten (wie den innerhalb des

Testszenarios verwendeten Spielzeugautos) mit einer Vielzahl von Merkmalspunkten ist u.U. ein nicht vollständiges Entfalten des *Neural Gas*-Netzes zu beobachten. Dies kann dann dazu führen, daß bestimmte Objektregionen gar nicht oder nur unzureichend repräsentiert werden mit negativen Folgen für die Diskriminierungseigenschaften des Modellnetzes.

Besonders die Probleme im Zusammenhang mit der Parametrisierung sind typisch für Erkennungssysteme und in ähnlicher Ausprägung in vielen Objekterkennungsansätzen wiederzufinden. Zusammenfassend ist festzustellen, daß Graphenstrukturen ein adäquates Mittel für die Repräsentation von Objekteigenschaften darstellen. Die Generierung dieser Repräsentation durch den angepaßten *Neural Gas*-Algorithmus stellt allerdings noch nicht das Optimum dar.

7.3.2 Erkennung

Die im vorigen Kapitel entworfenen Matchstrategien sollen nun an dem gewählten Anwendungsszenario getestet werden. Innerhalb des Testbetriebes wirkten sich vor allem die im vorigen Kapitel erläuterten Mängel des Repräsentationsschemas negativ aus. Sowohl das Hashverfahren als auch das Relaxations- und Graphensuchverfahren setzen stabil generierte Szenennetze voraus. Die mangelnde Reproduzierbarkeit ist hierbei der größte Störfaktor. Da sich die während der Netzgenerierung verwendeten Zufallsprozesse auf alle Eigenschaften der resultierenden Netze, wie Anzahl und Position der Neuronen bzw. Neuronenverbindungen, auswirken, gestaltet sich das Matchen von *Neural Gas*-Netzen der präsentierten Szene mit den Modellnetzes als äußerst instabil.

Neben den Problemen der eigentlichen Erkennung gelernter Objekte wegen der geringen Reproduzierbarkeit der *Neural Gas*-Netze konnten auch die Erwartungen an eine Transformationsschätzung für präsentierte Szenenobjekte innerhalb des Hashverfahrens nicht erfüllt werden. Diese Transformationsschätzung leitete sich aber ebenfalls aus den gefundenen Matchpartnern ab, so daß hier die gleichen Ursachen wie für die Erkennungsergebnisse ausschlaggebend sind.

Für das Relaxationsverfahren erwies sich weiterhin die Festlegung des globalen Fehlermaßes sowie der durchzuführenden Modifikationen innerhalb der Matchmatrix als äußerst kritisch. Auch bei gleichen Einstellungen entstanden oft sehr unterschiedliche Ergebnisse. Es ist fraglich, ob überhaupt die vorgeschlagene Matchrepräsentation innerhalb einer Matrix adäquat ist, um solche komplexen Objekte, wie die hier verwendeten, in Beziehung setzen zu können. Innerhalb der untersuchten, bereits realisierten Erkennungsansätze (siehe Kapitel 3.4) wurden meist nur sehr einfache geometrische Objekte untersucht.

Entsprechend den dargestellten Problemen wurde die Weiterentwicklung der oben genannten Matchverfahren zugunsten des Deformationsverfahrens nicht

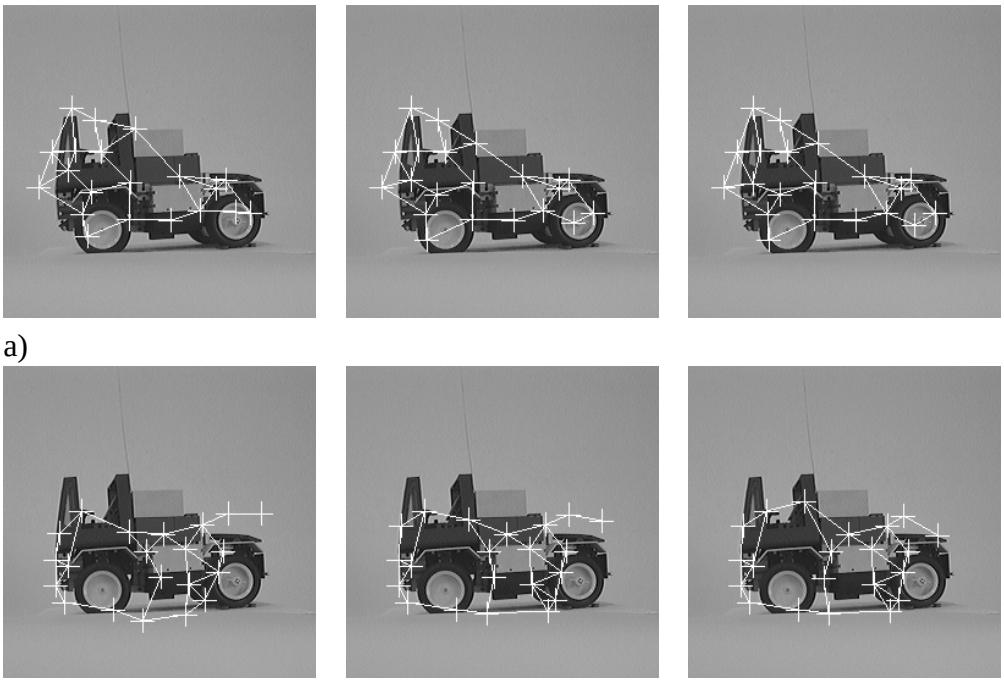
weiter verfolgt, da dieses nicht von der Generierung der *Neural Gas*-Netze für die präsentierten Szenen abhängt, sondern die Modellnetze mit den Merkmalsbildern der Szene in Beziehung setzt.

Zur Demonstration der prinzipiellen Funktionsweise des Deformationsmatchverfahrens werden im folgenden zwei exemplarische Matchvorgänge demonstriert. Tabelle 7.2 zeigt die dazu verwendeten Parameter. In Abbildung 7.14 sind jeweils Zwischenzustände für das Matchen der Ansicht 0 des Objektes 2 gegenüber den Modellen der Ansicht 1 von Objekt 2 bzw. Objekt 1 dargestellt. Ersteres Modell wird kaum deformiert, schnell haben die Neuronen passende Merkmalspositionen gefunden und verändern ihre Positionen nur marginal. Anders beim zweiten Modell. Dort finden vor allem im linken oberen und rechten oberen Bereich stärkere Veränderungen statt. Diese führen zu einem größeren Deformationsmaß für das Modell von Objekt 1 und damit zur korrekten Klassifikation als Objekt 2.

Tabelle 7.2: Parameterwerte zur Modellnetzdeformation

Parameter	Wert
f_{rez}	13
F_b	0.08
F_n	0.01
D_{straf}	180
β, γ	0.5

In Tabelle 7.3 ist eine Übersicht über die Ergebnisse des gesamten Testlaufes angegeben. Es wurden jeweils alle 52 als Szenenbilder deklarierte Objektansichten mit allen 48 Modellen in Beziehung gesetzt. Aufgeführt ist jeweils die Objekt- und Ansichtennummer für das präsentierte Szenenbild sowie die durch das Deformationsverfahren bestimmte Objekt- und Ansichtennummer. In Tabelle 7.4 ist ein Ausschnitt aus der gesamten Tabelle mit allen 2496 ($52 \cdot 48$) Matchwerten angegeben. Abgebildet sind die Matchwerte nach Gleichung 7.3 bei der Präsentation der Ansichten mit geradem Index von Objekt 2 unter Verwendung der Ansichten mit ungeradem Index von Objekt 2 als Modellansichten. Die ersten beiden Spalten geben das Szenenobjekt und dessen Ansicht an. Die ersten beiden Zeilen bezeichnen Modellobjekt und -ansicht. Maxima repräsentieren das Klassifikationsergebnis und sind fett markiert. Eine korrekte Modellzuordnung würde nicht nur die zutreffende Objektnummer sondern auch eine ähnliche Modellansicht liefern. Dementsprechend repräsentieren insgesamt 42 Klassifikationen das richtige Modellobjekt. 34 liefern ebenfalls die ähnlichste Objektnummer. In Anbetracht der geringen Unterschiede zwischen benachbarten Ansichten ist dies ein befriedi-



a)

b)

Abbildung 7.14: Matchen eines Szenenbildes (Ansicht 0 vom Objekt 2) gegen zwei verschiedene Modelle (Ansicht 1 jeweils vom Objekt 2 (*a*) und 1 (*b*)). Gezeigt sind die präsentierten Szenenbilder mit überlagertem *Neural Gas*-Netz.

gendes Ergebnis.

Ein besonderes Problem stellt die Hintergrundabhängigkeit des Erkennungssystems dar. Gerade hier bietet die im Kapitel 4.4 eingeführte Stereobildverarbeitung Ansatzpunkte. Durch eine Objekt-Hintergrund-Trennung kann irritierende Hintergrundinformation ausgeblendet werden. In den Abbildung 7.15 und 7.16 werden die Ergebnisse der Stereovorverarbeitung und die Klassifikationsergebnisse an einem Beispiel demonstriert. Präsentiert wurden das Objekt 0 und 1 vor einem stark strukturierten Hintergrund - hier ein Bücherregal. Die Objekte konnten gut aus der Szene herausgelöst werden. Zur Erkennung mußten die Parameter für den Deformationsprozeß angepaßt werden, da bisher alle Experimente ohne Hintergrund durchgeführt wurden. Tabelle 7.5 zeigt die hier angewendeten Einstellungen. Die Objekte konnten gut erkannt werden. Für das Objekt 0 wurde die ähnlichste Ansicht gefunden; die Modellzuordnung für Objekt 1 lieferte nicht die ähnlichste Ansicht, aber immer noch eine gute Näherung für die präsentierte Szene.

Hinsichtlich der Objekt-Hintergrund-Trennung mittels Stereobildverarbeitung

Tabelle 7.3: Modell- bzw. Ansichtenzuordnung für die 52 Szenenbilder. Die ersten beiden Spalten bezeichnen jeweils das Szenenobjekt S mit einer Ansichtennummer SA, die folgenden zwei Spalten das klassifizierte Objekt M mit der bestimmten Ansicht MA.

S	SA	M	MA	
0	0	0	13	
0	2	0	15	
0	4	1	5	
0	6	0	7	
0	8	0	19	[6]
0	10	1	11	
0	12	0	15	
0	14	0	17	[6]
0	16	0	15	
0	18	0	19	
0	20	0	17	
0	22	0	13	
0	24	0	23	[6]
1	0	0	15	
1	2	1	1	[6]
1	4	1	1	
1	6	0	7	[6]
1	8	1	11	
1	10	1	11	[6]
1	12	1	13	
1	14	1	13	[6]
1	16	1	17	
1	18	0	7	
1	20	1	21	
1	22	1	5	
1	24	1	1	

S	SA	M	MA	
2	0	2	21	
2	2	2	19	[6]
2	4	2	5	
2	6	1	7	
2	8	2	9	
2	10	2	11	[6]
2	12	2	11	
2	14	0	13	
2	16	2	19	
2	18	2	17	
2	20	2	21	
2	22	2	21	
2	24	2	23	
3	0	1	19	
3	2	1	19	[6]
3	4	3	17	
3	6	3	1	
3	8	3	7	[6]
3	10	3	11	
3	12	3	13	
3	14	3	15	
3	16	3	17	
3	18	3	17	
3	20	3	19	
3	22	3	23	[6]
3	24	1	1	[6]

[6] Klassifikation erfolgte nach der minimalen Iterationszahl

ergaben sich allerdings eine Reihe von Unzulänglichkeiten. Das eingesetzte Verfahren zur Disparitätsberechnung stellt ein erweitertes Verfahren zu dem im Kapitel 4.4 vorgestellten dar. Es wird z.Z. im Rahmen einer Diplomarbeit weiterentwickelt [LIEDER 1999]. Folgende, sich zum Teil gegenseitig bedingende, Probleme wurden beobachtet: Der eingesetzte Algorithmus geht von minimalen Unter-

Tabelle 7.4: Matchergebnisse für die winkelbasierte Deformationsbewertung (Ausschnitt; S=Szenenobjekt; SA=Szenenobjektansicht; M=Modellobjekt; MA=Modellobjektansicht; nach [SCHMALZ und MERTSCHING 1999])

	M	...	2												...
	MA	...	1	3	5	7	9	11	13	15	17	19	21	23	...
S	SA														
	⋮														
2	0		0.72	0.51	0.60	0.47	0.31	0.43	0.46	0.65	0.66	0.61	**0.82**	0.76	[7]
	2		0.81	0.51	0.57	0.50	0.26	0.43	0.53	0.55	0.51	**0.82**	0.75	0.56	[8]
	4		0.72	0.84	**0.90**	0.62	0.61	0.42	0.40	0.78	0.33	0.70	0.78	0.30	
	6		0.24	0.28	0.43	0.76	0.72	0.36	0.59	0.77	0.82	0.72	0.60	0.49	[9]
	8		0.31	0.43	0.40	0.56	**1.00**	0.82	0.67	0.89	0.85	0.66	0.66	0.39	
	10		0.45	0.39	0.39	0.46	0.81	**0.97**	0.69	0.90	0.62	0.76	0.63	0.43	[8]
	12		0.43	0.30	0.34	0.49	0.69	**0.99**	0.88	0.92	0.54	0.74	0.54	0.20	
	14		0.42	0.42	0.58	0.43	0.81	0.63	0.76	0.82	0.72	0.82	0.73	0.24	[9]
	16		0.34	0.53	0.63	0.67	0.63	0.59	0.70	0.85	0.88	**0.92**	0.54	0.50	
	18		0.43	0.54	0.53	0.69	0.68	0.45	0.73	0.77	**0.95**	0.83	0.68	0.44	
	20		0.62	0.57	0.68	0.35	0.50	0.44	0.35	0.58	0.71	0.87	**0.89**	0.28	
	22		0.44	0.64	0.46	0.47	0.36	0.52	0.59	0.70	0.41	0.86	**0.91**	0.78	
	24		0.87	0.61	0.62	0.49	0.48	0.55	0.28	0.53	0.59	0.80	0.65	**0.95**	
	⋮														

[7] Hinweis: Modellansicht Nr. 21 ist hier korrekt, da alle Modellansichten eine vollständige Rotation des Objekts beschreiben.

[8] Klassifikation erfolgte auf der Basis der minimalen Iterationszahl

[9] falsches Matchergebnis; maximaler Matchwert ergab sich nicht für Objekt Nr. 2

Tabelle 7.5: Parameterwerte zur Modellnetzdeformation bei vorhandenem Hintergrund

Parameter	Wert
f_{rez}	7
F_b	0.15
F_n	0.08
D_{straf}	180
β, γ	0.5

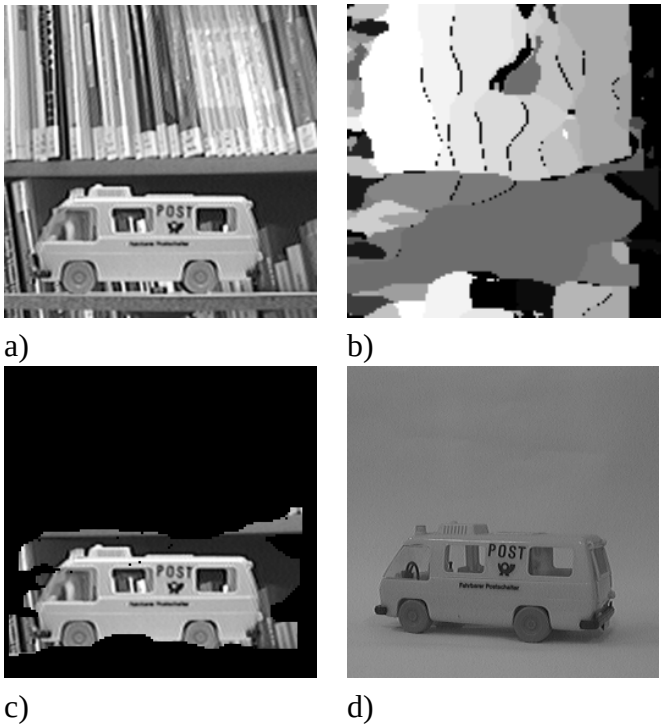


Abbildung 7.15: Experiment I mit Einbeziehung einer Objekt-Hintergrund-Trennung basierend auf einer Disparitätsanalyse: *a)* linkes Kamerabild; *b)* generiertes Disparitätsfeld; *c)* extrahierter Szenenbereich; *d)* ermittelte ähnlichste Ansicht

schieden zwischen dem linken bzw. rechten Kamerabild aus. Dies erfordert eine kleine Stereobasis (durch den vorhandenen Aufbau nicht gegeben) bzw. eine größere Objektentfernung. Durch die größere Entfernung werden jedoch die Objekte nur sehr klein abgebildet, wodurch wiederum keine ausreichende Merkmalsdetektion mit den hier verwendeten Verfahren möglich ist. Als Ausweg dazu wurden die Zoom-Möglichkeiten der Kameras eingesetzt. Leider machte sich beim Zoomeinsatz der Effekt der sinkenden Schärfentiefe stark bemerkbar. Die Objekte wurden dadurch relativ unscharf abgebildet. Natürlich werden in Zuge der Merkmalsberechnung Hochpaß-Verfahren eingesetzt. Jedoch zeigen sich trotzdem deutliche Effekte in den Merkmalsbildern vor allem hinsichtlich Position und Stärke der Merkmalspunkte. Zu erkennen ist jedoch, daß die Stereobildverarbeitung die Mittel liefert, um eine Objekt-Hintergrund-Trennung und weitergehend auch eine metrische Raumrekonstruktion zur Erhöhung der Skalierungsinvarianz durchzuführen. Die z.Z. auftretenden rein technischen Probleme sind jedoch in weiteren Ausbaustufen lösbar.

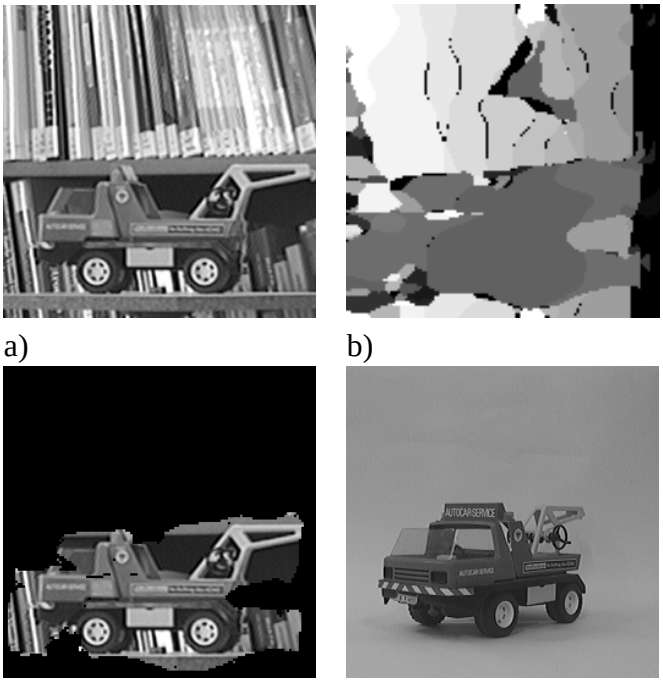


Abbildung 7.16: Experiment II mit Einbeziehung einer Objekt-Hintergrund-Trennung basierend auf einer Disparitätsanalyse: *a)* linkes Kamerabild; *b)* generiertes Disparitätsfeld; *c)* extrahierter Szenenbereich; *d)* ermittelte ähnlichste Ansicht

Auch wenn die erzielten Ergebnisse des rein merkmalsbasierten Systems aus Kapitel 6 besser waren (vergleiche Tabelle 6.1), so konnte doch das Potential des Erkennungssystems basierend auf der Kodierung der globalen Objektstruktur gezeigt werden. Trotz der Verwendung primitiver geometrischer Merkmale und einfacher heuristischer Regeln zur Aufbau und zur Deformation der Modellnetzwerke konnten in einer schwierigen Erkennungsaufgabe respektable Ergebnisse erzielt werden, die die prinzipielle Herangehensweise rechtfertigen.

Das Erkennungssystem mit zeitlichen Relationen beinhaltete bereits Mechanismen, um zum einen Fehlklassifikationen durch eine Analyse der Ansichtensequenzen zu kompensieren. Diese Sequenzauswertung bietet sich ebenfalls für das Erkennungssystem mittels Strukturbeschreibungen an. Auch hier finden Fehlklassifikationen für ähnliche Ansichten unterschiedlicher Objekte statt. Durch die Analyse des zeitlichen Kontexts können diese berichtigt werden. Zum anderen stellte das System auch Verfahren zur selbständigen Auswahl typischer und zu lernender Objektansichten bereit, deren Anwendung ebenso hier wichtig ist.

Die weiteren in Tabelle 3.3 aufgeführten Entwurfskriterien konnten nur zum Teil erfüllt werden. Invarianzen sind z.Z. durch die angestrebte Einbindung in ein aktives Sehsystem hinsichtlich Translation gegeben. Das Erkennungssystem selbst ist im gewissen Maße skalierungsinvariant, da die Modellnetzrepräsentation keine Aussage über die Größe des Objekts beinhaltet. Der Deformationsprozeß wiederum toleriert auch Größenänderungen, allerdings dürfte bei zu starker Skalierung die Merkmalsbasis nicht stabil bleiben und damit den Deformationsprozeß fehlleiten. Dieses Problem trat auch schon bei der Erkennung mittels Modellskalierungen (siehe Kapitel 5) auf. Bezüglich Rotationsinvarianz ist zu sagen, daß auch hier im Rahmen der Modelldeformationen Toleranzen akzeptiert werden. Prinzipiell wird aber ansichtenbasiert gearbeitet, d.h. jede charakteristische Ansicht wird als ein neues Modell aufgefaßt. Als weiteres Entwurfskriterium wurde eine hohe Einsatzflexibilität aufgeführt. Dies wird durch die Erkennung mittels Strukturbeschreibung gewährleistet. Eine Domänenabhängigkeit ist insofern allerdings immer gegeben, daß die verwendeten Merkmale auch detektierbar bzw. in genügender Anzahl vorhanden sein müssen. Betreffs der Parameterfestlegung ist anzumerken, daß dafür einfache Regeln (siehe Kapitel 7.3.1) aufgestellt wurden, die meist zuverlässig arbeiten. Abschließend ist dieses Problem, wie für die meisten Erkennungssysteme, jedoch nicht vollständig gelöst. Die Entwurfskriterien biologische Plausibilität, stabile Verarbeitung von Okklusionen und Geschwindigkeit konnten hier zunächst nicht umfassend eingearbeitet werden.

Hinsichtlich des eingangs des Kapitels formulierten Zieles, eine hohe Speichereffizienz zu gewährleisten, ist festzustellen, daß die gewählte, graphenähnliche Repräsentationsform eine sehr effektive Speicherung darstellt, da nur charakteristische Punkte und deren räumliche Relationen zueinander in die Datenbasis einfließen.

Kapitel 8

Zusammenfassung und Ausblick

Ziel der vorliegenden Arbeit war der Entwurf und die anschließende Evaluierung von Objekterkennungssystemen, wobei die Möglichkeiten einer aktiven Sehumgebung berücksichtigt werden sollten. Diese Untersuchungen wurden im Rahmen des Promotionsprojektes von Herrn Steffen Schmalz vorgenommen. Er übernahm innerhalb des NAVIS-Projektes die verantwortliche Rolle im Bereich der Objekterkennung. Dabei entstanden die dargestellten Erkennungssysteme mittels Modellskalierungen und unter Nutzung von zeitlichen Relationen sowie die Objekt-Hintergrund-Trennung durch eine Stereobildverarbeitung und die Generierung primärer Strukturmerkmale als Arbeiten von Studierenden unter seiner direkten Betreuung. Das Erkennungssystem unter Verwendung von Strukturbeschreibungen stellt demgegenüber ein selbständig entworfenes System dar. Unter seiner federführenden Arbeit konnten während seiner fünfjährigen Tätigkeit u.a. die dargestellten Fortschritte innerhalb der Objekterkennung erzielt werden. Es wurden dabei verschiedene Forschungsrichtungen eingeschlagen und bewertet, die alle auf das Ziel einer stabilen Objekterkennung in natürlichen Umgebungen gemäß den Kriterien aus Tabelle 3.3 ausgerichtet waren.

Den Ausgangspunkt der Untersuchungen bildete eine umfassende Analyse von internationalen Forschungsaktivitäten bzw. realisierten Ansätzen zum Thema (siehe Kapitel 2.2 und 3). Dabei wurde vor allem deutlich, daß der Umgang mit Objekterkennungssystemen in aktiven Sehumgebungen noch starken Forschungscharakter besitzt. Oft werden nur primitive Muster erkannt bzw. starke Einschränkungen bezüglich des Einsatzgebietes gemacht. In realisierten, mobilen Systemen wird visuelle Information meist für Navigationszwecke verwendet. Nur selten findet auch eine komplexe Objekterkennung statt. Der besondere Vorteil sichtgesteuerter Plattformen liegt aber gerade in den vielfältigen Anwendungsmöglichkeiten visueller Daten, wie Navigation, Hindernisvermeidung und Objekterkennung.

Die Objekterkennung als wesentliche Komponente visuell agierender Systeme wird noch lange Forschungsgegenstand bleiben. Z.Z. sind die entstandenen

Systeme nicht in der Lage, die Leistungsfähigkeit und Robustheit natürlicher Systeme auch nur annähernd zu erreichen. Für Teilproblematiken in abgegrenzten Anwendungsfeldern lassen sich jedoch gute Ergebnisse erzielen. Die im Kapitel 3 beschriebenen Ansätze verdeutlichen dies. Die Konzepte des aktiven Sehens, d.h. eine aufgabenabhängige, schrittweise Szenenanalyse mit der Möglichkeit, selbständig neue Objektinformationen zu generieren, bieten Möglichkeiten, die nötige Vorverarbeitung sowie den Suchraum für Objekterkennungsmodule günstiger zu gestalten. Dadurch ließen sich neue Anwendungsfelder erschließen.

Im Zusammenhang mit den Forschungsarbeiten innerhalb des NAVIS-Projektes wurden drei entstandene Erkennungssysteme vorgestellt, die auf die Erfahrungen aus dem 2D-Erkennungssystem von Götze und Mertsching (siehe Kapitel 4.3) aufsetzten. In ausgewählten Anwendungsfällen funktionierten alle gut. Die 2D-Erkennung hatte Probleme mit dem Umgang mit Objekttransformationen, konnte dafür aber durch ihren robusten Umgang mit stark strukturiertem Hintergrund und durch ihre effektive Anbindung an Aktorik überzeugen. Anschließend wurde untersucht, inwieweit eine stärkere Skalierungsinvarianz durch eine Stereobildanalyse im Rahmen des 2D-Erkennungssystems erreicht werden kann (siehe Kapitel 5). Dabei konnten hinsichtlich einer skalierungsinvarianten Objekterkennung gute Ergebnisse erzielt werden, allerdings blieben die prinzipiellen Restriktionen des 2D-Ansatzes erhalten. Der Übergang von einer 2D- auf eine 3D-Erkennung erfolgte anschließend durch das Erkennungssystem mit zeitlichen Relationen (siehe Kapitel 6). Hierdurch wurde eine stabile, ansichtenbasierte Objekterkennung zur Lösung erheblich komplexerer Aufgabenstellungen realisiert. Weiterhin wurden Verfahren entwickelt, um selbständig zu lernende Ansichten auszuwählen und zwischenzeitliche Fehlklassifikationen durch eine Analyse des temporalen Kontexts zu kompensieren. Die zunächst nötigen Einschränkungen bezüglich Hintergrund bzw. Anbindung an ein aktives Sehsystem entsprechen dem Entwicklungscharakter des Systems und sollten somit Gegenstand weiterer Untersuchungen sein. Im Kapitel 7 wurde sich der prinzipiellen Frage gewidmet, inwieweit strukturelle Objektbeschreibungen Vorteile gegenüber der Verwendung hochdimensionaler Merkmalsräume bieten. Dazu wurde konkret die Ansichtenerkennungsleistung des *Dynamic Link Matching* innerhalb des Erkennungssystems mit zeitlichen Relationen mit einem neu entworfenen System basierend auf strukturellen Repräsentationen und Deformationsverfahren verglichen. Festzustellen ist, daß das Erkennungssystem mit Strukturbeschreibungen trotz der Verwendung primitiver geometrischer Merkmale und einfacher heuristischer Regeln zur Ausbildung und Deformation von Modellnetzwerken respektable Ergebnisse erzielte. Besonders im Zusammenwirken mit einer stereobasierten Objekt-Hintergrund-Trennung konnte die Leistungsfähigkeit demonstriert werden (siehe Abbildungen 7.15 und 7.16). Die Grundidee, Objekte durch wenige charakteristische Punkte und deren räumliche Relationen zueinander zu beschreiben, sollte

somit beim Entwurf von Erkennungssystemen in aktiven Sehumgebungen weitere Beachtung finden. Die eingangs gestellte Frage, ob zukünftige Erkennungssysteme ihren Schwerpunkt auf die Verwendung hochdimensionaler Merkmalsräume setzen sollten oder besser auf strukturelle Objektbeschreibungen, kann jedoch nicht abschließend beantwortet werden und bedarf weiterer Untersuchungen.

Umfangreiche Experimente boten Raum für substantielle Kritik. In den einzelnen Ergebniskapiteln (5.2, 6.3 und 7.3) wird eine gründliche Analyse der erzielten Ergebnisse vorgenommen. Darüber hinaus ergeben sich folgende prinzipielle Verbesserungsmöglichkeiten:

- Die entworfenen Erkennungssysteme benötigen eine stärkere Einbindung in eine aktive Sehumgebung. Dadurch ließen sich analog [HERBIN 1996], [SIPE und CASASENT 1997] oder [GVOZDJAK und LI 1998] (siehe Kapitel 3.6) Systeme bauen, die selbständig neue Objektansichten generieren, um in unklaren Entscheidungssituationen zu gültigen Ergebnissen zu kommen.

- Gerade für das Verfahren mit Strukturbeschreibungen ist die Einbeziehung weiterer, auch komplexerer, Merkmale wünschenswert. Zu nennen wären hier Farbmerkmale, parallele Strukturen, Texturen, bezüglich des eingeschlossenen Winkels diskretisierte Ecken und verschiedene Kreisradien. In diesem Zusammenhang ist eine enge Kopplung mit Aufmerksamkeitsmechanismen anzustreben, die je nach Intention des Beobachters attraktive Punkte aus der Szene isolieren.

- In den dargestellten Erkennungssystemen wurde sich nicht mit dem Problem des Nachlernens beschäftigt, d.h. inwieweit Änderungen der Datenbasis für bereits gelernte Objekte möglich sind und ob neu aufgenommene Objekte eine Neugenerierung bereits vorhandener Modelle nötig machen.

- Es ist zu prüfen, ob die Strukturbeschreibungen von benachbarten Ansichten eines Objekts in ein Modell zusammengefaßt werden können. Der Aufbau eines Würfels aus den einzelnen Außenflächen wäre ein einfaches Beispiel für diese Vorgehensweise.

- Als wichtigster Kritikpunkt muß die für die *Neural Gas*-Matchverfahren (konkret für das Hash-, Relaxations- und Graphensuchverfahren) zum Teil noch nicht ausreichende Stabilität der entstehenden Netze genannt werden. Die oben vorgeschlagenen Modifikationen können zu deren Verbesserung beitragen.

Eine Reihe von interessanten Ansatzpunkten aus der Literatur konnten zunächst nicht mit eingearbeitet werden: Vielversprechend erscheint die Idee aus

[ZHANG et al. 1992], die Kanten von Graphenstrukturen mit einem Sichtbarkeitsindex zu versehen, um so neben der Objektidentität auch deren Lage bzw. Verdeckungen verarbeiten zu können. Damit ließen sich zum Beispiel auch Übergänge von benachbarten Ansichten modellieren. In [MAGGIONI und WIRTZ 1991] werden durch künstliche neuronale Netze Vorhersagen für Objekttransformationen gemacht. Dadurch wären Verbesserungen der Invarianzleistungen denkbar. In [AZENCOTT und YOUNES 1997] werden Objektbeschreibungen nicht direkt mit Szenen gematcht, sondern es wird hierarchisch vorgegangen. Objektbeschreibungen sind dabei ähnlich, wenn sich vereinfachte, zusammengefaßte Beschreibungen ähneln. Mit diesem Ansatz könnte eine starke Beschleunigung der Systeme erreicht werden, da so leicht zu diskriminierende Objekte schneller klassifiziert werden können.

Zusammengefaßt ist festzustellen, daß die Verwendung von Strukturbeschreibungen als Objektrepräsentation als eine vielversprechende Forschungsrichtung innerhalb der Objekterkennung anzusehehn ist. Diese symbolische Verarbeitung von Objektwissen stellt eine konträre Entwicklungsrichtung zu den rein merkmalsbasierten Systemen dar. Beide Varianten erzielen gute Ergebnisse, sind aber hinsichtlich zahlreicher Aspekte verbesserungsfähig. Die Frage nach dem robustesten bzw. effektivsten System kann nicht allgemeingültig beantwortet werden. Das Kriterium der praktischen Handhabbarkeit der Erkennungssysteme wird ausschlaggebend für deren Einsatz sein.

Literaturverzeichnis

[AIYER et al. 1990] AIYER, S. V. B., M. NIRANJAN und F. FALLSIDE (1990). *A theoretical investigation into the performance of the Hopfield Model*. IEEE Transactions on Neural Networks, 1(2):204–215.

[ALEXANDRE et al. 1991] ALEXANDRE, F., F. GUYOT, J.-P. HATON und Y. BURNOD (1991). *The Cortical Column: A new processing unit for multilayered networks*. Neural Networks, 4:15–25.

[ALOIMONOS 1990] ALOIMONOS, Y. (1990). *Purposive and qualitative active vision*. In: *Proc. 10th Int. Conf. on Pattern Recognition*, S. 346–360, Atlantic City.

[ALOIMONOS 1993] ALOIMONOS, Y. (1993). *Introduction: Active vision revisited*. In: ALOIMONOS, Y., Hrsg.: *Active perception*. Lawrence Erlbaum, Hillsdale.

[AMIN et al. 1998] AMIN, A., D. DORI, P. PUDIL und H. FREEMANN, Hrsg. (1998). *Advances in Pattern Recognition. Joint IAPR International Workshops SSPR'98 and SPR'98*, Session Structural Matching and Grammatical Inference, Sydney. Springer.

[ANDERSEN et al. 1985] ANDERSEN, R., G. ESSICK und R. SIEGEL (1985). *Encoding of spatial location by posterior parietal neurons*. Science, 230:456–458.

[ANDO 1996] ANDO, H. (1996). *3D Object Recognition using Bidirectional Modular Networks*. In: LI, S. Z., D. P. MITAL, E. K. TEOH und H. WANG, Hrsg.: *Recent Developments in Computer Vision*, S. 467–474. Springer, Berlin.

[ASADA und KITANO 1999] ASADA, M. und H. KITANO, Hrsg. (1999). *RoboCup-98: Robot Soccer World Cup II*, Paris. Springer.

[AZENCOTT und YOUNES 1997] AZENCOTT, R. und L. YOUNES (1997). *An energy minimization method for matching and comparing structured object*

representations. In: PELILLO, M. und E. R. HANCOCK, Hrsg.: *Energy minimization methods in computer vision and pattern recognition*, S. 441–456. Springer, Berlin.

[BAJCSY 1988] BAJCSY, R. (1988). *Active perception*. Proc. of the IEEE, 76(8):996–1005.

[BALLARD 1984] BALLARD, D. H. (1984). *Parameter Nets*. Artificial Intelligence, 22:235–267.

[BALLARD 1991] BALLARD, D. H. (1991). *Animate Vision*. Artificial Intelligence, 48(1):57–86.

[BALLARD und BROWN 1992] BALLARD, D. H. und C. M. BROWN (1992). *Priciples of animate vision*. Computer Graphics and Image Processing, 56(1):3–21.

[BARATOFF et al. 1999] BARATOFF, G., I. AHRNS, C. TOEPFER und H. NEUMANN (1999). *Ortsvariantes aktives Sehen: Von biologischer Motivation zu technischer Realisation*. KI, (1):33–35.

[BARROW und TENENBAUM 1978] BARROW, H. G. und J. M. TENENBAUM (1978). *Recovering Intrinsic Scene Characteristics from Images*. In: HANSON, A. R. und E. M. RISEMAN, Hrsg.: *Computer Vision Systems*, S. 3–26. Academic Press, New York.

[BARROW und TENENBAUM 1981] BARROW, H. G. und J. M. TENENBAUM (1981). *Computational Vision*. Proc. IEEE, 69(5):572–595.

[BASAK und PAL 1995] BASAK, J. und S. K. PAL (1995). *PsyCOP-a psychologically motivated connectionist system for object perception*. IEEE Transactions on Neural Networks, 6(6):1337–1354.

[BAYLISS et al. 1997] BAYLISS, J. D., C. M. BROWN, R. L. CARCERONI, C. EVELAND, C. HARMAN, A. SINGHAL und M. VAN WIE (1997). *Mobile Robotics 1997*. Technischer Bericht 661, Dept. of Computer Science, University of Rochester.

[BAYOUTH und THORPE 1996] BAYOUTH, M. und C. THORPE (1996). *An AHS Concept Based on an Autonomous Vehicle Architecture*. In: *Proc. of the 3rd Annual World Congress on Intelligent Transportation Systems*.

[BECKER et al. 1997] BECKER, C., H. GONZÁLEZ-BAÑOS, J. C. LATOMBE und C. TOMASI (1997). *An Intelligent Observer*. In: KHATIB, O. und J. K.

SALISBURY, Hrsg.: *Lecture Notes in Control and Information Sciences*, Nr. 223, S. 153–160. Springer, New York.

[BELLAIRE 1995] BELLAIRE, G. (1995). *Hashing with a Topological Invariant Feature*. In: *Proceedings of the ACCV 95*, S. 598–602, Singapure.

[BELLAIRE und LÜBBE 1995] BELLAIRE, G. und M. LÜBBE (1995). *Adaptive Hierachical Indexing and Constrained Localization: Matching Characteristic Views*. In: CHIN, R. T., H. H. S. IP, A. C. NAIMAN und T.-C. PONG, Hrsg.: *Image Analysis Applications and Computer Graphics*, S. 241–249. Springer, Berlin.

[BELLAIRE und SCHLÜNS 1996] BELLAIRE, G. und K. SCHLÜNS (1996). *3-D Object Recognition by Matching the Total View Information*. In: *Proceedings of the ICPR '96*, S. 534–538.

[BELLAIRE et al. 1996] BELLAIRE, G., K. SCHLÜNS, K. OPPERMANN und W. SCHIMKE (1996). *Matching using Object Models Generated from Photometric Stereo Images*. In: *Proceedings of the Photonics West Symposium*, Bd. IV - Machine Vision Applications in Industrial Inspections, San Jose.

[BENNAMOUN und BOASHASH 1997] BENNAMOUN, M. und B. BOASHASH (1997). *Structural-Description-Based Vision System for Automatic Object Recognition*. IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics, 27(6):893–906.

[BERGENER et al. 1997] BERGENER, T., C. BRUCKHOFF, P. DAHM, H. JANSSEN, F. JOUBLIN und R. MENZNER (1997). *Arnold: An Anthropomorphic Autonomous Robot for Human Environments*. In: GROSS, H.-M., Hrsg.: *Proc. of SOAVE '97 - Selbstorganisation von Adaptivem Verhalten*, S. 25–34, Ilmenau. VDI Verlag.

[BERTHOLD 1996] BERTHOLD, M. R. (1996). *A Probabilistic Extension for the DDA Algorithm*. In: *Proceedings of the IEEE International Conference on Neural Networks*, Bd. 1, S. 341–346.

[BERTHOLD und DIAMOND 1995] BERTHOLD, M. R. und J. DIAMOND (1995). *Boosting the Performance of RBF Networks with Dynamic Decay Adjustment*. In: TESAURO, G., D. S. TOURETZKY und T. K. LEEN, Hrsg.: *Advances in Neural Information Processing Systems*, Nr. 7, S. 521–528. MIT Press, Cambridge.

[BERTHOLD und DIAMOND 1998] BERTHOLD, M. R. und J. DIAMOND (1998). *Constructive Training of Probabilistic Neural Networks*. Neurocomputing, (19):167–183.

[BIEDERMANN 1985] BIEDERMANN, I. (1985). *Human image understanding: Recent research and a theory*. Comput. Vision Graphics Image Processing, 32:29–73.

[BIEDERMANN et al. 1993] BIEDERMANN, I., J. HUMMEL, E. COOPER und P. GERHARDSTEIN (1993). *Shape recognition in mind, brain, and machine*. In: *Proceeding of a U.S.-Mexico Seminar Neuroscience: From Neural Networks to Artificial Intelligence*, S. 282–293.

[BIESZCZAD 1994] BIESZCZAD, A. (1994). *Neurosolver: A Step Towards a Neuromorphic General Problem Solver*. In: *Proceedings of the IEEE International Conference on Neural Networks, IEEE World Congress on Computational Intelligence*, Bd. 3, S. 1313–1318.

[BISCHOF und CAELLI 1997] BISCHOF, W. F. und T. CAELLI (1997). *Visual Learning of Patterns and Objects*. IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics, 27(6):907–917.

[BLAKE und YUILLE 1992] BLAKE, A. und A. YUILLE (1992). *Introduction*. In: BLAKE, A. und A. YUILLE, Hrsg.: *Active vision*. MIT Press, Cambridge.

[BÜLTHOFF et al. 1995] BÜLTHOFF, H., S. EDELMAN und M. TARR (1995). *How are three-dimensional object represented in the brain*. Cerebral Cortex, 5(3):247–260.

[BODENDIEK und LANG 1995] BODENDIEK, R. und R. LANG (1995). *Lehrbuch der Graphentheorie*, Bd. 1. Spektrum, Heidelberg.

[BOLLMANN 1999] BOLLMANN, M. (1999). *Entwicklung einer Aufmerksamkeitssteuerung für ein aktives Sehsystem*. Doktorarbeit, Universität Hamburg, Fachbereich Informatik, AG IMA.

[BOLLMANN et al. 1999] BOLLMANN, M., R. HOISCHEN, M. JESIKIEWICZ, C. JUSTKOWSKI und B. MERTSCHING (1999). *Playing Domino: A Case Study for an Activ Vision System*. In: *Proceedings of the first International Conference on Vision Systems (ICVS'99)*, S. 392–411, Las Palmas de Gran Canaria. Springer.

[BOOKSTEIN 1984] BOOKSTEIN, F. L. (1984). *A statistical method for biological shape comparisons*. Journal of Theoretical Biology, 107:475–520.

[BRADSKI et al. 1992] BRADSKI, G., G. CARPENTER und S. GROSSBERG (1992). *Working memory networks for learning temporal order with application to three-dimensional visual object recognition*. Neural Computation, 4:270–286.

[BUHMANN et al. 1995] BUHMANN, J., W. BURGARD, A. B. CREMERS, D. FOX, T. HOFMANN, F. SCHNEIDER, S. J. und S. THRUN (1995). *The Mobile Robot Rhino*. AI Magazin, 16(2):31–38.

[BUNKE und MESSMER 1995] BUNKE, H. und B. T. MESSMER (1995). *Efficient Attributed Graph Matching and its Application to Image Analysis*. In: BRACCINI, C., L. DEFLORIANI und G. VERNAZZA, Hrsg.: *Image Analysis and Processing*, S. 45–55. Springer, Berlin.

[BURGARD et al. 1998] BURGARD, W., A. B. CREMERS, D. FOX, D. HÄHNEL, G. LAKEMEYER, D. SCHULZ, W. STEINER und S. THRUN (1998). *The Interactive Museum Tour-Guide Robot*. In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI'98)*, Madison.

[BURGER et al. 1996] BURGER, W., M. BURGE und W. MAYR (1996). *Learning to Recognize Generic Visual Categories using a Hybrid Structural Approach*. In: *Proc. IEEE Int. Conf. on Image Processing*, S. 321–324, Lausanne.

[CARPENTER et al. 1992] CARPENTER, G., S. GROSSBERG, N. MARKUZON, J. REYNOLDS und D. ROSEN (1992). *Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps*. IEEE Transactions on Neural Networks, 3(5):698–713.

[CHAO und DHAWAN 1992] CHAO, C.-H. und A. P. DHAWAN (1992). *Artificial Neural Networks and Model-based Recognition of 3-D Objects from 2-D Images*. In: *Proceedings of the SPIE: Applications of Artificial Neural Networks III*, Bd. 1709, S. 119–130.

[CHELAZZI et al. 1993] CHELAZZI, L., E. MILLER, J. DUNCAN und R. DESIMONE (1993). *A neural basis for visual search in inferior temporal cortex*. Nature, 363:345–347.

[CHRISTENSEN und CROWLEY 1998] CHRISTENSEN, H. I. und J. L. CROWLEY (1998). *Intelligent robotics systems*. Robotics and Autonomous Systems, 23:201–204.

[CHUNG 1995] CHUNG, R. (1995). *Object Locating using a Single Model View*. In: *Proceedings of the International Symposium on Computer Vision*, S. 545–550, Coral Gables.

[COSTA und SHAPIRO 1996] COSTA, M. S. und L. G. SHAPIRO (1996). *Relational Indexing*. In: PERNER, P., P. WANG und A. ROSENFELD, Hrsg.: *Advances in Structural and Syntactical Pattern Recognition*, S. 130–139. Springer, Berlin.

[CROSS und HANCOCK 1996] CROSS, A. D. J. und E. R. HANCOCK (1996). *Inexact Graph Matching with Generic Search*. In: PERNER, P., P. WANG und A. ROSENFELD, Hrsg.: *Advances in Structural and Syntactical Pattern Recognition*, S. 150–159. Springer, Berlin.

[CRUZ et al. 1990] CRUZ, V., G. CRISTOBAL, T. MICHAUX und S. BARQUIN (1990). *Distortion Invariant Image Recognition by Madaline and Back-Propagation Learning Multi-Networks*. In: SOULIE, F. F. und J. HERAULT, Hrsg.: *Neurocomputing*, Bd. F 68 d. Reihe *NATO ASI Series*. Springer, Berlin.

[DICKINSON und METAXAS 1997] DICKINSON, S. J. und D. METAXAS (1997). *Using Aspect Graphs to Control the Recovery and Tracking of Deformable Models*. International Journal of Pattern Recognition and Artificial Intelligence, 11(1):115–141.

[DICKMANNS 1998] DICKMANNS, E. D. (1998). *Expectation-based, Multi-focal, Saccadic Vision for Percieving Dynamic Scenes (EMS-Vision)*. In: POSCH, S. und H. RITTER, Hrsg.: *Dynamische Perzeption*, S. 47–54, Bielefeld. infix.

[DIESTEL 1996] DIESTEL, R. (1996). *Graphentheorie*. Springer, Berlin.

[DRÜE et al. 1994] DRÜE, S., R. HOISCHEN und R. TRAPP (1994). *Tolerante Objekterkennung durch das Neuronale Active-Vision-System NAVIS*. In: BISCHOF, H. und W. G. KROPATSCH, Hrsg.: *Mustererkennung 1994*, S. 251–264, Technische Universität Wien. Informatik Xpress 5.

[DRÜE und TRAPP 1996] DRÜE, SIEGBERT und R. TRAPP (1996). *Ein flexibles binokulares Sehsystem: Konstruktion und Kalibrierung*. In: MERTSCHING, B., Hrsg.: *Aktives Sehen in technischen und biologischen Systemen*, Nr. 4 in *Proceedings in Artificial Intelligence*, S. 32–39. Infix.

[DRYDEN und MARDIA 1998] DRYDEN, I. L. und K. V. MARDIA (1998). *Statistical Shape Analysis*. John Wiley & Sons, Chichester.

[DURBIN und WILLSHAW 1987] DURBIN, R. und D. WILLSHAW (1987). *An analogue approach to the traveling salesman problem using an elastic net method*. Nature, 326(6114):689–691.

[ENS und LI 1995] ENS, J. und Z. N. LI (1995). *Real-time motion stereo on SFU pyramid*. Real-time Imaging, (1):385–396.

[ESHERA und FU 1984] ESHERA, M. A. und K.-S. FU (1984). *A Graph Distance Measure for Image Analysis*. IEEE Transactions on System, Man, and Cybernetics, 14(3):398–408.

[FAUSETT 1994] FAUSETT, L. (1994). *Fundamentals of Neural Networks. Architectures, Algorithms and Applications*. Prentice Hall, Englewood Cliffs.

[FINCH und HANCOCK 1995] FINCH, A. M. und E. R. HANCOCK (1995). *Matching Deformed Delaunay Triangulations*. In: *Proceedings of the International Symposium on Computer Vision*, S. 31–36, Coral Gables.

[FISHER und MACKIRDY 1998] FISHER, R. B. und A. MACKIRDY (1998). *Integrating Iconic and Structured Matching*. In: BURKHARDT, H.: NEUMANN, B., Hrsg.: *Proceedings of the European Comference on Computer Vision ECCV '98*, Bd. II, S. 687–698, Freiburg. Springer.

[FONTANE 1896] FONTANE, T. (1896). *Effi Briest*. F. Fontane & Co., Berlin.

[FRITZKE 1993] FRITZKE, B. (1993). *Growing Cell Structures - A Self-Organising Network for Unsupervised and Supervised Learning*. Technischer Bericht TR-93-026, International Computer Science Institut, Berkeley.

[FRITZKE 1995a] FRITZKE, B. (1995a). *Growing Grid - A Self-Organizing Network with Constant Neighborhood Range and Adaptation Strength*. Neural Processing Letters, 2(5):9–13.

[FRITZKE 1995b] FRITZKE, B. (1995b). *A Growing Neural Gas Network Learns Topologies*. In: TESAURO, G., D. S. TOURETZKY und T. K. LEE, Hrsg.: *Advances in Neural Information Processing Systems*, Bd. 7. MIT Press, Cambridge.

[FRITZKE 1995c] FRITZKE, B. (1995c). *Incremental Laerning of Local Linear Mappings*. In: FOGELMANN, F., Hrsg.: *Proceedings of ICANN'95. International Conference on Artificial Neural Networks*, S. 217–222, Paris.

[FRITZKE 1996] FRITZKE, B. (1996). *Growing Self-organizing Networks - Why?*. In: VERLEYSEN, M., Hrsg.: *ESANN '96. European Symposium on Artificial Neural Networks*, S. 61–72, Brussels. D-Facto.

[FUJITA et al. 1992] FUJITA, I., K. TANAKA, M. ITO und K. CHENG (1992). *Columns for visual features of objects in monkey inferotemporal cortex*. Nature, 360:343–346.

[FUKUNAGA et al. 1996] FUKUNAGA, K., N. NISHIKAWA, T. MATSUMOTO und M. IZUMI (1996). *Model-Based Object Recognition using Camera Control*. In: *Proc. of the ETFA '96. IEEE Conf. on Emerging Technologies and Factory Automation*, Bd. 2, S. 707–711, Hawaii.

[FUSIELLO et al. 1997] FUSIELLO, A., V. ROBERTO und E. TRUCCO (1997). *Experiments with a New Area-Based Stereo Algorithm*. In: DEL BIMBO, A., Hrsg.: *Proc. 9th International Conference On Image Analysis and Processing, ICIAP'97*, Bd. I, S. 669–676, Florence. Springer.

[GEE et al. 1991] GEE, A. H., S. V. B. AIYER und R. PRAGER (1991). *A subspace approach to invariant pattern recognition using Hopfield Networks*. In: *Proc. of the Intern. Joint Conf. on Neural Networks*, S. 795–800, Singapure.

[GIEFING et al. 1991] GIEFING, G.-J., H. JANSSEN und H. MALLOT (1991). *A Saccadic Camera Movement System for Object Recognition*. In: *Proceedings of ICANN 91*, S. 63–68.

[GIEFING et al. 1992a] GIEFING, G.-J., H. JANSSEN und H. MALLOT (1992a). *Saccadic Object Recognition with an Active Vision System*. In: *Proceedings of the IEEE International Conference on Pattern Recognition ICPR 92*, S. 664–667.

[GIEFING et al. 1992b] GIEFING, G.-J., H. JANSSEN und H. MALLOT (1992b). *Saccadic Object Recognition with an Active Vision System*. In: *Proceedings of the ECAI 92*, S. 803–805.

[GMÜR und BUNKE 1989] GMÜR, E. und H. BUNKE (1989). *3-D Object Recognition Based on Subgraph Matching in Polynomial Time*. In: MOHR, R., T. PAVLIDIS und A. SANFELIU, Hrsg.: *Structural Pattern Analysis*, S. 131–147. World Scientific.

[GOLD und RANGARAJAN 1996] GOLD, S. und A. RANGARAJAN (1996). *Graph Matching by Graduated Assignment*. In: *Proceedings of CVPR 96. Computer Vision and Image Processing*, S. 239–244.

[GONZÁLEZ-BAÑOS et al. 1999] GONZÁLEZ-BAÑOS, H. H., J. L. GORDILLO, D. LIN, J. C. LATOMBE, A. SARMIENTO und C. TOMASI (1999). *The Autonomous Observer: A Tool for Remote Experimentation in Robotics*. In: *Proc. SPIE Conf. on Telemanipulator and Telepresence Technologies*, Bd. 3840. (to appear).

[GÖTZE et al. 1996] GÖTZE, N., B. MERTSCHING, S. SCHMALZ und S. DRÜE (1996). *Multistage Recognition of Complex Objects with the Active Vision System NAVIS*. In: MERTSCHING, B., Hrsg.: *Aktives Sehen in technischen und biologischen Systemen*, S. 186–193, Sankt Augustin. Infix.

[GÖTZE et al. 1998] GÖTZE, N., R. STEMMER, R. TRAPP, S. DRÜE und G. HARTMANN (1998). *Steuerung eines Demontageroboters mit einem aktiven Stereokamerasystem*. In: *Proc. Workshop Dynamische Perzeption*, Bielefeld.

[GRAF und WECKESSER 1998] GRAF, R. und P. WECKESSER (1998). *Room-service in a hotel*. In: SALICHS, M.A. und A. HALME, Hrsg.: *Proceedings of the Third IFAC Symposium on Intelligent Autonomous Vehicles (IAV 98)*, Bd. 2, S. 641–647.

[GROSS und BÖHME 1997] GROSS, H.-M. und H.-J. BÖHME (1997). *Verhaltensorganisation in interaktiven und lernfähigen Systemen - das MILVA Projekt*. In: GROSS, H.-M., Hrsg.: *Proc. of SOAVE '97 - Selbstorganisation von Adaptivem Verhalten*, S. 1–13, Ilmenau. VDI Verlag.

[GROSSBERG 1976] GROSSBERG, S. (1976). *Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors*. Biological Cybernetics, 23:121–134.

[GROSSBERG 1980] GROSSBERG, S. (1980). *How does a brain build a cognitive code*. Psychological Review, 87:1–51.

[GROSSBERG 1990] GROSSBERG, S. (1990). *A model cortical architecture for the preattentive perception of 3-D form*. In: SCHWARTZ, E. L., Hrsg.: *Computational Neuroscience*, S. 117–138. MIT Press, Cambridge.

[GÖRZ 1993] GÖRZ, G. (HG.), Hrsg. (1993). *Einführung in die künstliche Intelligenz*. Addison-Wesley, Bonn.

[GUYOT et al. 1990] GUYOT, F., F. ALEXANDRE, J.-P. HATON und Y. BURNOD (1990). *A potential powerful connectionist unit: The Cortical Column*. In: FOGELMAN-SOULIE, F. und J. HERAULT, Hrsg.: *Neurocomputing*, Bd. F 68 d. Reihe *NATO ASI Series*, S. 369–377. Springer, Berlin.

[GUZZONI et al. 1997] GUZZONI, D., A. CHEYER, L. JULIA und K. KONOLIGE (1997). *Many Robots Make Short Work*. AI Magazine, 18(1):55–64.

[GVOZDJAK und LI 1998] GVOZDJAK, P. und Z.-N. LI (1998). *From nomad to explorer: active objekt recognition on mobile robots*. Pattern Recognition, 31(6):773–790.

[HERBIN 1996] HERBIN, S. (1996). *Recognizing 3D Objects by Generating Random Actions*. In: *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, S. 35–40, San Francisco.

[HLAVÁČ und ŠÁRA 1995] HLAVÁČ, V. und R. ŠÁRA, Hrsg. (1995). *Computer Analysis of Images and Patterns. 6th International Conference, CAIP'95*, Session Structure and Matching, Prague. Springer.

[HOFMEYER 1998] HOFMEYER, S. (1998). *Mobiler Serviceroboter MORTIMER als Zimmerkellner*. UNIKATH. Universität Karlsruhe (TH), 29(4):12–13.

[HU 1962] HU, M. K. (1962). *Visual Pattern Recognition by Moment Invariants*. IRE Transactions on Information Theory, IT-8:179–187.

[HUBEL 1988] HUBEL, D. (1988). *Eye, Brain and Vision*. Scientific American Library.

[KAWAGUSHI und SETOGUCHI 1994] KAWAGUSHI, T. und T. SETOGUCHI (1994). *3-D Object recognition using Hopfield-Style Neural Networks*. IEICE Trans. Inf. & Syst., E77-D(8):904–917.

[KENDALL 1984] KENDALL, D G. (1984). *Shape manifolds, Procrustean metrics and complex projective spaces*. Bulletin of the London Mathematical Society, 16:81–121.

[KITANO 1998] KITANO, H., Hrsg. (1998). *Robot Soccer World Cup I / RoboCup-97*, Nagoya. Springer.

[KLEIN 1988] KLEIN, R (1988). *Inhibitory tagging system facilities visual search*. Nature, 334:430–431.

[KOLLIAS et al. 1991] KOLLIAS, S., A. TIRAKIS und T. MILIOS (1991). *An Efficient Approach to Invariant Recognition of Images using Higher-Order Neural Networks*. In: KOHONEN, T., K. MÄKISARA, O. SIMULA und J. KANGAS, Hrsg.: *Artificial Neural Networks*, S. 87–92. Elsevier Science.

[KONEN et al. 1994] KONEN, W., T. MAURER und C. VON DER MALSBURG (1994). *A fast link matching algorithm for invariant pattern recognition*. Neural Networks, 7:1019–1030.

[KONOLIGE 1997] KONOLIGE, K (1997). *Small Vision Systems: Hardware and Implementation*. In: *Proc. Eighth International Symposium on Robotics Research*, Hayama.

[LEE et al. 1995] LEE, J.-S., C.-H. CHEN, Y.-N. SUN und G.-S. TSENG (1995). *Occludes Objects Recognition using Multiscale features and Hopfield Neural Networks*. In: BRACCINI, C., L. DEFLORIANI und G. VERNAZZA, Hrsg.: *Image Analysis and Processing*, S. 171–176. Springer.

[LI und NASRABADI 1989] LI, W. und N. M. NASRABADI (1989). *Object Recognition based on Graph Matching implemented by a Hopfield-Style Neural Network*. In: *Proc. of Intern. Joint Conf. of Neural Networks*, Bd. II, S. 287–290, Washington DC.

[LIEDER 1996] LIEDER, T. (1996). *Untersuchungen zur Detektion von Ecken in Grauwertbildern*. Technischer Bericht Universität Hamburg, Fachbereich Informatik, AG IMA.

[LIEDER 1997] LIEDER, T. (1997). *Objekterkennung unter Ausnutzung von Tiefenhinweisen*. Studienarbeit, Universität Hamburg, Fachbereich Informatik, AG IMA.

[LIEDER 1999] LIEDER, T. (1999). *Integration von Tiefen- und Bewegungsschätzung bei der Analyse von Stereobildfolgen*. Diplomarbeit, Universität Hamburg, Fachbereich Informatik, AG IMA.

[LIEDER et al. 1998] LIEDER, T., B. MERTSCHING und S. SCHMALZ (1998). *Using Depth Information For Invariant Object Recognition*. In: POSCH, S. und H. RITTER, Hrsg.: *Dynamische Perzeption*, S. 9–16, St. Augustin. Infix.

[LIN et al. 1991] LIN, W.-C., F.-Y. LIAO, C.-K. TSAO und T. LINGUTLA (1991). *A Hierarchical Multiple-View Approach to Three-Dimensional Object Recognition*. IEEE Transactions on Neural Networks, 2(1):84–92.

[LIU et al. 1998] LIU, T., K. L. CHAN und S. Z. LI (1998). *Model-Based Active Object Recognition using MRF Matching and Sensor Planning*. In: CHIN, R. und T.-C. PONG, Hrsg.: *Computer Vision - ACCV '98. 3rd Asian Conf. on Computer Vision*, Bd. II, S. 225–232, Hong Kong. Springer.

[LOURENS und WÜRTZ 1998] LOURENS, T. und R. P. WÜRTZ (1998). *Object Recognition by Matching Symbolic Edge Graphs*. In: *Proc. of the Asian Conference on Computer Vision ACCV '98*, Bd. II, S. 193–200. Springer.

[LOWE 1986] LOWE, D. G. (1986). *Perceptual organization and visual recognition*. Kluwer, Boston.

[MAGGIONI und WIRTZ 1991] MAGGIONI, C. und B. WIRTZ (1991). *A Neural Net Approach to 3-D Pose Estimation*. In: KOHONEN, T., K. MÄKISARA,

O. SIMULA und J. KANGAS, Hrsg.: *Artificial Neural Networks*, S. 75–80. Elsevier Science.

[VON DER MALSBURG 1981] MALSBURG, C. VON DER (1981). *The correlation theory of brain function*. Technischer Bericht Max-Planck-Institut für Biophysikalische Chemie, Göttingen.

[MARR 1982] MARR, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Freeman, San Francisco.

[MARR und NISHIHARA 1978] MARR, D. und H. K. NISHIHARA (1978). *Representation and recognition of the spatial organization of three-dimensional shapes*. Proceedings of the Royal Society of London, 200:269–294.

[MARTINETZ und SCHULTEN 1991] MARTINETZ, T. M. und K. J. SCHULTEN (1991). *A "neural-gas" network learns topoligies*. In: KOHONEN, T., K. MÄKISARA, O. SIMULA und J. KANGAS, Hrsg.: *Artificial Neural Networks*, S. 397–402. North-Holland, Amsterdam.

[MASSAD 1997] MASSAD, A. (1997). *Entwicklung eines Bildverarbeitungssystems zur Erkennung von dreidimensionalen Objekten anhand von Ansichtensequenzen*. Diplomarbeit, Universität Hamburg, Fachbereich Informatik, AG IMA.

[MASSAD et al. 1998a] MASSAD, A., B. MERTSCHING und S. SCHMALZ (1998a). *Combining Multiple Views and Temporal Associations for 3-D Object Recognition*. In: BURKHARDT, H.: NEUMANN, B., Hrsg.: *Proceedings of the European Comference on Computer Vision ECCV '98*, Bd. 2, S. 699–715, Freiburg. Springer.

[MASSAD et al. 1998b] MASSAD, A., B. MERTSCHING und S. SCHMALZ (1998b). *Utilizing Temporal Associations for view-based 3-D Object Recognition*. In: *Proc. of the 24th Annual Conference of the IEEE Industrial Electronics Society (IECON'98)*, Bd. 4, S. 2074–2078.

[MAURER und DICKMANNS 1997] MAURER, M. und E. DICKMANNS (1997). *A System Architecture for Autonomous Visual Road Vehicle Guidance*. In: *IEEE Conf. on Intelligent Transportation Systems*, Boston.

[MERTSCHING und BOLLMANN 1997] MERTSCHING, B. und M. BOLLMANN (1997). *Visual Attention and Gaze Control for an Active Vision System*. In: KASABOV, N., R. KOZMA, K. KO, R. O'SHEA, G. COGHILL und T. GEDEON,

Hrsg.: *Progress in Connectionist-Based Information Systems*, Bd. 1, S. 76–79, Singapore. Springer.

[MERTSCHING et al. 1999] MERTSCHING, B., M. BOLLMANN, R. HOISCHEN und S. SCHMALZ (1999). *The Neural Active Vision System NAVIS*. In: JÄHNE, B., H. HAUSSECKER und P. GEISSLER, Hrsg.: *Handbook of Computer Vision and Applications*, Bd. 3, S. 543–568. Academic Press, San Diego.

[MERTSCHING et al. 1998] MERTSCHING, B., M. BOLLMANN, A. MASSAD und S. SCHMALZ (1998). *Recognition of Complex Objects with an Active Vision System*. In: *Proc. of Int. ICSC/IFAC Symposium on Neural Computation (NC'98)*, S. 469–475, Wien. ICSC Academic Press.

[MERTSCHING und SCHMALZ 1999] MERTSCHING, B. und S. SCHMALZ (1999). *Active Vision Systems*. In: JÄHNE, B., H. HAUSSECKER und P. GEISSLER, Hrsg.: *Handbook of Computer Vision and Applications*, Bd. 3, S. 197–219. Academic Press, San Diego.

[MORAN und DESIMONE 1985] MORAN, J. und R. DESIMONE (1985). *Selective attention gates visual processing in the extrastriate cortex*. Science, 229:782–784.

[MUSIAL et al. 1998] MUSIAL, M., U. W. BRANDENBURG und G. HOMMEL (1998). *MARVIN - Ein autonom fliegender Erkundungsroboter*. In: *Tagungsband Fachgespräch Autonome Mobile Systeme (AMS'98)*, Karlsruhe. Springer-Verlag.

[NELSON 1995] NELSON, R. C. (1995). *3-D Recognition via 2-stage Associative Memory*. Technischer Bericht TR 565, University of Rochester, Dept. of Computer Science.

[NEUMANN und STIEHL 1992] NEUMANN, H. und H. S. STIEHL (1992). *Emergent segmentation of monocular visual invariants for space perception*. In: *Proc. Int. Joint Conf. on Neural Networks*, Bd. III, S. 266–271.

[NEUMANN und STIEHL 1993] NEUMANN, H. und H. S. STIEHL (1993). *Aktives Sehen: Das neue Paradigma*. In: GÖRZ, G., Hrsg.: *Einführung in die künstliche Intelligenz*, Kap. 6.2.6, S. 652–656. Addison-Wesley, Bonn.

[NOLL 1996] NOLL, D. (1996). *Ein Optimierungsansatz zur Objekterkennung*. Fortschritt-Berichte, Reihe 10. VDI Verlag, Düsseldorf.

[NOLL et al. 1993] NOLL, D., M. SCHWARZINGER und W. V. SEELEN (1993). *Contextual Feature Similarities for Model-Based Object Recognition*. In: *Proceedings of the ICCV'93*, S. 286–290, Berlin.

[NOTON und STARK 1971] NOTON, D. und L. STARK (1971). *Scanpaths in saccadic eye movements while viewing and recognizing patterns*. Vision Research, 11(9):929–942.

[ONISHI et al. 1998] ONISHI, M., M. IYUMI und K. FUKUNAGA (1998). *Object Recognition using sequential Images and Application to Active Vision*. In: AMIN, A., D. DORI, P. PUDIL und H. FREEMANN, Hrsg.: *Advances in Pattern Recognition. Joint IAPR Intern. Workshops SSPR '98 and SPR '98*, S. 366–373, Sydney. Springer.

[OSWALD und LEVI 1997] OSWALD, N. und P. LEVI (1997). *Cooperative Vision in a Multi-Agent Architecture*. In: *Image Analysis and Processing*, LNCS 1310, S. 709–716. Springer.

[PALM und KRAETZSCHMAR 1997] PALM, G. und G. KRAETZSCHMAR (1997). *SFB 527: Integration symbolischer und subsymbolischer Informationsverarbeitung in adaptiven sensomotorischen Systemen*. In: JARKE, M., K. PASEDACH und K. POHL, Hrsg.: *Informatik '97 - Informatik als Innovationsmotor. Proceedings der 27. Jahrestagung der Gesellschaft für Informatik*, Berlin. Springer.

[PAREKH et al. 2000] PAREKH, R., J. YANG, und V. HONAVAR (2000). *Constructive Neural Network Learning Algorithms for Multi-Category Pattern Classification*. IEEE Transactions on Neural Networks. (in press).

[PELILLO und HANCOCK 1997] PELILLO, M. und E. R. HANCOCK, Hrsg. (1997). *Energy minimization Methods in Computer Vision and Pattern Recognition, International Workshop EMMCVPR'97*, Session Structural Models, Venice. Springer.

[PERRETT et al. 1991] PERRETT, D., M. ORAM, M. HARRIES, R. BEVAN, J. HIETANEN, P. BENSON und S. THOMAS (1991). *Viewer-centred and object-centred coding of heads in the macaque tempral cortex*. Experimental Brain Research, 86:159–173.

[PETERSON und SÖDERBERG 1989] PETERSON, C. und B. SÖDERBERG (1989). *A new method for mapping optimization problems onto neural networks*. International Journal of Neural Networks, 1(1).

[PINZ 1994] PINZ, A. (1994). *Bildverstehen*. Springer, Wien.

[PONCE et al. 1996] PONCE, J., A. ZISSERMANN und M. HEBERT, Hrsg. (1996). *Object Representation in Computer Vision II. ECCV'96 International Workshop*, Session Geometric and topological representations, Cambridge. Springer.

[RAO und BALLARD 1995] RAO, R. P. N. und D. H. BALLARD (1995). *An Active Vision Architecture based on Iconic Representations*. Artificial Intelligence, 78:461–505.

[RAY und MAJUMDER 1994] RAY, K. S. und D. D. MAJUMDER (1994). *Application of Hopfield Neural Networks and Canonical Perspectives to Recognize and Locate Partially Occluded 3-D Objects*. Pattern Recognition Letters, 15:815–824.

[RYBAK et al. 1994] RYBAK, I., V. GUSAKOVA, A. GOLOVAN, N. SHEVTSOVA und L. PODLADCHIKOVA (1994). *Modeling of a Neural Network System for Active Visual Perception and Recognition*. In: *Proc. of the 12th IAPR. Conference B: Pattern Recognition and Neural Networks*, Bd. II, S. 371–373, Jerusalem.

[SARLE 1995] SARLE, W. (1995). *Why statisticians should not FART*. ftp://ftp.sas.com/pub/neural/fart.doc.

[SCHAFFER und CHEN 1995] SCHAFFER, M. und T. CHEN (1995). *Object Parts Matching using Hopfield Neural Networks*. In: *Proc. of the Conf. on Computer Architectures for Machine Perception*, S. 438–442.

[SCHMALZ 1999] SCHMALZ, S. (1999). *Aktive Sehsysteme Serviceteil*. KI - Sonderheft Aktive Sehsysteme, (1):43–44.

[SCHMALZ und MERTSCHING 1999] SCHMALZ, S. und B. MERTSCHING (1999). *Object Recognition with Structural Descriptions and Deformable Models*. Neurocomputing. (in press).

[SCHWARTZ 1991] SCHWARTZ, E. (1991). *Personal Commmunication*. In: *Workshop on Active Vision*, University of Chicago.

[SCHWARZINGER et al. 1992] SCHWARZINGER, M., D. NOLL und W. V. SEELEN (1992). *Object Recognition with Deformable Models using Constrained Elastic Nets*. In: FUCHS, S. und R. HOFFMANN, Hrsg.: *Mustererkennung 1992*, S. 96–104, Berlin. Springer.

[SCHWARZINGER et al. 1995] SCHWARZINGER, M., D. NOLL und W. V. SEELEN (1995). *Object Recognition with Constrained Elastic Models*. Mathl. Comput. Modelling, 22(4-7):163–184.

[VON SEELEN et al. 1994] SEELEN, W. VON, S. BOHRER, C. ENGELS, W. GILLNER, H. JANSSEN, H. NEVEN, G. SCHÖNER, W. M. THEIMER und B. VÖLPEL (1994). *Visual Information Processing in Neural Architecture*. In: *Proceedings of the DAGM - Mustererkennung*.

[SHAFER 1985] SHAFER, S. A. (1985). *Using Color to Separate Reflection Components*. COLOR research and application, 10(4):210–218.

[SHAMS 1995] SHAMS, S. (1995). *Multiple Elastic Modules for Visual Pattern Recognition*. Neural Networks, 8(9):1439–1456.

[SHAO und KITTLER 1996] SHAO, Z. und J. KITTLER (1996). *Fuzzy Non-Iterative ARG Labeling with Multiple Interpretations*. In: *Proceedings of the ICPR '96*, S. 181–185.

[SHAPIRO und HARALICK 1981] SHAPIRO, L. G. und R. M. HARALICK (1981). *Structural Descriptions and Inexact Matching*. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-3(5):504–519.

[SHAPIRO et al. 1984] SHAPIRO, L. G., J. D. MORIARTY, R. M. HARALICK und P. G. MULGAONKAR (1984). *Matching Three-Dimensional Objects using a Relational Paradigm*. Pattern Recognition, 17(4):385–405.

[SIPE und CASASENT 1997] SIPE, M. A. und D. CASASENT (1997). *Best viewpoints for active vision classification and pose estimation*. In: *Proc. of the SPIE - The International Society for Optical Engineering*, Bd. 3208, S. 382–393.

[SOMMER 1999] SOMMER, G. (1999). *The global algebraic frame of the perception-action cycle*. In: JÄHNE, B., H. HAUSSECKER und P. GEISSLER, Hrsg.: *Handbook of Computer Vision and Applications*, Bd. 3, S. 221–264. Academic Press, San Diego.

[SWAIN und STRICKER 1991] SWAIN, M. J. und M. STRICKER (1991). *Promising directions in active vision*. Technischer Bericht CS 91-27, University of Chicago. Written by the attendees of the NSF Active Vision Workshop, University of Chicago 1991.

[TANAKA 1996] TANAKA, K. (1996). *Inferotemporal Cortex and Object Vision*. Annual Reviews in Neuroscience, 19:109–139.

[THOMPSON und MUNDY 1987] THOMPSON, D. W. und J. L. MUNDY (1987). *Three dimensional model matching from an unconstrained viewpoint*. In: *Proc. IEEE Int. Conf. Robotics and Automation*, S. 208–220.

[THRUN et al. 1998] THRUN, S., A. BÜCKEN, W. BURGARD, D. FOX, T. FRÖHLINGHAUS, D. HENNIG, T. HOFMANN, M. KRELL und T. SCHMIDT (1998). *Map Learning and High-Speed Navigation in RHINO*. In: KORTENKAMP, D., R. P. BONASSO und R. MURPHY, Hrsg.: *AI-based Mobile Robots: Case studies of successful robot systems*. MIT Press.

[TRAPP 1996] TRAPP, R. (1996). *Entwurf einer Filterbank auf der Basis neurophysiologischer Erkenntnisse zur orientierungs- und frequenzselektiven Dekomposition von Bilddaten*. Technischer BerichtNAVIS 01/96, Heinz-Nixdorf-Institut, Universität-GH-Paderborn.

[TRAPP et al. 1998] TRAPP, R., S. DRÜE und G. HARTMANN (1998). *Stereo Matching with Implizit Detection of Occlusions*. In: BURKHARDT, H. und B. NEUMANN, Hrsg.: *Proc. of the 5th European Conference on Computer Vision (ECCV'98)*, Bd. II, S. 17–33, Freiburg. Springer.

[TRAPP et al. 1995] TRAPP, R., S. DRUEE und B. MERTSCHING (1995). *Korrespondenz in der Stereoskopie bei räumlich verteilten Merkmalsrepraesentationen im Neuronalen-Active-Vision-System NAVIS*. In: SAGERER, G., S. POSCH und F. KUMMERT, Hrsg.: *Mustererkennung 1995*, S. 492–499. Springer.

[TREISMAN 1986] TREISMAN, A. (1986). *Features and Objects in Visual Processing*. Scientific American, 255:114–125.

[ULLMANN 1976] ULLMANN, J. R. (1976). *An Algorithm for Subgraph Isomorphism*. Journal for Computing Machinery, 22(1):31–42.

[UNGERLEIDER und MISHKIN 1982] UNGERLEIDER, L. und M. MISHKIN (1982). *Object vision and spatial vision: Two cortical pathways*. Trends Neurosci., 6:414–417.

[VESTLI und TSCHICHOLD 1996] VESTLI, S. und N. TSCHICHOLD (1996). *MoPS, a System for Mail Distribution in Office-Type Buildings*. Service Robot: An International Journal, 2(2):29–35.

[VETTER et al. 1997] VETTER, V., T. ZIELKE und W. VON SEELEN (1997). *Integrating Face Recognition into Security Systems*. In: BIGÜN, J., G. CHOLLET und G. BORGEFORS, Hrsg.: *Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science 1206, S. 439–448. Springer.

[WATSON 1986] WATSON, G. S. (1986). *The shape of a random sequence of triangles*. Advances in Applied Probability, 18:156–169.

[WILLIAMSON 1996a] WILLIAMSON, J. (1996a). *Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps*. Neural Networks, 9(5):881–897.

[WILLIAMSON 1996b] WILLIAMSON, J. R. (1996b). *Neural Network for Dynamic binding with Graph Representation: Form, Linking, and Depth-from-Occlusion*. Neural Computation, 8:1203–1225.

[WILSON et al. 1996] WILSON, R. C., A. D. J. CROSS und E. R. HANCOCK (1996). *Sensitivity Analysis for Structural Matching*. In: *Proceedings of the ICPR '96*, S. 62–66.

[WILSON und HANCOCK 1995] WILSON, R. C. und E. R. HANCOCK (1995). *An Integrated Approach to Grouping and Matching*. In: BRACCINI, C., L. DE-FLORIANI und G. VERNAZZA, Hrsg.: *Image Analysis and Processing*, S. 62–67. Springer, Berlin.

[WILSON und HANCOCK 1996] WILSON, R. C. und E. R. HANCOCK (1996). *Gauging Relational Consistency and Correcting Structural Errors*. In: *Proc. of CVPR 96. Computer Vision and Image Processing*, S. 47–54.

[WISE und DESIMONE 1988] WISE, S. und R. DESIMONE (1988). *Behavioural neurophysiology: Insigths into seeing and grasping*. Science, 242:736–741.

[WOODHAM 1980] WOODHAM, R. J. (1980). *Photometric Method for Determining Surface Orientations from Multiple Images*. Optical Engineering, 19(1):139–144.

[YANG et al. 1997] YANG, J., R. PAREKH und V. HONAVAR (1997). *DistAl: An Inter-pattern Distance-based Constructive Learning Algorithm*. Technischer Bericht ISU-CS-TR 97-05, Iowa State University.

[YEUNG 1993] YEUNG, D. Y. (1993). *Constructive neural networks as estimators of Bayesian discriminant functions*. Pattern Recognition, 26(1):189–204.

[ZELL 1994] ZELL, A. (1994). *Simulation Neuronaler Netze*. Addison-Wesley, Bonn.

[ZELL et al. 1997a] ZELL, A., S. FEYRER, M. EBNER, A. MOJAEV, O. SCHIMMEL und K. LANGENBACHER (1997a). *Systemarchitektur der autonomen mobilen Roboter Robin und Colin der Universität Tübingen*. In: GROSS, H.-M., Hrsg.: *Proc. of SOAVE '97 - Selbstorganisation von Adaptivem Verhalten*, S. 151–155, Ilmenau. VDI Verlag.

[ZELL et al. 1997b] ZELL, A., J. HIRCHE, D. JUNG und V. SCHEER (1997b). *Erfahrungen mit einer Gruppe mobiler Kleinroboter in einem Roboterpraktikum*. In: *21. Deutsche Jahrestagung für Künstliche Intelligenz*, Workshop 08 Kognitive Robotik, Freiburg.

[ZHANG et al. 1992] ZHANG, S., G. D. SULLIVAN und K. D. BAKER (1992). *Using Automatically Constructed View Independent Relational Model in 3D Object Recognition*. In: SANDINI, G., Hrsg.: *Computer Vision - ECCV '92*.

Proceedings of the 2nd European Conference on Computer Vision, S. 768–786, Santa Margherita Ligure, Italy.

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, diese Arbeit selbständig verfaßt und keine anderen als die angegebenen Quellen benutzt zu haben. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Hamburg, den 17.4.2000

(Steffen Schmalz)

Anhang A

Objektansichten

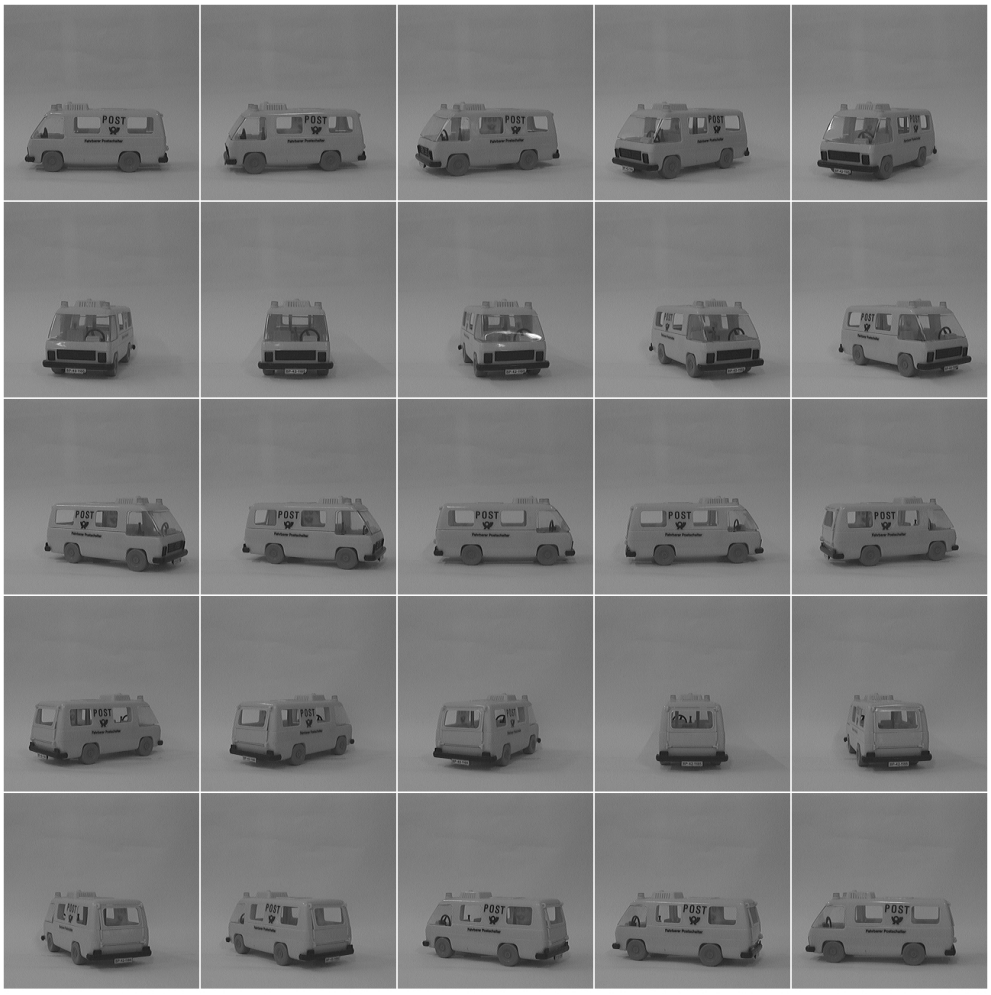


Abbildung A.1: Alle Ansichten des Objekts 0 (Index $0 \ldots 24$)

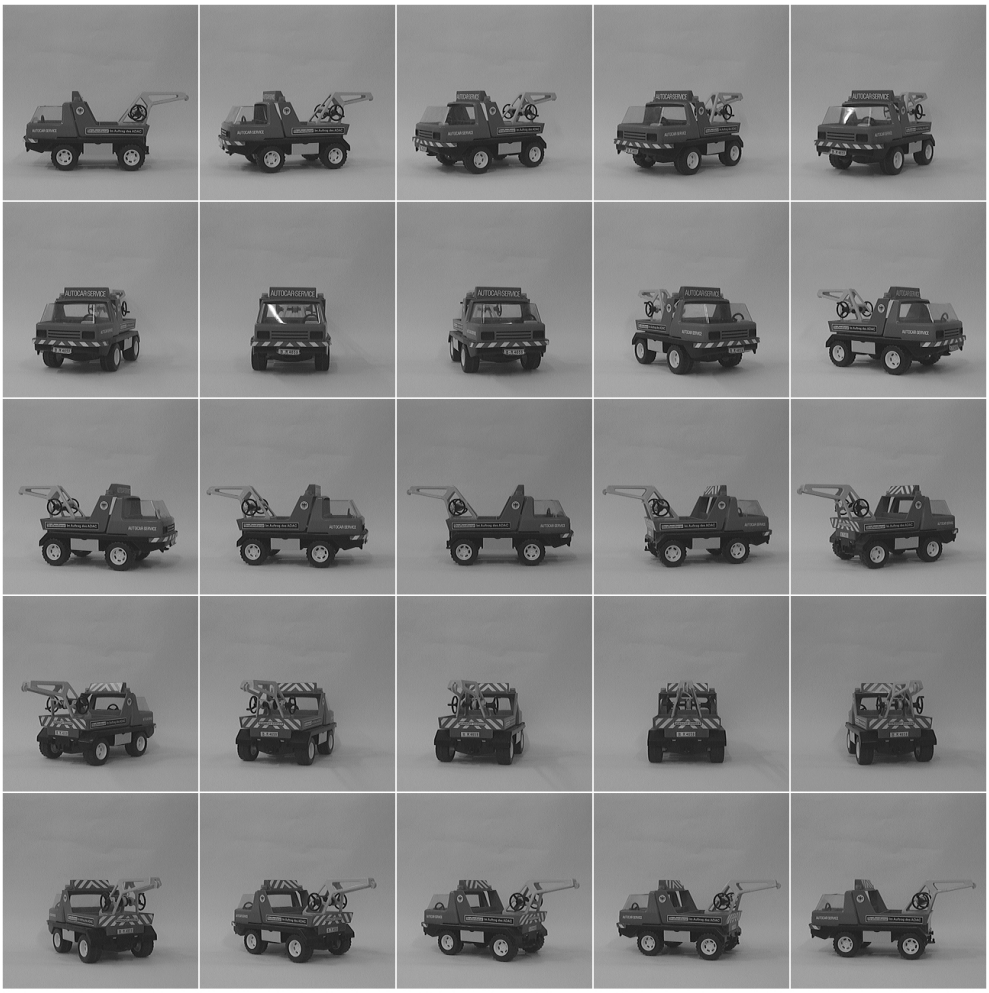


Abbildung A.2: Alle Ansichten des Objekts 1 (Index $0 \ldots 24$)

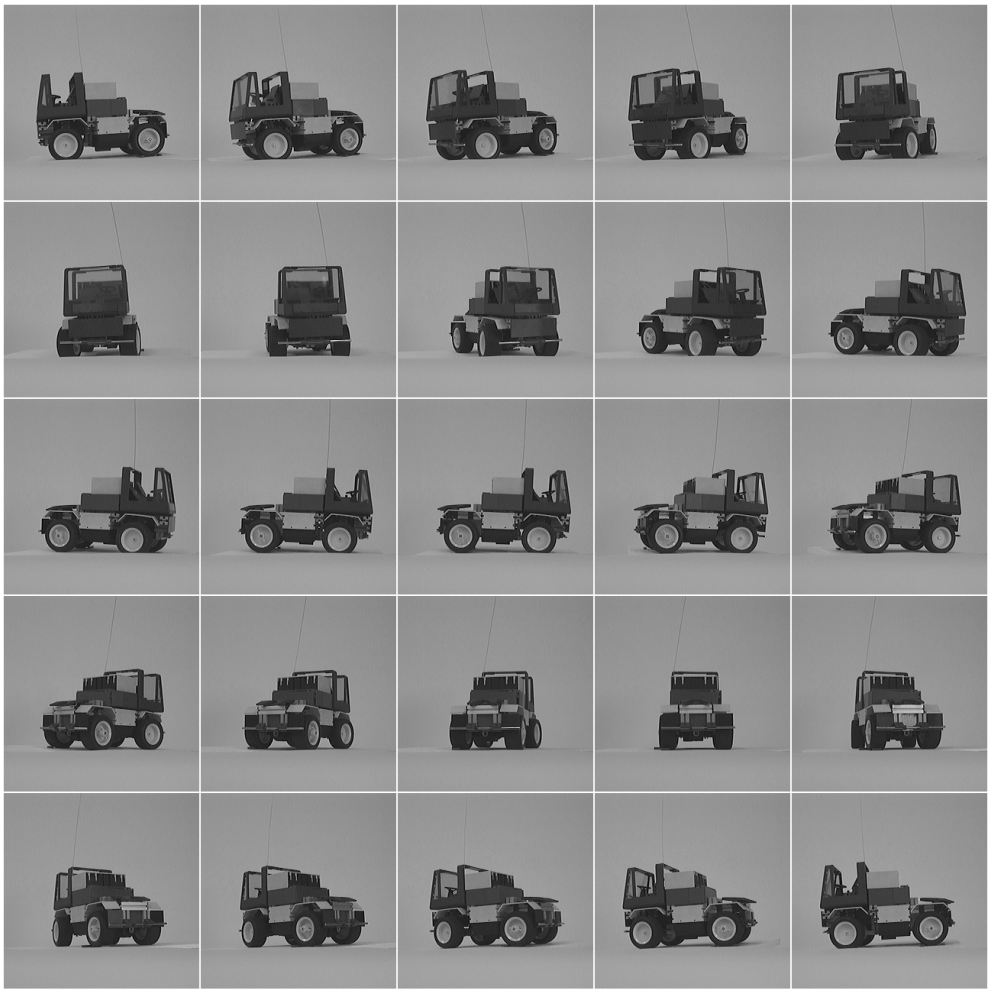


Abbildung A.3: Alle Ansichten des Objekts 2 (Index $0 \ldots 24$)

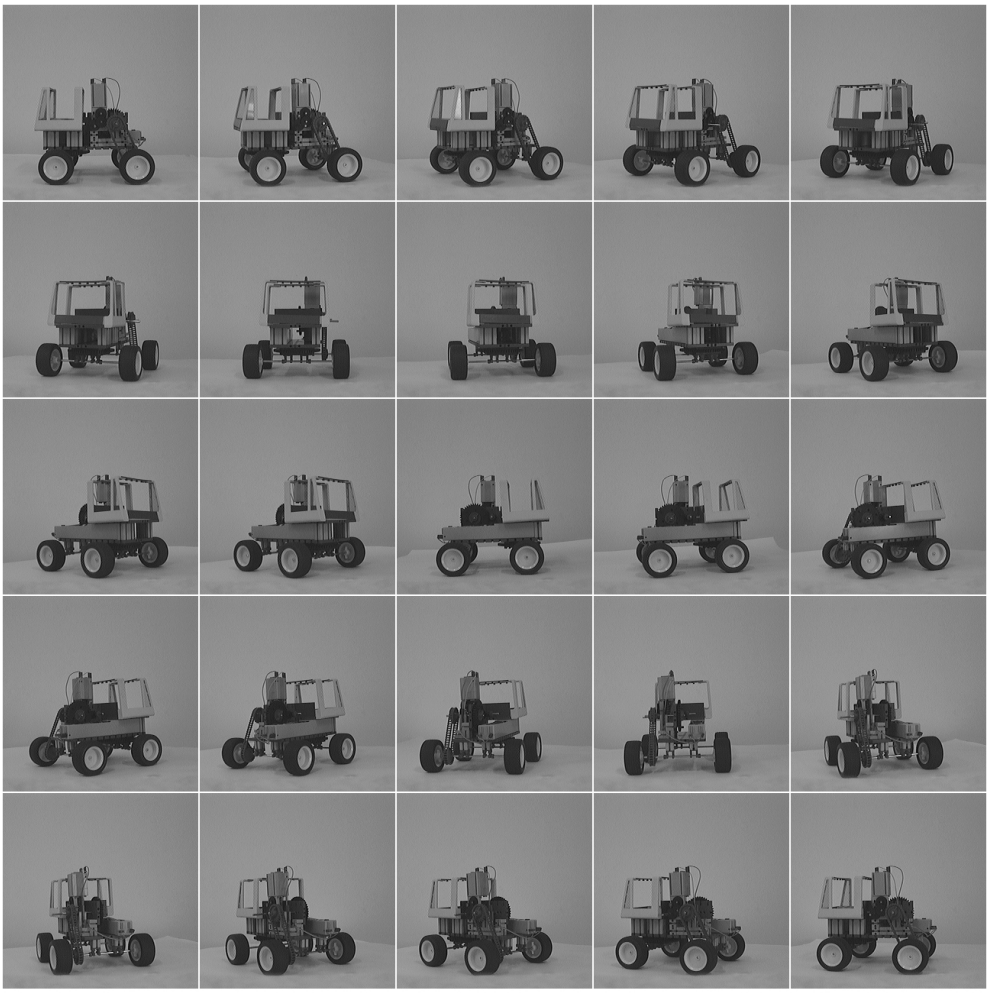


Abbildung A.4: Alle Ansichten des Objekts 3 (Index $0 \ldots 24$)

Lebenslauf

17.01.1970	geboren in Erfurt
1988	Abitur
1988-1989	Wehrdienst
1989-1994	Studium der Informatik an der Technischen Universität Dresden
1992-1994	Beschäftigung (u.a. auch Anfertigung der Studien- und Diplomarbeit) am Fraunhofer-Institut für Mikroelektronische Schaltungen und Systeme Dresden mit dem Schwerpunkt Mustererkennung
1994-1999	Mitarbeiter in der Arbeitsgruppe IMA am Fachbereich Informatik der Universität Hamburg
seit 1999	Beschäftigung in der GAO mbH München als Sensorsoftwareentwickler