Modellierung visueller Aufmerksamkeit im Computer-Sehen: Ein zweistufiges Selektionsmodell für ein Aktives Sehsystem

Dissertation

zur Erlangung des Doktorgrades

am Fachbereich Informatik

der Universität Hamburg

vorgelegt von

Gerriet Backer
aus Emden

Hamburg 2003

Genehmigt vom Fachbereich Informatik der Universität Hamburg
auf Antrag von Prof. Dr. Bärbel Mertsching
und Prof. Dr. Jianwei Zhang

Hamburg, den 06.05.2004

Prof. Dr. Siegfried Stiehl (Dekan)

Kurzfassung

Die visuelle Aufmerksamkeit ist zentraler Bestandteil der menschlichen visuellen Informationsverarbeitung. Sie findet zunehmend auch in der Modellierung von Systemen des Computer-Sehens Verwendung. Aufmerksamkeit ist immer da entscheidend, wo es um die Verteilung von Ressourcen, die Auswahl von relevanten Informationen und die Priorisierung von Aufgaben geht. Die positiven Effekte von Aufmerksamkeit liegen in der effizienteren Verarbeitung visueller Informationen sowie in der Unterdrückung von ablenkenden Informationen. Schließlich spielt Aufmerksamkeit eine wichtige Rolle, wenn es um die Verbindung von Wahrnehmung und Handlung und somit um die Lösung des Bindungsproblems geht.

Konventionelle Modellierungen visueller Aufmerksamkeit zeichnen sich jedoch entweder durch eine Fixierung auf statische zweidimensionale Bilder aus oder zeigen eine stark vereinfachte Modellierung der Selektion. Dabei sind es gerade dreidimensionale dynamische Umgebungen, in denen der Einsatz visueller Aufmerksamkeit den größten Nutzen verspricht.

Als Konsequenz sollten einerseits fortgeschrittene Aspekte der Selektivitätsmodellierung wie objektbasierte Aufmerksamkeit, Integration datengetriebener und modellgetriebener Aufmerksamkeit und dynamische Selektion und Inhibition in einem Modell visueller Aufmerksamkeit berücksichtigt werden. Andererseits sollen die Selektionsmechanismen dahingehend modifiziert werden, dass sie einer dynamischen dreidimensionalen Umgebung gerecht werden. Eine zentrale Herausforderung ist es dabei, in einer dynamischen Umgebung mit einer seriellen attentiven Verarbeitungsstufe ein internes Modell mit den wichtigsten Objekten aktuell zu halten.

Die vorliegende Arbeit stellt ein Modell visueller Aufmerksamkeit für Systeme des Aktiven Sehens vor, das sich durch einen neuartigen Selektionsmechanismus auszeichnet. Seine Relevanz wird sowohl aus Sicht der Effektivität im Computer-Sehen als auch hinsichtlich der Modellierung menschlicher visueller Aufmerksamkeit belegt. Dieser Selektionsmechanismus ist durch die Anwendung einer Aufmerksamkeitssteuerung auf dynamische räumliche Szenen motiviert, die sich durch konventionelle Modellierungen nicht ohne weiteres erreichen lässt. Das Modell ist in der Lage, seltener beachtete experimentelle Daten zum *multiple object tracking* oder der objektbasierten *inhibition of return* zu erklären. Wichtiger jedoch ist die Leistungsfähigkeit als Bestandteil eines Computer Vision-Systems. Sie zeigt sich daran, dass mit wenigen Ressourcen ein Weltmodell der wichtigsten Objekte in dynamischen Szenen bestimmt und aufrecht erhalten werden kann.

Abstract

Visual attention is a substantial aspect of the way humans perceive and process visual information. Increasing use of it is made in models of computer vision. Attention is relevant whenever ressources are to be distributed, relevant information has to be selected, and tasks have to be priorized. The positive effects of attention are due to the more effective processing of visual information as well as the suppression of distracting elements. At last, attention plays an important role in connecting perception and action and thus in solving the binding problem.

Conventional models are mostly either focussed on static two-dimensional images, or are equipped with a strongly simplified selection mechanism. But especially in three-dimensional dynamic environments the use of attention seems most profitable.

As one consequence advanced aspects in modelling visual attenion like object-based attention, the integration of data-driven and model-driven attention, and the dynamic selection and inhibition should be integrated. On the other side, selection mechanisms need to be modified in order to cope with dynamic three-dimensional environments. A central challenge lies in using a serial attentive computations stage in a dynamic environment providing an up-to-date world model of the most relevant objects.

This work introduces a model of visual attention for active vision systems using a novel selection mechanisms. Its quality will be established regarding the effectiveness as a computer vision process as well as regarding the modelling of natural visual attention. The selection mechanism is motivated by applying attentional mechanisms to dynamic spatial scenes that cannot be accomplished by conventional models. The model serves to explain some of the more unregarded experimental data on multiple object tracking and object-based inhibition of return. Even more important is its usefulness as a module of computer vision systems. This usefulness is evident in its ability to compute and update a world model of the most relevant objects in a dynamic scene with most efficient use of ressources.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ADD, ADHD	*Attention deficit (hyperactivity) disorder*; Aufmerksamkeitsstörung
CIE	Comission Internationale De L'Eclairage; Internationales Kommitee und Standardisierungsgremium im Bereich Farbe und Licht [CIE03]
CIELab	s. LAB
CIEXYZ	s. XYZ
DNF	Dynamisches Neuronales Feld, *dynamic neural field*; Netzwerkmodell nach Amari [Ama77]
DoG	*Difference of Gaussians;* Funktion die sich als Differenz zweier Gaußverteilungen ergibt
FINST	*FINgers of INSTantiation*; Theorie des parallelen Zugriffs auf mehrere Items in der visuellen Aufmerksamkeit, eingeführt von Pylyshyn (u.a. [PS88, PBF+94, SP00])
fMRI	*functional magnetic resonance imaging*; funktionelle Magnetresonanztomographie - bildgebendes Verfahren zur Sichtbarmachung von Neuronenaktivität, die von äusseren Reizen abhängt.
FOA	*Focus of attention*; singulärer ausgewählter Bereich, dem bevorzugte attentive Verarbeitung zugeordnet ist
FPGA	*Field Programmable Gate Array*; reprogrammierbare integrierte Schaltung.
HLS	*Hue Light Saturation*; Farbmodell
HSI	*Hue Saturation Intensity;* Farbmodell
KNN	Künstliches Neuronales Netzwerk; Computersimulation der Struktur von Nervenzellen
Lab	Von der CIE definierter Farbraum
MT	Medio-temporaler Kortex; Areal des visuellen Kortex
MTM	Farbraum
NAVIS	*Neural Active VIsion System*; Modell für ein Neuronales Aktives Sehsystem von den Universitäten Hamburg (AG IMA) und Paderborn (GET)
NF	*Neural field*; Kurzform für Dynamisches Neuronales Feld (DNF), Künstliches Neuronales Netzwerk nach Amari [Ama77]
NN	Neuronales Netzwerk; siehe KNN
OF	*Object file*; Symbolische Datenstruktur mit Verweis auf ein Objekt im Bild
PET	Positronen-Emissions-Tomographie; bildgebendes Verfahren
ROI	*Region of interest;* ausgewählter Bildbereich

RSVP	*Rapid serial visual presentation*; Psychophysisches Experimentalparadigma, bei dem es um die schnell aufeinanderfolgende Darbietung mehrerer Reize am selben Ort ankommt
SOA	*Stimulus onset asynchrony*; Zeit, um die zwei Reize versetzt dargeboten werden
UML	*Unified Modeling Language*; Standard zur objektorientierten Modellierung.
VS	*Visual search;* zentrales Paradigma der Experimente zur visuellen Aufmerksamkeit
VSTM	*Visual short term memory*; visuelles Kurzzeitgedächtnis
V1, V2, V3, V4, V5	Bezeichnungen von Arealen des visuellen Kortex
WTA	*Winner-take-all*; Prozess zur Selektion eines einzelnen aus mehreren Kandidaten
XYZ	Von der CIE definierter Farbraum

Kapitel 1

Einleitung

1.1 Einführung und Motivation

Visuelle Aufmerksamkeit beschreibt den Teil der visuellen Verarbeitung, der für die Auswahl von Informationen zur weiteren Verarbeitung zuständig ist. Sie steht zwischen der reinen Aufnahme von Information und der Ausführung von Berechnungsschritten. Ihre Funktionsweise lässt sich als Filter, als Scheinwerfer oder als Verteilung von Ressourcen beschreiben. In der Verarbeitung visueller Information kommt ihr eine zentrale Rolle zu, denn alles, was diese Selektion nicht passiert, erhält keinen Zugang zu Erkennung, Gedächtnis, Aktionen und Bewusstsein. Andererseits sind viele Aufgaben nur schwer zu lösen, wenn die zu bearbeitende Datenbasis nicht zuvor durch einen Aufmerksamkeitsmechanismus begrenzt wurde. Aufmerksamkeit beschränkt den Aufwand für diese komplexen Operationen, indem nur ein Element nach dem anderen ausgewählt wird, anstatt die gesamte zur Verfügung stehende Information gleichzeitig und gleichmäßig zu verarbeiten.

Für Systeme des Computer-Sehens ist die Beschränkung des Aufwandes von besonderer Bedeutung. Die große Datenmenge, die mit visuellen Informationen einhergeht, bedeutet im Zusammenspiel mit komplexen Bildverarbeitungsalgorithmen (z.B. Suche, Vergleich, Selbstorganisation, Faltung, Korrespondenzbildung oder Filterung) und dem häufigen Bedarf nach redundanten Verarbeitungen zur Erhöhung der Robustheit eine Herausforderung für alle aktuellen Rechner und auch für die in den nächsten Jahrzehnten zu erwartenden Systeme. Bedenkt man, dass etwa die Hälfte des menschlichen Gehirns mit der Verarbeitung visueller Informationen befasst ist, wird deutlich, wie komplex und aufwendig diese Aufgabe ist. Hier erscheint die umgangssprachlich als „Eins nach dem anderen und das Wichtigste zuerst"-Heuristik als naheliegende Erleichterung. Sie erlaubt es, Algorithmen unterschiedlicher Komplexität angemessen zu verwenden, indem einfachere, aber damit auch in der Qualität schwächere Verfahren zur Lokalisierung relevanter Bestandteile des verarbeiteten Bildes dienen, während für diese ausgewählten Bereiche sehr viel komplexere Verfahren in Frage kommen. Es wird also zuerst bestimmt, **wo** sich wichtige Elemente befinden, um danach zu erkennen, **was** diese Elemente sind.

Dies ist die Arbeitsweise natürlicher visueller Aufmerksamkeit, wie man sie beim Menschen, aber auch bei vielen Tieren beobachten kann. Seit einigen Jahren wird versucht, die Mechanismen der visuellen Aufmerksamkeit auch im Computer-Sehen umzusetzen. Dabei profitiert man zum Beispiel von der Reduktion der Datenmenge in der Erkennung wichtiger Elemente und kann die gleichen Ergebnisse schneller erzielen, beziehungsweise in derselben Zeit bessere Ergebnisse erhalten. Der

Ansatz, Lösungen aus der Natur in die Technik zu übertragen, findet auf vielen Gebieten statt, ist aber nicht unumstritten. Hier haben wir jedoch einerseits die Situation, dass die technischen Systeme auf dem untersuchten Gebiet nach jahrelanger Forschung und Entwicklung hinsichtlich bestimmter Eigenschaften noch immer unbefriedigend arbeiten. Andererseits gibt es natürliche Vorbilder, die die gewünschten Eigenschaften aufweisen. Somit ist es eine vielversprechende Möglichkeit, Aspekte dieser natürlichen Vorbilder nachzubilden, um einige ihrer Eigenschaften auf technische Systeme zu übertragen. Dies ist der Ansatz der Bionik.

Speziell interessant an dem Bereich der Aufmerksamkeit ist dabei der Zusammenhang von paralleler und serieller Verarbeitung. Auf der einen Seite arbeiten Computer zwar mit hohem Takt, aber immer noch primär seriell und sind damit der Leistung des menschlichen Gehirns, das bei sehr langsamem „Takt" hochgradig parallel arbeitet, weit unterlegen. Auf der anderen Seite zeigt sich gerade in der visuellen Aufmerksamkeit, dass auch das menschliche Gehirn bestimmte Aufgaben seriell abarbeitet, obwohl es prinzipiell in der Lage wäre, viele Dinge gleichzeitig durchzuführen. Um wieviel mehr macht es daher im technischen Bildverstehen, das auf einer seriellen Implementierung beruht, Sinn, schwierige Aufgaben zu serialisieren. Die Umsetzung ist die zentrale Frage der Modellierung visueller Aufmerksamkeit im Computer-Sehen, mit der sich diese Arbeit beschäftigt.

Der Zusammenhang zwischen natürlichem Sehen und Computer-Sehen wird von vielen Forschern betont. Dazu seien hier nur zwei ausgewählte Zitate hervorgehoben:

> „We expect that computer vision research in the future will progress in tight collaboration with many other disciplines that are concerned with empirical approaches to vision, i.e. the understanding of biological vision." (Fermüller und Aloimonos, [FA95]).
>
> „To gain further insights it is necessary to observe and analyse the ways in which the system performs. Thus, the current work suggests a modelling process which adheres explicitly to the principles of operation of natural vision systems as demonstrated by psychophysical experimentation." (Leavers, [Lea94]).

Die Betrachtung von Aufmerksamkeit im Spannungsfeld von Ingenieurs- und Humanwissenschaften ist gerade deshalb angezeigt, weil die visuelle Wahrnehmung, ihre Steuerung und Verknüpfung mit Handlungen immer mehr als bedeutender Bestandteil von Intelligenz angesehen wird. Während früher Intelligenz stark mit Logik, symbolischer Berechnung assoziiert wurde, zeigt sich heute, dass diese letzteren Aspekte im menschlichen Gehirn ein vergleichsweise kleiner Bestandteil sind. Gleichzeitig gelingt es immer mehr, auf früher mit Intelligenz assoziierten Gebieten wie Logik oder Schach Computer mit einer Leistungsfähigkeit zu konstruieren, die der menschlichen zumindest nahe kommt. Im Bereich des Sehens ist dies keineswegs der Fall. Empiriker wie Ingenieure schließen daraus, dass das Geheimnis von Intelligenz und Bewusstsein auch wesentlich in Wahrnehmung und Handlung zu suchen ist. Empirisch ist das Interesse darin begründet, den Menschen als komplexes Wesen verstehen zu können. Gerade Intelligenz und Bewusstsein sind es, die die besondere Rolle des Menschen in der Evolution begründen. Ingenieure hingegen versuchen, Systeme zu konstruieren, die Leistungen des Menschen übernehmen können. Hinsichtlich Mobilität und Manipulation ist dies durch Fahrzeuge und Maschinen weitgehend gelungen, zum Teil werden die Leistungen weit übertroffen. Diese Maschinen sind jedoch weiterhin entweder direkt vom Menschen gesteuert oder auf einfaches Verhalten vorprogrammiert. Die Herausforderung ist es also, intelligentes Verhalten anhand intelligenter

Sensordatenverarbeitung zu ermöglichen und damit autonome mobile Systeme mit Manipulationsmöglichkeiten auszustatten. Ein Baustein dafür ist die Modellierung visueller Aufmerksamkeit als Kontroll- und Steuermechanismus in der visuellen Wahrnehmung und als Filter für die handlungsrelevanten Sensordaten.

Diese Arbeit ist keineswegs der erste Ansatz, visuelle Aufmerksamkeit in technischen Sehsystemen umzusetzen. Vielmehr gibt es in Folge des starken Interesses an der Funktionsweise der Aufmerksamkeit in den Humanwissenschaften eine Anzahl unterschiedlicher Ansätze, die sich hinsichtlich des Anspruches und der Umsetzung stark unterscheiden. Der Ansatz dieser Arbeit soll im folgenden beschrieben werden.

1.2 Aufgabenstellung und Lösungsansatz

Die vorliegende Arbeit hat den Anspruch, ein technisches Modell visueller Aufmerksamkeit zu entwickeln, das als Bestandteil eines aktiven Sehsystems dienen kann. Dieses Modell soll in erster Linie nützlicher Bestandteil eines aktiven Sehsystems sein. Gleichzeitig soll jedoch die Modellierung natürlicher visueller Aufmerksamkeit Bestandteil des Entstehungsprozesses sein, solange es dem Ziel der technischen Verwendbarkeit nicht entgegensteht. Der spezifische Beitrag dieses Modelles soll in der besonderen Beachtung von Dynamik und Tiefe der Umgebung liegen. Es ist zu untersuchen, inwieweit die in der Literatur vorgestellten Modelle zur visuellen Aufmerksamkeit sich direkt auf Szenen, in denen Tiefe und Dynamik eine Rolle spielen, anwenden lassen. Von der Repräsentation der Auffälligkeit bis zur Ausführung der Selektion von Elementen der Umgebung sind die Eigenschaften der Umgebung von großer Bedeutung. Die üblichen Modellierungen konzentrieren sich jedoch auf statische 2D-Bilder. Eine zentrale Frage der Arbeit ist also, welche der bekannten Mechanismen sich auf komplexere Umgebungen übertragen lassen und welche neuen Ansätze gefunden werden müssen. Neben der Integration der Umgebungskomplexität soll das Modell auch Ansätze zur objektbasierten Aufmerksamkeit enthalten, so dass die Selektion Informationen einbezieht, die sich nicht allein auf lokale Bildinformationen, sondern auf ganze Objekte oder Objektkandidaten beziehen. Im Vergleich zum klassischen „Scheinwerfermodell" mit einfachen Bildmerkmalen soll so eine *intelligentere* Selektion anhand komplexerer Bildmerkmale erreicht werden.

Im Rahmen dieser Arbeit wird aufgezeigt, dass ein neues Selektionsmodell benötigt wird, das eine zweite Stufe der Selektion und damit auch eine weitere Berechnungsstufe einführt. Für die Integration von daten- und modellgetriebenen Aspekten in der Steuerung von Aufmerksamkeit ist ein neuer Lösungsansatz zu finden, der zu dieser zweistufigen Selektion passt. Das Ziel sollte dabei sein, modellgetriebene Einflüsse auf einer symbolischen Stufe zu konzentrieren, um eine möglichst einfache Schnittstelle für die Interaktion mit anderen Systemteilen zu erhalten.

Das Modell will auch als Modell menschlicher visueller Aufmerksamkeit ernst genommen werden. Dies gilt vor allem hinsichtlich der Umsetzungen moderner und in bekannten Modellierungen zum Teil ignorierter Konzepte der Aufmerksamkeit als Ergebnis vor allem psychophysischer Untersuchungen. Eine Computerimplementierung von Modellen menschlicher Verarbeitung hat dabei immer den Vorteil, gezwungenermaßen konkret zu sein, um so Verifikationen zuzulassen.

Schließlich ist die Evaluation eines solchen Modelles zu untersuchen. Diese hat die unterschiedlichen Ziele (Modellierung des natürlichen Vorbildes und technische Leistungsfähigkeit) im Auge

zu behalten. Problematisch ist, dass man keine**korrekte** Zuweisung von Aufmerksamkeit definieren kann. Es ist somit nicht möglich, den Fehler oder Abstand zu einem solchen Ideal als Gütemass anzusetzen. Die Prüfung der verbleibenden Evaluationsmöglichkeiten wird die Entwicklung eines Simulationsrahmenwerks nahelegen, das zur Evaluation eingesetzt wird.

Ein wesentlicher Teil der Arbeit entstand im Rahmen des DFG-geförderten ESAB-II-Projektes (Entwicklung von Systembausteinen der Aktiven Bildanalyse II) in der AG IMA am FB Informatik der Universität Hamburg. Einige Vorarbeiten beruhen auf Untersuchungen im Rahmen des DFG-geförderten Projektes „Der Einfluss von Aufmerksamkeit und dem Ortsfrequenzgehalt der Reize auf die sensorische Gruppierung" im FB Psychologie der TU Braunschweig[1].

1.3 Beitrag der Arbeit

Diese Arbeit stellt ein Modell zur visuellen Aufmerksamkeit vor, das in starker Korrespondenz zum menschlichen Vorbild arbeitet. Dadurch ist eine explizite technische Modellierung sonst eher vernachlässigter Aspekte der natürlichen visuellen Aufmerksamkeit und eine Rückkopplung in die empirische Untersuchung von Aufmerksamkeit möglich. Zu den Aspekten, die dieses Modell im Gegensatz zu bekannten Modellen besonders hervorhebt, gehören die über ein statisches zweidimensionales Bild hinausgehenden Dimensionen: Tiefe und Dynamik. Es konnte aufgezeigt werden, dass sich die bestehenden Modelle nicht unmittelbar erweitern lassen, sondern tiefgreifende Veränderungen vorgenommen werden müssen. Diese resultieren in einer neuartigen Selektionsstruktur, die die Auswahl eines einzelnen Fokus der Aufmerksamkeit aus dem Bild in zwei Selektionsstufen aufteilt. Die erste Stufe ist dabei für die subsymbolische Selektion und Verfolgung einiger weniger Regionen hoher Auffälligkeit zuständig. Die zweite Stufe selektiert anschließend auf symbolische Weise eine dieser Regionen als klassischen Fokus der Aufmerksamkeit.

Zur technischen Modellierung der ersten Selektionsstufe gehören Mechanismen, die die Bestimmung lokaler Auffälligkeit oder Salienz erlauben. Eines der Ziele ist es, Bereichen Auffälligkeit zuzuordnen, die potenziell mit visuellen Objekten korrespondieren. Somit geht die präattentive Verarbeitung über klassische Filtermodelle hinaus. Dabei ist zu beachten, dass die präattentive Verarbeitung naturgemäß keine allzu aufwendigen Verfahren enthalten darf. Die zweistufige Selektion bringt eine zusätzliche semiattentive Berechnungsstufe mit sich. Sie wird Mechanismen beinhalten, die nicht in die klassische Dichotomie von präattentiver paralleler und attentiver serieller Verarbeitung passen. Dazu gehört die Verfolgung einer kleinen Anzahl von Elementen. Diese erlaubt die objektbasierte Verarbeitung in dynamischen Szenen zu verbessern und unter anderem eine objektbasierte Inhibition vor kurzem selektierter Objekte zu erreichen.

Die Aufteilung der Selektion in zwei Stufen sowie die Integration von Verfolgungsmechanismen und einer dreidimensionalen Salienzrepräsentation bewähren sich mehrfach. Sie erlauben eine bessere Bindung der Information an die Objekte, eine reichhaltigere Grundlage zur Auswahl der Objekte und schließlich eine bessere Möglichkeit zur Integration datengetriebener Aufmerksamkeit mit unterschiedlichen Verhalten, die modellgetrieben vorgegeben werden können. Das Aufmerksamkeitsmodell integriert also nicht nur modellgetriebene und datengetriebene Aspekte der Aufmerksamkeit sondern gleichzeitig eine raumbasierte und objektbasierte Selektion. Damit vereinfacht und verbessert sich

[1]Der Deutschen Forschungsgemeinschaft (DFG) sei für die Förderung im Rahmen der Projekte Me1289/3-2 und Hu532/5-1 gedankt.

die mögliche Integration dieses Modells visueller Aufmerksamkeit in Aktive Sehsysteme. Somit stellt das Modell einen signifikanten Fortschritt in der Modellierung visueller Aufmerksamkeit dar.

1.4 Gliederung der Arbeit

Als Grundlagen der Arbeit werden die allgemeine visuelle Wahrnehmung (Kap. 2) und daran anschließend genauer die visuelle Aufmerksamkeit und ihre Modellierung (Kap. 3 und 4) beschrieben. Für beide Gebiete wurde eine Aufteilung in die jeweils natürliche und technische Ausprägung vorgenommen, um dem Vorwissen des Lesers auf den entsprechenden Gebieten gerecht zu werden. Die visuelle Wahrnehmung ist ein derart komplexes und umfangreiches Gebiet, dass hier nur kurze einführende Informationen zu den relevanten Themenbereichen dargestellt werden können.

Das im Rahmen dieser Arbeit entwickelte Modell visueller Aufmerksamkeit wird dann in der Reihenfolge der Verarbeitungsschritte beschrieben. Beginnend beim Bild findet eine datengetriebene Berechnung von Salienzinformationen, d.h. von der lokalen Auffälligkeit der Bildbereiche, parallel für das ganze Bild statt. Diese Berechnung der Salienz getrennt anhand verschiedener Merkmale und deren Integration in verschiedene Formen der Salienzrepräsentation wird in Kapitel 5 dargestellt. Dieser Schritt wird klassischerweise als präattentive Berechnung bezeichnet, also als Berechnung, die keine Zuweisung von Aufmerksamkeit voraussetzt. Darauf folgt die erste, subsymbolische Selektionsstufe zur Auswahl einer kleinen Anzahl von Elementen. Die Stufe verwendet für die Selektion, aber auch für Aufgaben der hier semiattentiv genannten Berechnungsstufe sogenannte Dynamische Neuronaler Felder (Kap. 6). Zu den bedeutenden Aufgaben der semiattentiven Stufe gehört die modellfreie Verfolgung mehrerer Elemente. Es werden dabei unterschiedliche Architekturen dieser Felder diskutiert, die auch auf die unterschiedlichen Salienzrepräsentationen eingehen. Darauf folgt die zweite symbolische Selektionsstufe zur Auswahl eines einzelnen Fokus der Aufmerksamkeit (Kap. 7). Diese Selektionsstufe wird durch Verhaltensmodelle gesteuert, die für unterschiedliche Aufgaben gedacht sind. Der Fokus der Aufmerksamkeit enthält ein einzelnes Element, auf das komplexe Operationen wie Objekterkennung angewandt werden können. Diese Operationen werden der attentiven Verarbeitung zugeordnet, sie setzen also die Zuweisung von Aufmerksamkeit voraus. Die Verhaltensmodelle sind auch für die Auslösung von Blickbewegungen zuständig und werden in Kap. 8 beschrieben. Abb. 1.1 gibt die Einordnung dieser Bestandteile in die Systemarchitektur wieder.

Es schließt sich in Kapitel 9 eine Diskussion der Möglichkeiten zur Evaluation von Aufmerksamkeit an. Sie hat zur Entwicklung und Verwendung eines Simulationsrahmenwerks geführt, das die Umsetzung einer Vielzahl der vorgestellten Experimente ermöglichte. Daneben belegen allgemeine Analysen der Leistungsfähigkeit und Vergleiche zur natürlichen Aufmerksamkeit die Eigenschaften des Modells. Eine zusammenfassende Bewertung und ein Ausblick auf künftige Entwicklungen in Kap. 10 schließen die Arbeit ab.

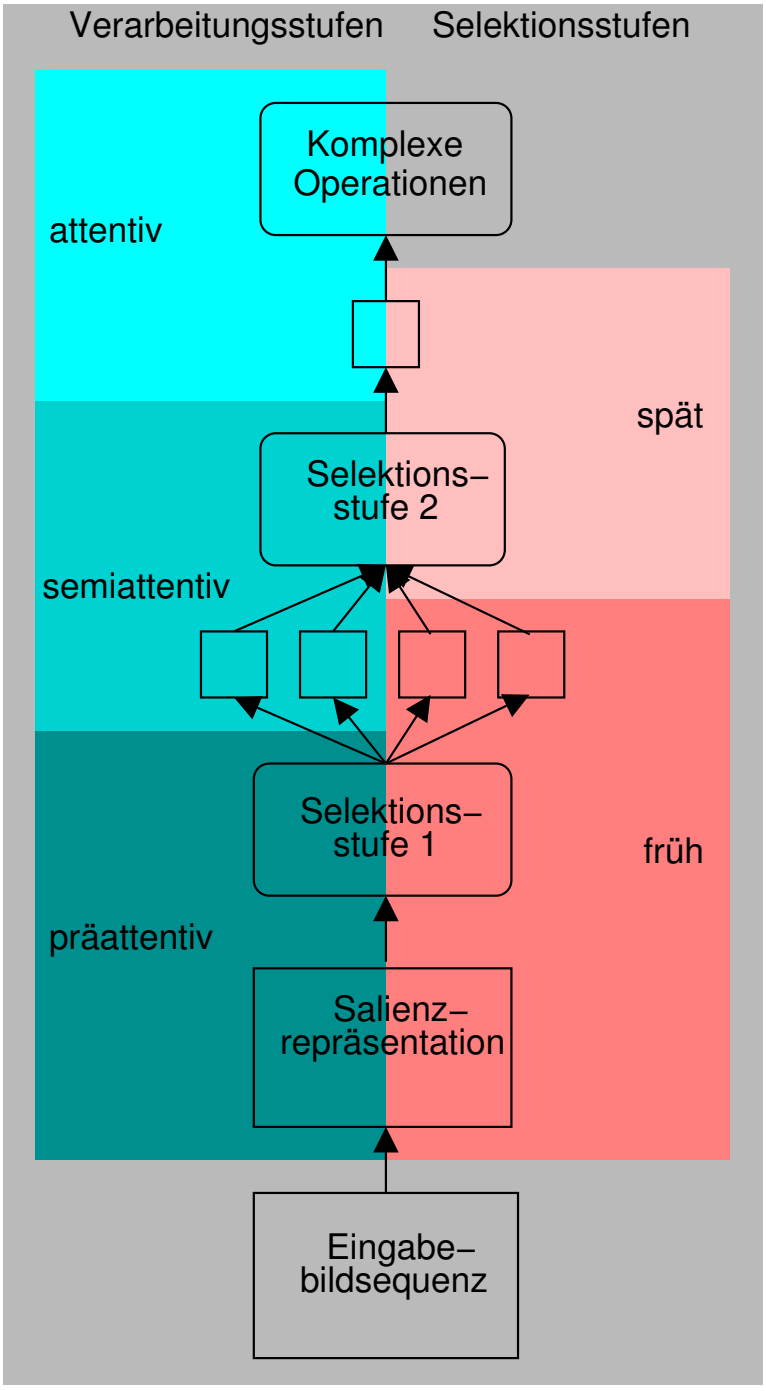


Abbildung 1.1: Architektur der Aufmerksamkeitssteuerung mit einer Einteilung in Verarbeitungs- und Selektionsstufen.

Teil I

Status quo

Kapitel 2

Visuelle Wahrnehmung

Die visuelle Wahrnehmung umfasst Aspekte, die vom physikalischen Prozess der Bildentstehung bis hin zum kognitiven Prozess des Bildverstehens reichen. Obgleich das Sehen für den Menschen den wichtigsten Sinn darstellt, der unser *Bild der Welt* bestimmt, sind die Systeme zum technischen Bildverstehen trotz erheblicher Forschungsanstrengungen vergleichsweise simpel und auf stark eingeschränkte Domänen begrenzt. Wie komplex die Aufgabenstellung des Erkennens und Verstehens anhand von Bildern der Umgebung ist, wird deutlich, wenn man sich vor Augen hält, dass der größte Teil des Kortex eben dieser Aufgabe dient. Um zu analysieren, inwieweit Aufmerksamkeit eine Rolle in der Lösung dieser Aufgabe spielt, ist zu klären, welcher Art die Aufgabe ist, deren Lösung durch Verwendung von attentiven Verfahren vereinfacht werden soll.

Da die Modellierung von Aufmerksamkeit sowohl für Ingenieure als auch für Humanwissenschaftler von Bedeutung ist, werden im folgenden kurz Aspekte der visuellen Wahrnehmung beim Menschen und beim Computer beschrieben. Die grundlegenden Mechanismen und Modelle werden benannt. Die Darstellung beschränkt sich dabei auf Aspekte der Wahrnehmung mit Bezug zum vorgestellten Aufmerksamkeitsmodell. Für weitergehende Betrachtungen wird jeweils auf entsprechende Übersichtsliteratur verwiesen.

2.1 Menschliche visuelle Wahrnehmung

2.1.1 Einführung

Was wir als Bild unserer Umgebung empfinden, ist die Reflektion von Licht an Oberflächen, das über die Linse auf unsere Retina fällt und dort entsprechend der Energie entlang bestimmter Wellenlängen Rezeptoren unterschiedlicher Typen anregt. Die Rezeptoren signalisieren Informationen über diese Anregung an unterschiedliche Regionen im Gehirn, die für die Verarbeitung visueller Informationen zuständig sind.

Abb. 2.1 gibt die Anatomie des Auges wieder, die den Weg des Lichtes bis zu den Sehnerven bestimmt. Einfallendes Licht wird zuerst an der sogenannten Cornea gebrochen, muss dann die Iris passieren, die je nach Helligkeit die Menge des Lichtes im Verhältnis von 1 zu 10^4 regulieren kann. Das Licht wird dann noch einmal in der Linse gebrochen, wobei Muskeln die Form der Linse so verändern können, dass unterschiedliche Entfernungen scharf abgebildet werden. Dieser Vorgang wird als Akkomodation bezeichnet. Das Licht fällt nun auf die Retina, die aus einer hohen Anzahl von

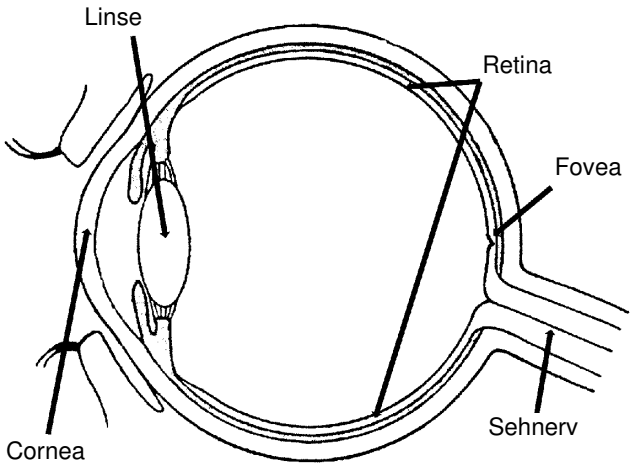


Abbildung 2.1: Anatomie des menschlichen Auges.

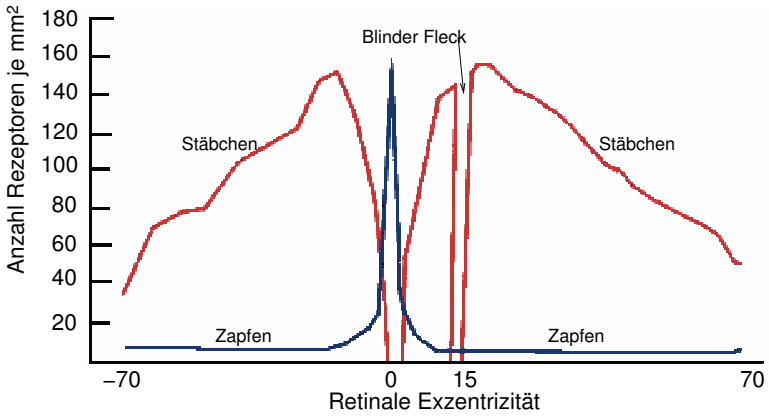


Abbildung 2.2: Die Verteilung von Zapfen und Stäbchen auf der Retina in Abhängigkeit des Abstandes von der Fovea (nach Osterberg [Ost35]).

Rezeptoren besteht. Diese teilen sich in etwa 6 bis 7 Millionen Zapfen und mehr als 100 Millionen Stäbchen auf. Während die Stäbchen nur Helligkeitsunterschiede auswerten, sind die Zapfen für das Farbsehen zuständig. Die Rezeptoren sind nun keineswegs gleichmäßig auf der Retina verteilt. Vielmehr gibt es einen kleinen Bereich scharfen Sehens, die sogenannte *Fovea centralis*. In diesem Bereich, der ca. 5° des Sichtfeldes abbildet, finden sich keine Stäbchen, die Dichte der Zapfen ist hier jedoch am höchsten. Sie nimmt nach außen hin immer weiter ab, man spricht von einer ortsvarianten Auflösung. Die Dichte der Stäbchen hingegen wächst außerhalb der Fovea, bis sie bei etwa 20° ihr Maximum erreicht und von da an wieder abnimmt (s. Abb. 2.2).

Die ortsvariante Auflösung der Retina wird in der Technik meist durch eine logarithmisch-polare Koordinatentransformation (siehe [BL98]) modelliert. Sie macht Bewegungen der Augen notwendig, die dafür sorgen, dass der interessante Bildbereich auf die Fovea abgebildet wird. Die Bewegungen teilen sich im wesentlichen in langsame Folgebewegungen und schnelle Blicksprünge, die als Sakkaden bezeichnet werden.

Die Dynamik der Rezeptoren selbst ist mit 1 zu 100 recht eingeschränkt. Zusammen mit der Pupillenadaption durch die Iris erhält man aber einen Dynamikbereich von 1 zu 10^6. Die Kon-

trastempfindung ist für Farbe und Grauwerte erwartungsgemäß unterschiedlich: während Farbe als Tiefpass wirkt, so dass nur Ortsfrequenzen unterhalb einer gewissen Grenze erkennbar sind, zeigt sich die Grauwertwahrnehmung als Bandpass mit durchweg höheren Frequenzen als im Farbbereich. Das bedeutet, dass für feine Strukturen die Helligkeitsunterschiede von Bedeutung sind. Langsame Helligkeitsveränderungen, wie sie durch Beleuchtungsvariation entstehen, können schlecht wahrgenommen werden. Zeitlich verhält sich das System wie ein Tiefpass. Ab etwa 15 Bildern pro Sekunde werden die Bilder nicht mehr einzeln für sich wahrgenommen, sondern als kontinuierliche Veränderung angesehen. Die Arbeitsweise von Zapfen und Stäbchen unterscheidet sich auch anhand der Lichtstärke; Farbsehen operiert im Bereich hoher Lichtstärken, während die Stäbchen auch mit relativ wenig Licht auskommen - daher sind „des Nachts alle Katzen grau".

Auf der Retina werden die Informationen über drei Schichten von Interneuronen an sogenannte Ganglienzellen weitergegeben. Deren Axone verlassen im nasalen Bereich, etwa 15 Grad von der Fovea entfernt, das Auge und bilden den Sehnerv. An dieser Stelle liegt der „blinde Fleck", an dem keine Rezeptoren vorhanden sind. Dass er im täglichen Leben nicht auffällt, liegt einerseits an der Redundanz der beiden Augen, andererseits an einem Mechanismus, der zum Auffüllen dieses Bereiches anhand der umliegenden Sensorinformationen dient.

Es gibt nur etwa 1,5 Millionen Ganglienzellen, die Informationen an das Gehirn weitermelden können, also eine - im Vergleich mit technischen Systemen - geringe Auflösung. Dabei ist jedoch zu beachten, dass die Information bereits einer gewissen Verarbeitung unterworfen wurde. Ganglienzellen integrieren Informationen von mehreren Zapfen bzw. Stäbchen. Dabei führen sie jedoch keine einfache Summation durch und reagieren damit proportional zur Menge des eingefallen Lichtes. Vielmehr kann man sie einteilen in sogenannte *On-Center*-Zellen und *Off-Center*-Zellen. Die *On-Center*-Zellen reagieren vor allem dann, wenn die Lichtintensität im Zentrum der verbundenen Rezeptoren höher ist als im Umfeld. Bei den *Off-Center*-Zellen verhält es sich gerade umgekehrt. Man erkennt also die Kontrastbildung als einen wichtigen Schritt in dieser frühen Verarbeitung. Die Dichte der Ganglienzellen nimmt kontinuierlich von der Fovea nach außen hin ab.

Beschreibt man die Eigenschaften der verschiedenen Zelltypen (Neurone) ist der Begriff des rezeptiven Feldes entscheidend. Er bezeichnet denjenigen Bereich, in dem Veränderungen eine Veränderung der Feuerrate dieser Zelle bewirken können. Während diese rezeptiven Felder in den frühen visuellen Arealen recht klein sind, nimmt ihre Größe im Laufe der Verarbeitung weiter zu. Die Sehnerven beider Augen überkreuzen sich im sogenannten Chiasma. Dabei führen die für den rechten Teil des visuellen Feldes zuständigen Strukturen beider Augen in die linke Hirnhälfte, die für den linken Teil zuständigen Strukturen analog in die rechte Hirnhälfte. Die Ganglienzellen lassen sich in zwei Typen unterscheiden: die M-Zellen und die P-Zellen. Sie sind in den Kniehöckern (lateral geniculatus nuclei, LGN) entsprechend mit dem magnozellulären und dem parvozellulären System verbunden. Die meisten Zellen (etwa 80 %) sind P-Zellen mit kleinen rezeptiven Feldern. Sie sind für das Farbensehen und die hohe Auflösung zuständig, während die M-Zellen im Vergleich sehr viel schneller reagieren. Schließlich erreichen die Sehbahnen den visuellen Kortex. Man findet hier sogenannte retinotope Karten, d.h. neuronale Strukturen, bei denen die räumliche Anordnung der Neuronen zum Bereich der Retina korrespondiert, von dem sie Informationen erhalten.

Folgende weiterführende Werke seien für einen umfangreicheren Einstieg in das Thema empfohlen. Wandell [Wan95] beschreibt die Leistungen des Sehsystems systematisch durch Anwendung der

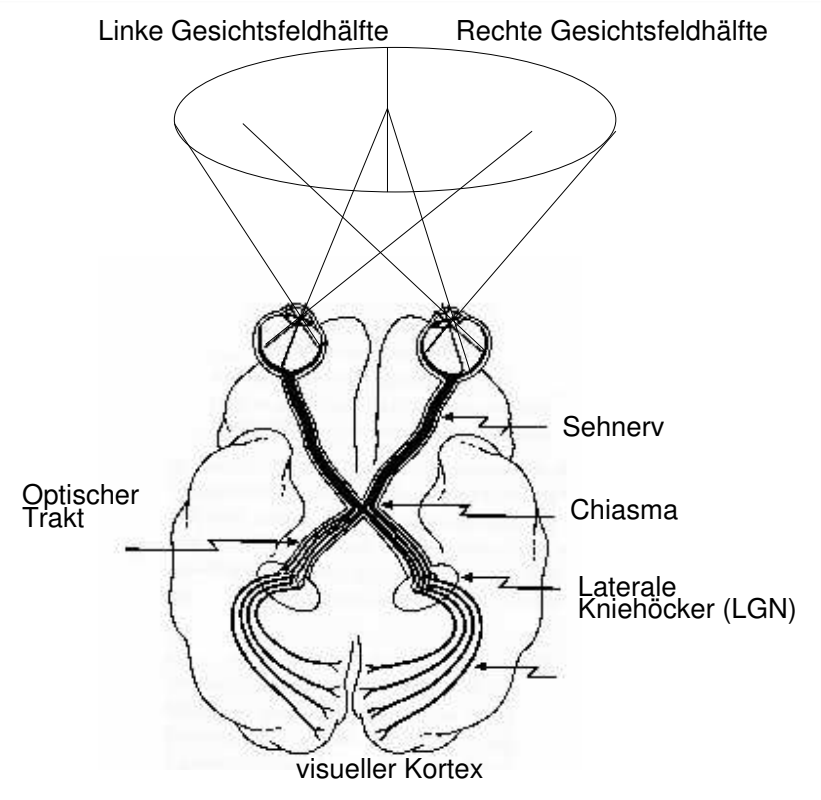


Abbildung 2.3: Vom Objekt zur visuellen Verarbeitung (nach Owen [ODR+99]).

linearen Systemtheorie. Rock [Roc98] bietet einen sehr anschaulichen Einstieg in das Thema, geht dabei funktional von der Retina aus vor. Mallot [Mal98] untersucht die frühe visuelle Verarbeitung als informationsverarbeitenden Prozess im Sinne der „Computational Intelligence" und versucht, gleichzeitig technische und natürliche Bildverarbeitung zu erläutern. Der neuronalen Grundlage widmen sich allgemein Kolb und Whishaw [KW96], spezieller auf die visuelle Wahrnehmung bezieht sich Tovee [Tov96]. Einen sehr breiten Ansatz, der frühe Verarbeitung ebenso wie stärker interpretative oder handlungssteuernde Aspekte beachtet, bieten Bruce et al. [BGG96].

2.1.2 Farbwahrnehmung

Die Zapfen erlauben das Farbensehen, weil sich drei Typen mit unterschiedlicher Sensitivität für die verschiedenen Frequenzen des Lichts finden. Abb. 2.4 zeigt die relative Absorption in Abhängigkeit von der Wellenlänge. Die Zentren der Empfindlichkeit liegen für die drei Typen bei den wahrgenommenen Farben violett (419 nm), grün (531 nm) und gelb (558 nm). Jeder Rezeptor integriert die Menge der von ihm absorbierten Energie, was zur Konsequenz hat, dass viele unterschiedliche Spektralverteilungen zur selben Farbwahrnehmung führen. Der Raum aller möglichen Empfindungen ist auf maximal drei Dimensionen reduziert, ein Wert der auch experimentell bestätigt wurde [SP75].

Die relativen Reaktionen dieser Photosensoren überführen beliebige Frequenzverteilungen in einen wahrgenommenen dreidimensionalen Farbraum. Als Konsequenz daraus ergibt sich, dass sich dieselbe Farbwahrnehmung auf unterschiedliche Spektralverteilungen zurückführen lässt, der Effekt der *Metamerie*.

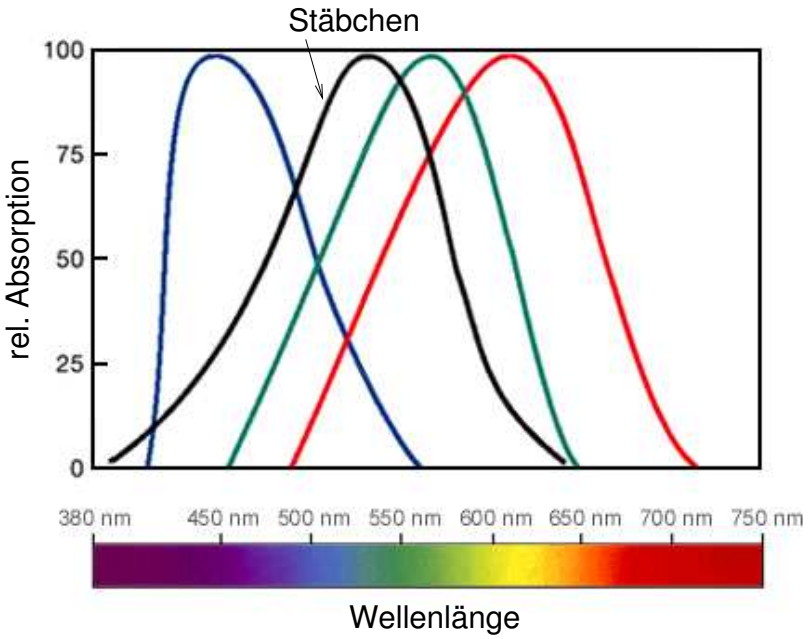


Abbildung 2.4: Empfindlichkeit der Rezeptoren (Stäbchen sowie drei Typen von Zapfen) für Licht unterschiedlicher Wellenlänge [Ask03].

Für die menschliche Wahrnehmung sind auch noch die Komplementärfarben von Bedeutung. Während einige Farben koexistieren können und man sich Mischungen aus ihnen vorstellen kann, wie z.B. ein grünliches Blau oder ein gelbliches Rot, geht das für andere Farbpaare nicht. Weder rot und grün noch blau und gelb können koexistieren. Dass dies so ist, lässt sich nicht direkt aus den Rezeptortypen ableiten, sondern gibt einen Zusammenhang in der Folgeverarbeitung von Farben wieder. Bereits in den Ganglienzellen werden Differenzen oder Summen der Reaktionen unterschiedlicher Zapfentypen berechnet, die für eine Dekorrelation sorgen [BG83]. Es findet eine Weiterverarbeitung in drei Kanälen statt:

- Der achromatische Kanal aus Zellen des magnozellulären Typs als Summe der Reaktionen. Er weist den bei weitem größten Informationsgehalt auf.
- Der Rot-Grün-Kanal aus parvozellulären Strukturen enthält deutlich weniger Informationen.
- Der parvozelluläre Blau-Gelb-Kanal hat den geringsten Informationsgehalt.

Die Wahrnehmung von Farbe stellt insgesamt einen Interferenzprozess dar. Zuerst findet anhand der auf drei Rezeptortypen reduzierten Information die Schätzung einer Wellenlängenverteilung statt. Daraus muss die Reflexionseigenschaft eines Objektes erschlossen werden, welches das Licht einer nicht bekannten Wellenlängenverteilung reflektiert. Wie bei vielen anderen Problemen der visuellen Wahrnehmung handelt es sich um ein schlecht gestelltes Problem, d.h. es stehen prinzipiell nicht genug Informationen zur Verfügung, um die Aufgabe eindeutig zu lösen. Dass der Mensch solche Aufgaben tatsächlich nicht lösen kann, spiegelt sich in den optischen Täuschungen wieder. Jedoch gelingt es, in vielen „natürlichen" Situationen gute Lösungen mit Hilfe von Heuristiken zu finden.

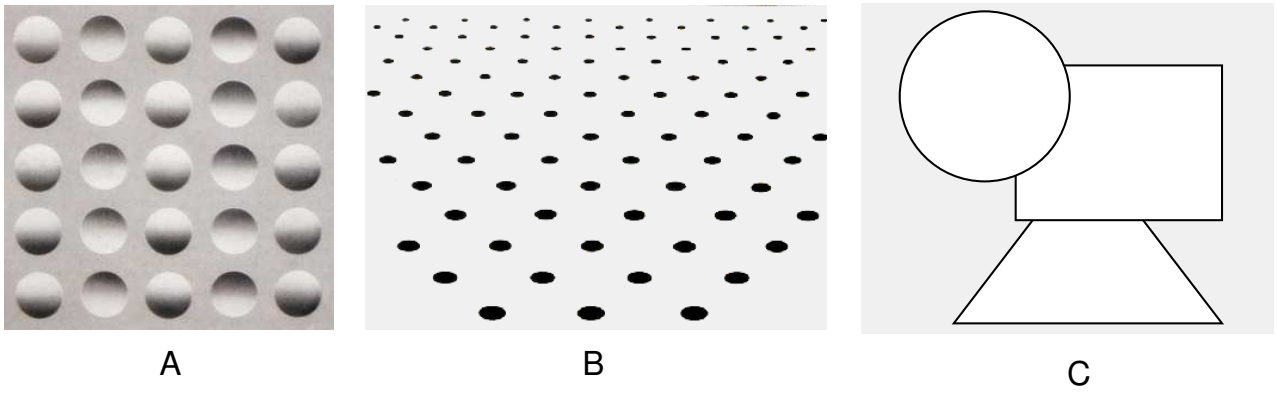


Abbildung 2.5: Beispiele für monokulare Hinweise auf Tiefeninformation. A) Die Schattierung legt eine Interpretation der Kreise als Erhebungen und Löcher nahe - konsistent mit der Annahme von oben einfallenden Lichts. B) Die kleiner werdenden Ellipsen deuten die Perspektive einer sich entfernenden Ebene an. C) Die Verdeckung der einfachen Formen (als die sie interpretiert werden) sorgt dafür, dass der Kreis vor dem Rechteck und dieses wiederum vor dem Dreieck wahrgenommen werden.

2.1.3 Tiefenwahrnehmung

Das Problem der Rekonstruktion von Tiefeninformationen aus den zweidimensionalen Abbildern auf der Retina geschieht durch die Auswertung unterschiedlicher Hinweise. In der Kunst werden ganz unterschiedliche Hinweise wie partielle Verdeckung von Bildelementen, die bekannte Größe von Objekten, Perspektive oder Schatten bewusst eingesetzt, um in zweidimensionalen Bilden einen Tiefeneindruck zu erzeugen[1]. Einige Beispiele dazu finden sich in Abb. 2.5. Durch Okklusion etwa kann die relative Anordnung von Objekten in der Tiefe bestimmt werden, ohne dass dies jedoch eine Information über absolute Distanzen liefern würde (relative Tiefe). Wichtig ist hier auch die Halb-Okklusion, die dazu führt, dass bestimmte Teile der Szene nur für ein Auge sichtbar sind. Ihr starker Einfluss wurde von Nakayama [Nak96] nachgewiesen. Die bekannte Größe eines Objektes hingegen kann in Zusammenhang mit der Größe seiner retinalen Abbildung zur absoluten Tiefenbestimmung genutzt werden. Aber auch die Textur kann dazu dienen, Verläufe von Tiefe zu berechnen. Schattierung trägt häufig dazu bei, die lokale Tiefenstruktur von Objekten zu bestimmen. Die Perspektive, als Projektion dreidimensionaler Objekte auf eine zweidimensionale Fläche ist ein Hinweis, der nur unter Annahmen über die dreidimensionale Struktur der Objekte hilfreich ist. So hilft die Hypothese eines flach nach hinten verlaufenden Bodens zur Schätzung von Entfernungen bei Objekten, die sich auf dem Boden befinden. Auch die Akkomodation der Augen liefert Information über die Entfernung eines Objektes. Ist sie bekannt, kann die Entfernung scharf abgebildeter Objekte ungefähr abgeleitet werden. Dieser Hinweis wird jedoch nur in einem Bereich bis zu etwa 2 Metern Entfernung benutzt.

Neben diesen vielfältigen Hinweisen muss wohl die versetzte Abbildung von Strukturen auf beide Augen als wichtigste Quelle gelten. Diese von der Tiefe abhängige Abbildung derselben Struktur auf horizontal verschobene Bereiche der Retina wird als Stereo- oder Querdisparität bezeichnet. Sie steht bei konstanter Vergenz der Augen in direktem Zusammenhang mit der Entfernung. Die Stereodisparität wird nur in einem Bereich von ± 12 Bogenminuten zuverlässig berechnet, dem Panumschen

[1] Umgekehrt kann die Entfernung eines Objektes zur Bestimmung seiner Größe dienen, was etwa im Film „Herr der Ringe - Die Gefährten" dazu genutzt wurde, die Hobbits kleiner darzustellen, indem man sie weiter von der Kamera entfernt positionierte, als es den Anschein hatte. Dadurch wirkte es so, als ob die Schauspieler wesentlich kleiner wären.

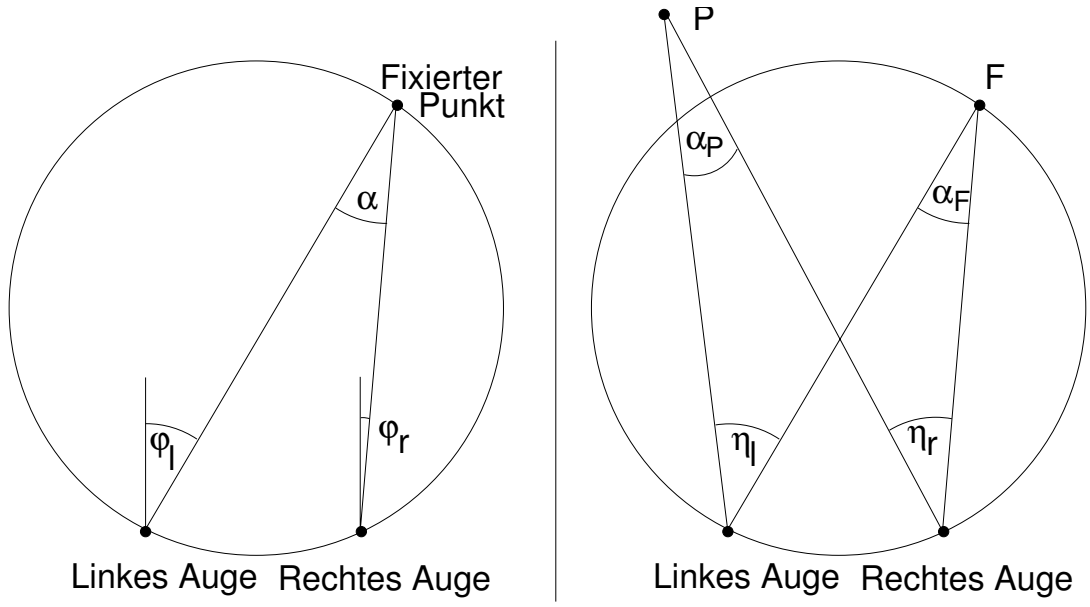


Abbildung 2.6: Links: Geometrie der binokularen Wahrnehmung bei Konvergenz der Augen nach Mallot [Mal98]. Für alle Punkte auf dem Kreis (Vieth-Müller-Kreis) gilt derselbe Konvergenzwinkel α als Differenz der Winkel φ_l und φ_r. Rechts: Berechnung der Disparität eines Punktes P bei Fixation von F als Winkelgröße: $\delta = \eta_r - \eta_l = \alpha_F - \alpha_P$.

Bereich.

Als Horopter bezeichnet man die Punkte in der Welt, die mit derselben Disparität abgebildet werden. Bei zentraler Fixation $\varphi_l = \varphi_r$ bildet der Vieth-Müller-Kreis den Horopter in der Ebene (s. Abb. 2.6). Die absolute Tiefe kann nur unter Kenntnis des Vergenzwinkels bestimmt werden. Das Hauptproblem in der Berechnung der Disparität liegt in der Bestimmung der zum selben Objekt gehörenden retinalen Abbildungen, es wird als Korrespondenzproblem bezeichnet. Bei komplexeren Strukturen kann selbstverständlich die Ähnlichkeit der Strukturen verwendet werden, um die Korrespondenzen zu identifizieren. Auch sind bestimmte Konstellationen in sich nicht konsistent (z.B. mehrfache Zuordnungen).

Dass es keiner komplexen Strukturen wie z.B. Kanten oder Flächen im Bild bedarf, um einen Tiefeneindruck zu erzeugen, demonstrieren die random-dot Stereogramme nach Julesz [Jul71]. Sie ergeben sich, indem man von einem Bild zufällig verteilter Punkte ausgehend, einen Bereich horizontal versetzt und die entstehende Lücke wieder mit zufälligen Punkten füllt. Auf diese Weise wird das Bild für das zweite Auge präpariert. Für den versetzten Bereich wird eine Disparität wahrgenommen.

Sofern höhere Strukturen vorhanden sind, werden sie auch genutzt, um die Tiefe zu ermitteln. So wird man ein leuchtendes Trapezoid in einem ansonsten dunklen Raum als geneigtes Rechteck wahrnehmen. Die Prozesse der Tiefenwahrnehmung und der perzeptuellen Organisation können sich also so wechselseitig beeinflussen, dass die Tiefeninformation zur Bildung einer Form beiträgt wie bei den random-dot Stereogrammen oder die Formwahrnehmung einen Tiefeneindruck erzeugt wie im Falle des Trapezoids.

Die ersten Zellen, die Stereoinformationen repräsentieren, finden sich im Areal V1. Diese Neuronen verfügen über je ein rezeptives Feld in beiden Augen, die beide Informationen liefern müssen. Natürlich gibt es viele andere Reize, die ebenfalls zur Reaktion eines solchen Neurons führen, doch

wird über eine geeignete Verschaltung erreicht, dass solche Fehlinterpretationen unterdrückt werden.

Schließlich kann auch die dynamische Veränderung des Bildes als sogenannte Bewegungsparallaxe Informationen über die Tiefe liefern. Bei einer Eigenbewegung des Beobachters verändert sich die retinale Abbildung der Objekte in Abhängigkeit von ihrem Abstand. Je näher die Objekte, desto stärker die Bewegung. Veranschaulichen lässt sich dies bei einer Autofahrt: die Bäume in der Nähe „bewegen sich schnell", die Brücke in einiger Entfernung nur langsam und Sterne oder Mond praktisch gar nicht. Der geleistete Beitrag wird laut Rock [Roc98] jedoch als gering eingeschätzt. Die Auflösung für Tiefe liegt unter günstigsten Bedingungen bei 3 bis 10 Bogensekunden.

2.1.4 Bewegungswahrnehmung

Aus der zeitlichen Veränderung der Bildes lassen sich Informationen über die Bewegung von Objekten der Umgebung, aber auch über die Bewegung des Beobachters selbst ableiten. Jedoch ist nicht jede Veränderung des Bildes auf eine Bewegung zurückzuführen. Eine wichtige Aufgabe der Bewegungswahrnehmung ist es, dies zu unterscheiden. Neben Veränderungen der Beleuchtung ist es vor allem auch die Eigenbewegung des Beobachters, die eine starke Veränderung des Sinneseindruckes verursacht.

Neuronale Grundlage sind raum-zeitliche rezeptive Felder, die man durch die Verschaltung von Neuronen mit räumlich versetzten rezeptiven Feldern erhält, wobei der Eingang eines der Neuronen mit einer zeitlichen Verzögerung versehen wird. Man erhält Neuronen, die lokal auf Bewegungen bestimmter Geschwindigkeit reagieren, indem ein raum-zeitlicher Gradient gebildet wird. Andere Verschaltungen wie verzögerte Inhibition oder Verwendung von zeitlich und räumlich differenzierenden Neuronen sind ebenfalls möglich. Das resultierende Verhalten entspricht den Reaktionen simpler Zellen. Was man erhält, ist ein Indiz dafür, dass eine derartige Bewegung stattgefunden hat.

Im weiteren werden die vielen möglichen Bewegungen, die so im Bild wahrgenommen werden, miteinander abgeglichen, um zu einem eindeutigen optischen Fluss zu gelangen, der aus der wahrgenommenen retinalen Verschiebung besteht. Er wird herangezogen, um die Bewegungen von Objekten, aber auch die Eigenbewegung zu erschließen und daraus weiterhin die räumliche Struktur abzuleiten (*structure from motion*). Letzteres wird besonders deutlich in der Betrachtung von random-dot Kinematogrammen, die weniger bekannt sind als die entsprechenden Stereogramme. Sie bestehen aus einer Menge von Punkten, die sich entlang definierter Trajektorien bewegen. Entsprechen die Trajektorien einer dreidimensionalen Oberfläche, so wird diese Oberfläche wahrgenommen, ohne dass es zusätzlicher Hinweise wie Kanten, Schattierung oder Stereodisparität bedarf. Die Wahrnehmung hat selbst dann Bestand, wenn die einzelnen Punkte jeweils nur kurz dargeboten und dann gelöscht werden, um später an anderen Positionen wieder aufzutauchen.

Viele Untersuchungen befassen sich mit Scheinbewegungen, die nicht aus kontinuierlichen Veränderungen bestehen, sondern bei denen diskrete Sprünge ab einer gewissen Geschwindigkeit und bis zu einer gewissen Distanz den Eindruck einer kontinuierlichen Bewegung hervorrufen (Fernseher und Monitore beruhen etwa auf dieser Technik). Bewegungs- und Tiefenwahrnehmung sind eng miteinander verwandt, denn aus dem Bewegungsfeld lassen sich Tiefeninformationen ableiten. Andererseits sind Tiefenhinweise wichtig zur Bestimmung der räumlichen Bewegung von Objekten.

Auch gilt Bewegung als Merkmal, das besonders stark die Aufmerksamkeit auf sich zieht. Neuronal wird die Bewegungswahrnehmung vor allem in den Arealen V5 und MT lokalisiert [Zek93]. Die

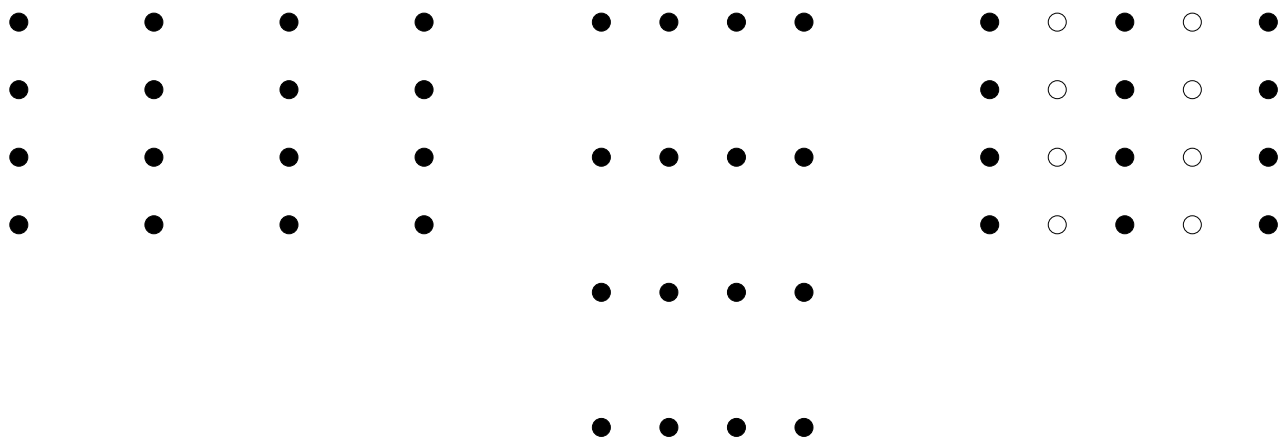


Abbildung 2.7: Illustration zweier Gestaltgesetze. Während die **Nähe** der Elemente dafür sorgt, dass die Punkte links in Spalten, in der Mitte aber in Zeilen organisiert wahrgenommen werden, sorgt rechts bei gleichem Abstand die **Ähnlichkeit** der Elemente für eine Einteilung in Spalten.

Psychophysik der Bewegungswahrnehmung wurde ausführlich von McKee und Watamaniuk [MW94] analysiert.

2.1.5 Segmentierung und Gruppierung

Nachdem bisher hauptsächlich die Auswertung von lokalen Informationen beschrieben wurde, stellt sich die Frage, wie diese Informationen zu zusammenhängenden Objekten gruppiert werden können bzw. eine Aufteilung des Bildes in einzelne Segmente vorgenommen wird. Die Gestaltgesetze nach Wertheimer [Wer23] beschreiben phänomenologisch die Arbeitsweise der Gruppierung. Darunter befinden sich die Gesetze der Nähe, der Ähnlichkeit, des gemeinsamen Schicksals, der Prägnanz, der guten Gestalt und der Geschlossenheit, von denen zwei in Abb. 2.7 illustriert werden.

Die Vorstellung von einem einzelnen Prozess, der zur Gruppierung führt, wurde von Zucker [Zuc87] verworfen. Er zeigt vielmehr, dass eine Vielzahl unterschiedlicher Mechanismen zum Phänomen Gruppierung beitragen. So können die bis jetzt diskutierten Reizeigenschaften Farbe, Tiefe und Bewegung einen starken Einfluss auf die Einteilung des Stimulus ausüben.

Im Zusammenhang von Aufmerksamkeit und Gruppierung ist eine der wichtigsten Fragen die nach der Reihenfolge. Operiert Aufmerksamkeit auf bereits gruppierten Elementen, ist Gruppierung nur durch Zuweisung von Aufmerksamkeit möglich oder kann man sich eine Unabhängigkeit und Parallelität beider Prozesse vorstellen? Moore und Egeth [ME97] haben gezeigt, dass Gruppierung ohne Aufmerksamkeit stattfinden kann. Auch sie nehmen jedoch einen Zusammenhang an, der darin bestehen könnte, dass Aufmerksamkeit zwar nicht für die Gruppierung, wohl aber für die Speicherung der Gruppierungsergebnisse im Gedächtnis notwendig sei. Anhand eines Patienten mit intakter räumlicher Aufmerksamkeit, aber gestörter Gruppierung und Symmetriewahrnehmung, konnten Vecera und Behrmann [VB97] zeigen, dass umgekehrt Aufmerksamkeit keine Gruppierung der Objekte voraussetzt.

Den scheinbaren Widerspruch, dass Gruppierung und Aufmerksamkeit trotzdem nicht unabhängig erscheinen, lösten Trick und Enns [TE97] durch die Annahme einer zweistufigen Gruppierung. Von diesen zwei Stufen soll die erste, das *Clustering,* vor der Zuweisung von Aufmerksamkeit - also

präattentiv - stattfinden. Die eigentliche Gruppierung zu einer Form, die die zweite Stufe darstellt, soll hingegen Aufmerksamkeit voraussetzen und somit attentiv stattfinden. Diese Unterscheidung in einen Prozess, der Einheiten in einer Gruppe zusammenfasst und einen zweiten, der die Form der Gruppe bestimmt, geht schon auf Koffka [Kof35] zurück, wurde jedoch seitdem weitgehend ignoriert.

Als Übersicht zum Thema sei noch auf die Arbeit von Kehrer und Meinecke[KM96] verwiesen.

2.1.6 Objekterkennung

In einer segmentierten Szene ein Objekt als solches zu erkennen und von zu anderen unterscheiden, erscheint in vielen Kontexten als die entscheidende, gleichzeitig auch die schwierigste Aufgabe, die ein Sehsystem zu lösen hat. Zwei bedeutende Schulen der Objekterkennung unterscheiden, ob primär Modelle der Objekte mit dem visuellen Reiz abgeglichen werden (z.B. bei Biedermann [Bie85, Bie87]) oder ob für ein Objekt mehrere Ansichten repräsentiert sind, wofür etwa Bülthoff et al. [BET95] experimentelle Evidenz anbringen. Neuronal wird die Objekterkennung vor allem im inferotemporalen Kortex (IT) lokalisiert [You95], einem Teil des sogenannten *Was*-Pfades im Gegensatz zum *Wo*-Pfad zur Lokalisation [UM82]. Zellen in IT, die *elaborate cells,* reagieren auf die Präsenz einfacher Formen unabhängig von Größe und Position dieser Formen [FTIC92]. In einigen Fällen hat man sehr spezifische Neuronen gefunden, die zum Beispiel auf Gesichter reagieren.

Ein Modell zur attentiven Objekterkennung, das sich eng an der Unterscheidung eines Wo- und eines Was-Kanals hält, stammt von Carpenter, Grossberg und Lesher [CGL98]. Es wird auf neuronale Weise nicht nur der Ort, sondern auch Skalierung und Orientierung eines Objektes bestimmt. Damit lässt sich die Ortsinformation von einer normalisierten Repräsentation des Objektes trennen, die zum Abgleich mit gespeicherten Objekten geeignet ist.

Der Zusammenhang von Objekterkennung und Aufmerksamkeit ist in zweierlei Hinsicht relevant: einerseits wird im allgemeinen fokale Aufmerksamkeit vorausgesetzt, um Objekte erkennen zu können, andererseits stellen visuelle Objekte einen Kandidaten als Einheit der attentiven Selektion dar. Es stellen sich also ähnliche Fragen wie im Verhältnis von Gruppierung und Aufmerksamkeit, wobei der Zusammenhang zur Erkennung von Objekten unter dem Stichwort „frühe vs. späte Selektion" später ausführlich diskutiert wird (Kap. 3.3.1).

2.1.7 Augenbewegungen

Aufgrund der stark varianten Auflösung der Retina stellt die Ausrichtung des Blicks und die daraus resultierende Möglichkeit zur Wahrnehmung eines Bereiches mit hoher Auflösung einen wichtigen Aspekt der Aufmerksamkeit dar, der als offene Zuweisung von Aufmerksamkeit beschrieben wird. Der varianten Auflösung der Retina entspricht auch die Anzahl von Neuronen, die für die Verarbeitung einer Retinaposition verantwortlich sind. Dies wird als kortikaler Abbildungsmaßstab (M-Skalierung) beschrieben [RV79]. Darüber hinaus kennt man qualitative Unterschiede zwischen der Fovea und extrafovealen Bereichen gerade hinsichtlich des Lernens und Erkennens von Objekten [RJ96]. Somit ist also für das Lösen komplexer Aufgaben eine Fovealisierung interessanter Bildbereiche unumgänglich. Neben der Ausrichtung des Kopfes oder des ganzen Körpers, die mit einem hohen Zeit- und Energieaufwand verbunden sind, ist es die Bewegung der Augen zu einem Ziel, die diese Fovealisierung herbeiführt.

Bei der Ausrichtung der Augen unterscheidet man primär langsame Folgebewegungen und sogenannte Sakkaden. Während die Folgebewegungen kontinuierlich einem sich bewegenden Objekt folgen (anders sind kontinuierliche Augenbewegungen nicht möglich), stellen Sakkaden eine ballistische Bewegung zu einem entfernten Ort dar. Neben den beschriebenen Typen gibt es noch weitere Bewegungen der Augen, so z.B. Vergenzbewegungen, gegenläufige Bewegungen beider Augen, die zur Fixation eines Punktes in der Tiefe dienen sowie kompensatorische Bewegungen zum Ausgleich von Kopf- und Körperbewegungen. Folgebewegungen erreichen Geschwindigkeiten bis etwa 20 bis 30 Grad pro Sekunde erreichen, wogegen Sakkaden, die nur etwa 20 bis 100 ms dauern, mit 20 bis 600 Grad je Sekunde stattfinden [Mal99]. Die meisten Sakkaden sind vergleichsweise kurz, in der Untersuchung von Malinov et al. [MEHS00] liegen mehr als die Hälfte unter 5°, 83 % unter 15°.

Im Normalfall liegt die Vorbereitungszeit für eine Sakkade bei etwa 200 ms. Jedoch kennt man auch sogenannte Expresssakkaden, die vor allem von Fischer und Kollegen [FB83, Fis98] untersucht wurden. Unter sehr spezifischen Bedingungen - ein Verschwinden des fixierten Objektes bei gleichzeitigem plötzlichen Auftauchen eines neuen Objektes - können diese deutlich schneller ablaufen. Damit ist die Geschwindigkeit von Sakkaden meist wesentlich langsamer als die von Aufmerksamkeitswechseln, die bei etwa 50 ms liegen [SJ91]. Allerdings bezeichnet Ward [War01] die allgemeine Schätzung von 50 ms als zu niedrig und argumentiert für eine vergleichbare Dauer von Aufmerksamkeitswechsel und Sakkade bei etwa 200 ms.

2.1.8 Visuelles Gedächtnis

Bei den Gedächtnisstrukturen für visuelle Informationen unterscheidet man üblicherweise zumindest das Kurzzeit- oder Arbeitsgedächtnis vom Langzeitgedächtnis. Während das Kurzzeitgedächtnis dazu dient, die gerade zur Verarbeitung benötigten Einheiten vorzuhalten, ist das Langzeitgedächtnis ein dauerhafter Speicher, dessen Inhalt im Bedarfsfall in das Arbeitsgedächtnis übertragen wird.

Der kurzfristige Speicher lässt sich weiter differenzieren. Man kennt das ikonische Gedächtnis (*iconic memory*, so genannt von Neisser [Nei67]), das an retinale Koordinaten gebunden ist und die letzte Wahrnehmung zur Verfügung stellt, also etwa zur Erinnerung der Szene dient, wenn die Augen geschlossen werden. Durch sogenannte Maskierung, d.h. eine Veränderung oder kurzzeitiges Darbieten eines anderen Reizes an derselben Position kann das ikonische Gedächtnis gelöscht werden. Es arbeitet weitgehend unabhängig von der Komplexität der Reize. Auf der anderen Seite gibt es das Arbeitsgedächtnis VSTM (*visual short term memory*), das von Maskierung unbeeinflusst, jedoch abhängig von der Reizkomplexität arbeitet. Es ist nicht vom letzten visuellen Eindruck determiniert, sondern kann auch durch Vorstellungen (*visual imagery*) bestimmt werden. Der Repräsentationsrahmen ist nicht an retinale Koordinaten gebunden.

Luck und Vogel [LV97] versuchten die Kapazität des visuellen Arbeitsgedächtnisses einzuschätzen. Erste Experimente zeigten eine Grenze, die bei der Speicherung von vier Objekten lag. Interessanterweise werden jedoch zu den vier Objekten jeweils mindestens vier Merkmale zuverlässig gespeichert. Es standen damit insgesamt mindestens 16 Merkmale zur Verfügung unter der Bedingung, dass sie auf höchstens vier Objekte verteilt waren. Es handelt sich demnach um eine objektbasierte Strukturierung.

Interessant ist, dass einerseits Aufmerksamkeit dazu dient, die Exploration der Umgebung zu serialisieren, andererseits Veränderungen der Umgebung ohne fokale Aufmerksamkeit häufig unbemerkt

bleiben. Dies deutet darauf hin, dass Veränderungen unter natürlichen Umständen Aufmerksamkeit anziehen. Die Repräsentation der Umgebung, die uns als vollständiges „Bild" der Szene erscheint, scheint dabei viel stärker nicht bildlich strukturiert zu sein. Eine gute Übersicht zum visuellen Gedächtnis gibt Logie [Log95].

2.2 Computerimplementationen visueller Wahrnehmung

Die digitale Bildverarbeitung folgt nicht notwendigerweise denselben Strukturen wie die natürliche visuelle Wahrnehmung, denn die grundlegende „Hardware" ist eine andere. Die Sensorik weist im natürlichen Fall zum Beispiel eine ortsvariante Auflösung auf, der eine homogene Auflösung der üblichen technischen Bildsensoren gegenübersteht. Die weitere Verarbeitung wird durch die unterschiedliche Anzahl und Geschwindigkeit der Berechnungseinheiten unterschieden. Auf der einen Seite zeichnet sich das Gehirn mit einem vergleichsweise langsamem Takt (Feuerraten der Neurone liegen in der Größenordnung von 1 kHz) durch einen extrem hohen Grad an Parallelität aus. Dieser wird durch die Verschaltung von mehr als 10 Milliarden Neuronen erreicht. Im Gegensatz dazu arbeiten Computer weitgehend seriell (selbst Parallelrechner verfügen über eine kleine Anzahl von Prozessoren) bei sehr hohem Takt von derzeit mehreren GHz. Trotzdem sind die zu lösenden Probleme vergleichbar und man kann eine abstraktere Sichtweise auf das Problem des Bildverstehens als informationstechnischer Fragestellung einnehmen.

Die Verwendung von Computern zur Auswertung von Bilddaten reicht schon recht lange zurück. Einen wissenschaftlichen Anspruch begründete aber erst Marr [Mar82] durch die Beschreibung der visuellen Wahrnehmung als *computational vision*.

Im weiteren sollen Systeme der industriellen und medizinischen Bildverarbeitung vernachlässigt werden. Sie befassen sich mit sehr stark eingeschränkten Domänen, in denen die allgemeinen Probleme eines flexiblen und robusten Sehsystems keine Rolle spielen. Durch die Möglichkeit zur Festlegung von Beleuchtung und Aufnahmegeometrie sowie Einschränkungen im Hinblick auf die Menge der möglichen Eingabedaten wird die Verarbeitung hier erleichtert.

Als Übersichtswerke seien vor allem die Lehrbücher von Jähne [JMNS96, Jäh97], Mallot [Mal98], Pitas [Pit00] und Jain [JKS95] empfohlen.

2.2.1 Architekturen

Sehsysteme lassen sich als passiv oder aktiv klassifizieren. Eine weitere Differenzierung soll später durchgeführt werden. Die passiven Systeme leiten sich von den genannten Arbeiten von Marr [Mar82] ab. Ziel ist eine Rekonstruktion der Umgebung, die erzielt wird, indem vom aufgenommenen zweidimensionalen Bild ausgehend schrittweise immer komplexere und abstraktere Strukturen abgeleitet werden, die bis zu einer dreidimensionalen Repräsentation von geometrischen Formen reichen (*primal sketch*). Erst diese Repräsentation wird interpretiert. Es handelt sich um einen rein datengetriebenen Prozess, der unbeeinflusst von Zustand und Intention des Systems abläuft.

Als Reaktion auf die Probleme mit diesem Ansatz entwickelte sich das Paradigma des Aktiven Sehens ([AWB87, Baj88, Mer96], ein Überblick wird in [MS99] gegeben). Es zeichnet sich dadurch aus, den Bildaufnahmeprozess aktiv so zu gestalten, dass die aktuell zu lösende Aufgabe möglichst weit vereinfacht wird, indem zum Beispiel eine geeignete Aufnahmeposition gewählt wird. Dadurch

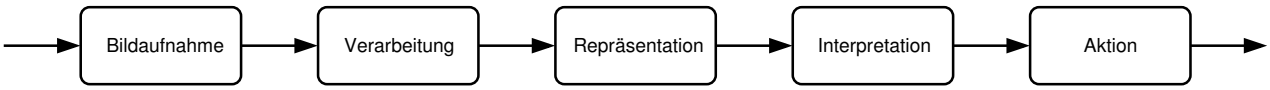


Abbildung 2.8: Klassischer Ablauf der Bildinterpretation nach Marr [Mar82]

sollen Mehrdeutigkeiten, die in einzelnen Aufnahmen auftreten können, beseitigt sowie beste Voraussetzungen zur Vereinfachung der Berechnung von Szeneneigenschaften geschaffen werden. Wichtig ist im Aktiven Sehen die Betonung des Zweckes, einer Aufgabe, die aktuell gelöst werden muss, während im Passiven Sehen die vollständige Rekonstruktion der Szene im Vordergrund steht. Genauer werden Aktive Sehsysteme im Kapitel 4.2 behandelt.

Einen interessanten Ansatz zum modularen Aufbau von Sehsystemen in Anlehnung an das menschliche Sehen findet man in den Visuellen Routinen nach Ullman [Ull84]. In dieser Sichtweise besteht ein Sehsystem aus einer Vielzahl von einfachen Operationen auf Bildausschnitten oder bereits verarbeiteten Bildausschnitten, Verfahren zur Selektion solcher relevanter Bildausschnitte und einer einfachen Struktur zur Speicherung der verarbeiteten Daten. Eine Anwendung solcher Visueller Routinen findet sich z.B. bei Salgian und Ballard [SB98] für die Steuerung von Fahrzeugen.

Leavers [Lea94] beschreibt exemplarisch eine häufige Einteilung der Verarbeitung. Eine frühe Verarbeitungsstufe dient dazu, parallel Hinweise auszuwerten, die zur Lokalisierung interessanter Bereiche dienen, die dann seriell in eine höhere Bearbeitungsstufe eingehen. Vielfach wird aufgrund der bisherigen Beobachtungen ein Modell der Umgebung erstellt, aus dem Vorhersagen über zukünftige Sensordaten abgeleitet werden. Diese Vorhersagen werden dann im aktuellen Bild validiert, auftretende Differenzen fließen in das Modell ein. Die Unterschiede führen nach entsprechender Interpretation zur Aktualisierung des internen Modells.

2.2.2 Farbe

Obwohl Farbe einen wichtigen Hinweis darstellt, der allein eine Unterscheidung bei konstantem Grauwert erlaubt, tendiert man bei Computer Vision-Systemen dazu, sich auf Monochrombilder zu beschränken. Dies hängt vor allem mit der Varianz gegenüber Einflüssen der Beleuchtung zusammen [Ebn01].

Ausgenutzt wird Farbe immer dort, wo gerade die Farbe für relevante Objekte charakteristisch ist, wie etwa Hautfarben zur Lokalisation und Erkennung von Menschen [BBC+97, BBB+98], Erkennung von (Verkehrs-)schildern [PKL+94, Ras02] oder der Detektion von Landmarken [YKLH01]. Meist wird dabei Farbe mit einem weiteren Hinweis kombiniert, wie im Segmentierungsverfahren von Shor und Kiryati [SK01].

In technischen Systemen wird der RGB-Standard verwendet, der sich aus additiver Zusammensetzung dreier Grundfarben mit den Wellenlängen

$$\lambda_R = 700\,nm;\ \lambda_G = 546,1\,nm;\ \lambda_B = 435,8\,nm$$

ergibt. Er hat sich jedoch aufgrund mangelnder Konstanzleistungen als für die Bildinterpretation ungeeignet erwiesen. Zur Standardisierung von Farben werden speziell die vom CIE definierten Farbräume XYZ und Lab herangezogen [Ill86]. Für die technische Verarbeitung werden jedoch häufig

Alternativen verwendet, die getrennte Komponenten für Helligkeit, Sättigung und Farbton enthalten. Von besonderer Bedeutung ist hier der Munsell-Farbraum, wie er im *Munsell Book of Colors* [Mun66] definiert wird. Es handelt sich um eine zylindrische Repräsentation, die die Farben in den drei Komponenten *Value* für die Helligkeit, *Chroma* für die Sättigung und *Hue* für den Farbton darstellt. Der Farbton wird dabei durch einen Kreis repräsentiert. Alternativen stellen der HSI-(*hue, saturation, intensity*) und der HLS-Farbraum (*hue, lightness, saturation*) dar. Die notwendigen Transformationen werden zum Beispiel von Harrington [Har87] beschrieben. Eine Einführung in die Farbwissenschaften für Computeranwendungen geben Sharma et al. [SVT98].

2.2.3 Tiefe

Abgesehen von zusätzlichen Modifikationen der Aufnahmegeometrie stehen den technischen Systemen zur Bestimmung der Tiefe prinzipiell dieselben Möglichkeiten zur Verfügung wie den natürlichen Systemen. So kann man die Tiefe aus der Textur, der Schattierung, der Perspektive, der Verdeckung, der Bewegung und aus Stereoinformationen berechnen. Textur, Schattierung, Perspektive und Verdeckung liefern jedoch nur Hinweise in eingeschränkten Bildbereichen. Bewegung setzt voraus, dass sich günstigerweise der Beobachter gerade so bewegt, dass sich möglichst viele Informationen über die Bewegungsparallaxe ergeben. Während es Verfahren gibt, die auf diesen Hinweisen beruhen [Mal98], ist die hauptsächlich genutzte Quelle jedoch die Querdisparität (s. Abb. 2.6) mit dem zu lösenden Korrespondenzproblem für die Strukturen in beiden Bildern. Hat man das Korrespondenzproblem einmal gelöst, ergibt sich die Tiefe durch einfache Triangulation.

Um die Suche nach den korrespondierenden Elementen auf eine Dimension einschränken zu können, ist es wichtig, die Epipolargeometrie der Kameras zu kennen. Für einen gegebenen Punkt in einem Bild können die korrespondierenden Punkte im anderen Bild nur auf einer Linie liegen, der sogenannten Epipolarlinie. Daher werden die Bilder vor der Korrespondenzberechnung üblicherweise rektifiziert, d.h. so transformiert, dass die Epipolarlinien den Zeilen der Bilder entsprechen. Die Ansätze zur Lösung des Korrespondenzproblems teilen sich in intensitätsbasierte, merkmalsbasierte und phasenbasierte Verfahren.

Intensitätsbasierte Verfahren beruhen auf einer direkten Korrelation der Grauwerte mit einer Fensterfunktion, die für eine lokale Beschränkung und Gewichtung sorgt. Das Verfahren beruht auf der Annahme, dass korrespondierende Strukturen auf weitgehend konstante Grauwerte in beiden Bildern abgebildet werden. Während sich so eine sehr dichte Disparitätskarte erhalten lässt, ist die Hypothese konstanter Grauwerte aus unterschiedlicher Perspektive wegen des unterschiedlichen Effektes der Beleuchtung für verschiedene Beobachtungspunkte problematisch.

Die merkmalsbasierten Verfahren versuchen dagegen Strukturen im Bild zu identifizieren, die so gut unterscheidbar sind, dass sich die Bestimmung der Korrespondenz entsprechend vereinfacht. Typische Merkmale wären Kanten, Ecken oder Liniensegmente. Das Aperturproblem sorgt hier dafür, dass nur Strukturen mit einer sichtbaren vertikalen Komponente zur Disparitätsschätzung beitragen. Einer horizontalen Geraden lässt sich keine Disparität zuordnen (s. Abb. 2.9).

Entscheidend ist die Auswahl der Merkmale. Sehr spezifische Merkmale resultieren in schneller zu findenden eindeutigeren Zuordnungsergebnissen, sind jedoch nicht überall im Bild präsent und sorgen so für eine dünn besetzte Disparitätskarte. Einfache Merkmale sind im Bild zwar stärker vertreten, daher aber auch weniger eindeutig einander zuzuordnen. Hier bestimmt die weitere Verwendung der

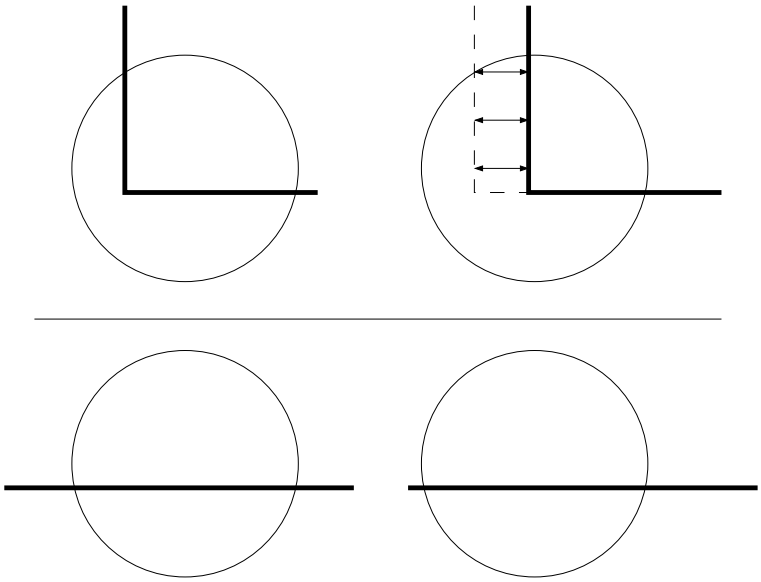


Abbildung 2.9: Illustration des Aperturproblems. Während in der oberen Hälfte im betrachteten Ausschnitt (Kreis) jeweils eine vertikale Komponente vorhanden ist, die die Disparitätsbestimmung erlaubt, fehlt sie in der unteren Hälfte und verhindert eine eindeutige Bestimmung der Disparität.

Tiefendaten die Auswahl der geeigneten Merkmale. In einem beschränkten Suchbereich werden nun die Korrespondenzen bestimmt, wobei die Ähnlichkeit der Strukturen und Nebenbedingungen wie die lokale Ähnlichkeit oder die Vollständigkeit und Eindeutigkeit der Zuordnung in die Berechnung eingehen können. Häufig werden solche Verfahren in einer Multiskalenrepräsentation durchgeführt, wobei von den gröberen Auflösungen ausgegangen wird. Der Suchraum wird dabei um die bereits gefundenen Disparitäten verschoben, so dass er für die feineren Auflösungen verkleinert werden kann. Auf die Triangulation der Entfernung folgt häufig eine Rekonstruktion der Oberfläche, die anhand einer Interpolation oder Füllung der fehlenden Werte eine dichte Entfernungskarte erzeugt.

Die Stärke phasenbasierter Verfahren nach Sanger [San88] liegt darin, dass sie ohne eine explizite Suche nach der maximalen Korrelation oder Ähnlichkeit von Merkmalen auskommen. Durch frequenzselektive Filterung und Extraktion der lokalen Phaseninformation, die voneinander subtrahiert wird, erhält man ein der Disparität proportionales Ergebnis. Dieses ist jedoch bis auf ein Vielfaches der Filterbreite unbestimmt. Auch ist die Information in schwach strukturierten Bereichen oft nicht ausreichend, was zu instabilen Phasenschätzungen führt. Beide Probleme werden durch mehrfache oder adaptive Filterung gelöst, wie sie unter anderem Theimer und Mallot [TM94] vorstellen. Einen ausführlicheren Einblick in die Bestimmung von Tiefeninformationen bieten Jiang und Bunke [JB97] sowie Mallot [Mal99] und Brown et al. [BBH03].

2.2.4 Bewegung

Die Rekonstruktion von Bewegungsinformationen beginnt mit dem Optischen Fluss. Dieser stellt eine Annäherung an das zweidimensionale Bewegungsfeld auf der Sensorebene dar. Das Bewegungsfeld entsteht wiederum durch eine Projektion der eigentlich interessierenden dreidimensionalen Bewegungen, dem Bewegungsflussfeld. Diesen Prozess zu invertieren ist Aufgabe der Bewegungserkennung.

Im Vergleich zur Stereoanalyse stehen bei der prinzipiell beliebig ablaufenden Bewegung nicht so

viele Einschränkungen zur Verfügung, die die Berechnung vereinfachen könnten. Zentrale Annahme ist hier, dass alle zeitlichen Änderungen im Grauwertbild auf Bewegung zurückzuführen ist, so dass der Optische Fluss die zeitlich versetzten Bilder ineinander überführen kann. Die Suche nach korrespondierenden Bereichen ist von zwei einander widersprechenden Einflüssen geprägt. Während große Bereiche für eine Überwindung des Aperturproblems sorgen, vermeiden kleinere Bereiche das Überschreiten von Grenzen unterschiedlicher Bewegungen. In der Bewegungswahrnehmung besteht das Aperturproblem darin, dass für beobachtete eindimensionale Strukturen nur der sogenannte normale Flussvektor berechnet werden kann, der orthogonal zur Struktur verläuft.

Eine Klasse von Lösungen basiert auf der Hornschen Bedingung [HS81], die eine Linearisierung der Grauwerte einer lokalen Umgebung anhand des ersten Elementes einer Taylorreihenentwicklung vornimmt. Problematisch sind Diskontinuitäten an den Grenzen unterschiedlicher Bewegungen. Phasenbasierte Verfahren finden in vergleichbarer Form wie bei der Disparitätsbestimmung Verwendung; es findet eine lokale richtungsselektive Filterung etwa mittels eines orientierten Gaborfilters [Gab46] statt. Anhand der Phasendifferenz erhält man jeweils die Richtungsvektoren orthogonal zur Filterorientierung. Es gelten die zuvor erwähnten Einschränkungen bezüglich der Uneindeutigkeit des Ergebnisses.

Neben den Verfahren zum Optischen Fluss kennt man auch korrelationsbasierte Verfahren, die jedoch mit großem Rechenaufwand einhergehen und allenfalls im Zusammenhang mit Auflösungspyramiden Verwendung finden. Analog zur Disparitätsbestimmung gibt es auch merkmalsbasierte Verfahren, die von der Hypothese der Grauwertkonstanz abstrahieren können, aber die Probleme einer zweidimensionalen Suche einer mangelhaften Merkmalsdichte lösen müssen.

Der Optische Fluss ist nun Basis weiterer Verfahren, die aus ihm weitere Informationen zur Struktur der Umgebung und ihrer räumlichen Veränderung ableiten. So geht man davon aus, dass die lokale Bewegungsinformation aus der dreidimensionalen Bewegung von Objekten und der Eigenbewegung des Beobachters hervorgegangen ist. Sofern die Eigenbewegung nicht bekannt ist, kann man versuchen, sie zu berechnen, wodurch jedoch die Tiefe nur bis auf einen konstanten Faktor zu bestimmen ist [BJT90]. Die verbleibende Bestimmung der Objekte und ihrer räumlichen Bewegung ist nicht eindeutig möglich. Die Ableitung kann jedoch durch Einschränkungen der Bewegungsart (etwa rein translatorisch) oder des Objektzusammenhanges (Annahme weniger Objekte) unterstützt werden. Die Komplexität des Problems liegt darin, dass gleichzeitig eine Segmentierung und ein Satz von sechs Parametern für jedes segmentierte Objekt (drei translatorische und drei rotatorische Bewegungskomponenten) ermittelt werden müssen.

Verwandt mit Verfahren zur Bewegungsdetektion sind solche zur Verfolgung. Verfolgungsverfahren legen den Schwerpunkt darauf, dauerhaft die Position eines oder weniger Objekte zu bestimmen und eventuell die Sensoren oder den ganzen Beobachter auf dieses Objekt auszurichten. Man differenziert zwischen Verfahren, die über ein Modell der zu verfolgenden Elemente verfügen und modellfreien Verfolgungsverfahren. Trotz der großen Datenmenge, die mit Bildfolgen einhergeht, gibt es einen Trend dazu, nicht nur zwei aufeinanderfolgende Bilder zu betrachten, sondern die Zeit als weitere Dimension in die Analyse von Ortszeitbildern mit einzubeziehen. Neben der Einführung von Jähne [Jäh97] sei für eine ausführliche Diskussion der Problematik vor allem auf die ausführliche Übersicht von Haußecker und Spies [HS99] mit einer weitergehenden Abwägung der Vor- und Nachteile einzelner Verfahren verwiesen.

2.2.5 Segmentierung und Gruppierung

Die Zuordnung der lokalen Informationen zu Segmenten oder Gruppen stellt einen entscheidenden Schritt zur Interpretation des Bildes dar. Schwerpunkt der Segmentierung ist die Zuordnung von Pixeln zu zusammengehörigen Segmenten, während es bei der Gruppierung eher um die Feststellung von Konstellationen aus mehreren Segmenten geht. Der Übergang ist allerdings fließend.

Für die Segmentierung unterscheidet man solche Verfahren, die sich eher auf die Eigenschaften der Fläche beziehen, von jenen, die sich auf die Grenzen zwischen Segmenten konzentrieren. Zu den flächenbasierten Verfahren gehören das Regionenwachstum, *split-and-merge* oder der *Watershed*-Algorithmus, um nur einige Beispiele zu nennen. Beim Regionenwachstum (*region growing*) wird von einem oder mehreren Startpunkten aus ein Wachstum in die Nachbarschaft gestartet, das auf einem Ähnlichkeitsmaß beruht. Stoßen zwei Wachstumsfronten aneinander, werden die Regionen entweder verschmolzen oder das Wachstum an der Stelle eingestellt. Dabei wird immer eine Statistik bezüglich der Eigenschaften der Region aktuell gehalten. *Split-and-merge*-Verfahren fangen umgekehrt mit einem einzigen Segment an. Dieses wird aufgeteilt (*split*), wenn es ein Homogenitätskriterium nicht erfüllt, worauf die Segmente auf Kompatibilität untersucht und eventuell verschmolzen werden (*merge*). Das Verfahren wird rekursiv auf die entstandenen Segmente angewandt.

Der *Watershed*-Algorithmus [VS91] beruht auf einer Interpretation der Bildwerte als Höhen, für die sich eine Gradientenberechnung eignet. Für einen gegebenen Schwellwert werden nun maximale zusammenhängende Bereiche unterhalb des Schwellwertes bestimmt. Es findet ein „Fluten" des Bildes statt, bei dem von den Bereichen ausgehend der „Pegel des Wassers" Schritt für Schritt erhöht wird und so die Einflussbereiche lokaler Minima als Segmente bestimmt werden können.

Kantenbasierte Verfahren beruhen auf der Suche nach Konturen. Die Positionen maximaler Veränderung findet man dabei als Maxima lokaler Gradienten. Diese Konturen müssen geschlossen werden, um Segmente zu erhalten. Eine Formulierung dieses Problems als Optimierung einer Energiefunktion stellen die sogenannten Snakes [KWT87] und ähnliche aktive Konturmodelle dar. Sie verwenden eine Energiefunktion, die die Einfachheit der Kontur und die Stärke des Intensitätsgradienten einbezieht. Als Merkmale zur Segmentierung dienen neben der direkten Verwendung der Grauwerte auch die zuvor besprochenen Merkmale Farbe, Bewegung und Tiefe (z.B. [BSP01]).

Bei der Gruppierung oder perzeptuellen Organisation liegt der Schwerpunkt eher auf der angemessenen Zusammenfassung von Bildteilen, zum Beispiel Merkmalen oder Segmenten. Hier finden komplexere Bildeigenschaften wie Verdeckung, gute Form und Gestaltgesetze Berücksichtigung. Die Ansätze zur Lösung sind überaus vielfältig und stellen einen wichtigen Schritt zur Bildinterpretation hinsichtlich Figur-Hintergrund-Trennung oder Objekterkennung dar. Einen Review perzeptueller Organisation bieten Sarkar und Boyer [SB93, SB94]. Sie strukturieren die Verfahren einmal nach Ausnutzung der Informationen 2D vs. 3D, mit oder ohne Zeit und zum anderen nach der Repräsentation, auf der sie operieren: Signal, einfache Merkmale, Strukturen oder Gruppen.

2.2.6 Objekterkennung

Hauptprobleme der Objekterkennung als ultimativem Schritt zum Verstehen des Bildes sind Robustheit, Skalierung und Aufwand. Prinzipiell geht es um den Abgleich von Bildinformationen mit einer Datenbank gelernter Objekte. Man unterscheidet dabei die ansichtenbasierte von der modellbasierten Erkennung. Ansichtenbasiert werden mehrere 2D-Ansichten des Objektes aus unterschiedlichen

Blickwinkeln gespeichert, die mit dem aktuellen Bild(-ausschnitt) abgeglichen werden. Bei der modellbasierten Erkennung wird hingegen versucht, ein 3D-Modell des Objektes mit den Bildinformationen in Übereinstimmung zu bringen.

Schmalz [Sch00] unterscheidet in einem Review folgende Verfahren:

- geometrische Ansätze, bei denen Objekte über Graphenstrukturen oder Formbeschreibungen repräsentiert werden,
- Optimierungsverfahren, die ein Maß an Übereinstimmung zwischen Modell und Bild durch eine hochdimensionale Energiefunktion bestimmen, die es zu minimieren gilt,
- unterschiedliche biologienahe Ansätze, häufig konnektionistisch aufgebaut,
- Aktive Sehsysteme, die die Sensoren so anzusteuern versuchen, dass das Erkennungsproblem vereinfacht wird.

Das Objekterkennungsproblem wird bei den meisten Verfahren in Isolation betrachtet, indem also für das segmentierte Abbild eines nicht verdeckten Objektes ein Abgleich mit der Datenbank vorgenommen wird. Diese Segmentierung stellt einen entscheidenden Schritt dar, um nur genau die Merkmale abzugleichen, die tatsächlich zum Objekt gehören. Während im industriellen Kontext eine solche Segmentierung noch leicht zu erreichen scheint [WP99], ist sie in natürlichen Szenen nur mit erheblichem Aufwand zu erreichen. Aktive Sehsysteme stellen einen Ansatz dar, die Erzielung geeigneter Sensordaten für die Objekterkennung innerhalb einer natürlichen Szene als Bestandteil des Gesamtsystems zu modellieren.

Allgemeine Objekterkennung, wie das natürliche Vorbild sie leistet, ist derzeit noch nicht erreichbar. Den bestehenden Verfahren mangelt es derzeit an Invarianz, Flexibilität, Robustheit, Skalierbarkeit und Effizienz. Das beschränkt die Verwendung von Objekterkennungsverfahren auf begrenzte Domänen oder kontrollierbare Bedingungen. Aufmerksamkeit kann verwendet werden, um die Rahmenbedingungen, unter denen die Objekterkennung operiert, zu optimieren.

2.2.7 Anwendungskontext: Mobile Autonome Systeme

Systeme der Bildverarbeitung finden in unterschiedlichen Kontexten Verwendung. Besonders anspruchsvoll ist derzeit die Verwendung für mobile autonome Systeme. So müssen sich mobile Roboter in unbekannten Umgebungen zurechtfinden, sich in einer dynamischen Umgebung bewegen und dabei Objekte erkennen und klassifizieren [KR99]. Für alle diese Aufgaben kann die Interpretation von Bilddaten herangezogen werden. Hier wird ersichtlich, dass neben der reinen Leistungsfähigkeit vor allem Robustheit, Flexibilität und Effizienz von Bedeutung sind.

Eine Schwierigkeit in solchen Anwendungen liegt darin, dass wenig Kontrolle über die Umgebung existiert. In der Industriellen Bildverarbeitung werden Beleuchtung, Lage der Objekte, ablenkende oder störende Objekte, Verdeckungen oder Aufnahmegeometrie weitgehend kontrolliert und optimal eingestellt. Im Unterschied dazu muss bei mobilen autonomen Systemen die Umgebung so hingenommen und verarbeitet werden, wie sie ist. Es lassen sich meist nur allgemeine Annahmen treffen, sei es, dass sich das System auf Straßen bewegt oder in Büroumgebungen. Die Autonomie des Systems hat auch Einschränkungen bezüglich einer exakten Kalibrierbarkeit zur Folge, denn diese kann in ungünstigen Umgebungen negativ beeinflusst werden. Die Autonomie des Systems impliziert einen hohen

Anspruch an das Verhalten, das auch unter unerwarteten Bedingungen die Sicherheit des Systems und der Umgebung gewährleisten muss, ohne die Notwendigkeit für einen kontrollierenden Eingriff seitens eines menschlichen Operators.

Einen umfangreichen Review zum Einsatz verschiedener Sensoren für Mobile Roboter bietet Everett [Eve95].

2.3 Zusammenfassung

Während sich dem natürlichen und dem technischen System zur Verarbeitung visueller Informationen vergleichbare Aufgaben stellen, können die Lösungsansätze sehr unterschiedlich sein. Der Hauptunterschied liegt in der Verwendung von Alltags- oder Weltwissen durch das natürliche System. Das Weltwissen vereinfacht die Lösung vieler Probleme, indem bestimmte Lösungen, die auch der Sensorinformation entsprechen würden, als unwahrscheinlich bewertet werden, da sie nicht dem Weltwissen entsprechen. Dieser Weg ist aufgrund der unüberschaubaren Menge an nötigem Weltwissen und der problematischen Formalisierbarkeit schwer auf den Computer zu übertragen. Ein möglicher Ansatz liegt jedoch in der Übernahme von wenigen generellen Heuristiken, die in vielen Bereichen anwendbar sind. Zu diesen Heuristiken kann man auch die visuelle Aufmerksamkeit zählen, die darauf beruht, einfache Hinweise zur Lokalisierung von relevanten Elementen auszunutzen, um komplexe Verarbeitungsschritte auf diese Elemente reduzieren zu können. Zur technischen Verwendung ist es notwendig, eine geeignete Formalisierung der Heuristik in einem implementierbaren Modell vorzunehmen. Im folgenden Kapitel wird die Arbeitsweise natürlicher visueller Aufmerksamkeit knapp dargestellt.

Kapitel 3

Natürliche visuelle Aufmerksamkeit

3.1 Einführung

Mit der natürlichen Ausprägung der visuellen Aufmerksamkeit zu beginnen ist angezeigt, da von ihrer Beschreibung und Modellierung die Umsetzungen im technischen Bereich mehr oder weniger abgeleitet sind.

Der Leser sei für umfassendere Einführungen in die natürliche visuelle Aufmerksamkeit auf die leicht lesbare, einführende Arbeit von Styles [Sty97], die auf Visuelle Suche konzentrierten Übersichtsartikel von Wolfe [Wol96, Wol00] und besonders auf das umfassende und aktuelle Werk von Pashler [Pas98] verwiesen. In diesem Kapitel soll nach einer Einführung in das Thema zuerst die empirische Untersuchung von Aufmerksamkeit gewürdigt werden mit besonderer Beachtung von Dynamik und Tiefe. Auf diese Befunde aufbauend werden Modellierungen visueller Aufmerksamkeit betrachtet, wie sie aus theoretischer, psychophysischer und konnektionistischer Sicht erstellt wurden. Anschließend werden die Daten und Modelle in Bezug gesetzt zur offenen Aufmerksamkeit, wie sie sich in Blickbewegungen ausdrückt. Vor allem der Zusammenhang von verdeckter und offener Aufmerksamkeit ist hier von Interesse. Schließlich sollen Befunde Erwähnung finden, die den klassischen Modellen wie dem Scheinwerfer (scheinbar) widersprechen. Dabei soll einerseits die Frage untersucht werden, ob Selektion immer oder hauptsächlich räumlich abläuft oder auch andere Einheiten in Frage kommen. Andererseits wird die Singularität des Fokus der Aufmerksamkeit in Frage gestellt und analysiert, ob mehrere Foki existieren oder eine andere Art von Selektion existiert.

Es gibt leider keine allgemein akzeptierte Definition visueller Aufmerksamkeit. Dies ist umso problematischer, als der wissenschaftliche Aufmerksamkeitsbegriff der umgangssprachlichen Verwendung zwar ähnelt, aber einen anderen Schwerpunkt setzt. Die umgangssprachliche Verwendung des Begriffes steht meist im Zusammenhang mit einem „Mehr" oder „Weniger" an Aufmerksamkeit. Etwas, das im Englischen eher dem Begriff der *alertness* (Wachsamkeit) entspricht und zum Beispiel auch in der immer häufiger bei Kindern diagnostizierten *attention deficit disorder* (ADD, auch ADHD: *attention deficit/hyperactivity disorder*) als mangelnde dauerhafte Aufmerksamkeit zum Ausdruck kommt [Bar95]. Aufmerksamkeit beschreibt dort einen Zustand von Konzentration, der mehr oder weniger stark ausgeprägt ist. Dagegen wird in der wissenschaftlichen Untersuchung von Aufmerksamkeit der Schwerpunkt auf die Selektions- und Auswahlaspekte gelegt. Man untersucht, wie Aufmerksamkeit die Wahrnehmung beeinflusst, indem bestimmte Teile der Umwelt bevorzugt oder vernachlässigt werden, beziehungsweise deren Verarbeitung mit mehr oder weniger Ressourcen ausgestattet wird. Es

geht also darum, bestimmte Teile der Umwelt oder interne Repräsentation zu beachten oder zu ignorieren und für ihre Verarbeitung mehr oder weniger Ressourcen zur Verfügung zu stellen. Für diese Arbeit soll im Wesentlichen die Definition von Corbetta [Cor90] herangezogen werden:

> „Attention defines the mental ability to select stimuli, responses, memories, or thoughts that are behaviorally relevant, among the many others that are behaviorally irrelevant."

Visuelle Aufmerksamkeit soll im Sinne dieser Definition als derjenige Aspekt der visuellen Wahrnehmung verstanden werden, der für eine bevorzugte oder eingeschränkte Verarbeitung der visuellen Reize oder aus visuellen Reizen abgeleiteten Repräsentationen sorgt und die Auswahl trifft, welche Reize auf welche Art und Weise verarbeitet werden. In dieser Hinsicht stellt die Aufmerksamkeit eine entscheidende Stufe der Kognition dar. Ihre Steuerung entscheidet darüber, welche Information der Erkennung, dem Gedächtnis, dem Bewusstsein, dem Lernen oder zu steuernden Aktivitäten zugeführt und welche ihnen vorenthalten wird. Im folgenden wird diese Umschreibung genauer betrachtet, wobei grundsätzlich von visueller Aufmerksamkeit die Rede ist. Es gibt jedoch sensorische Aufmerksamkeit auch für andere Sinne. Speziell für den auditiven Bereich gibt es ebenfalls eine Vielzahl empirischer Arbeiten [Che53, Bro58, Tre60].

Es finden sich in der Literatur verschiedene Metaphern, die den Effekt und die Arbeitsweise der visuellen Aufmerksamkeit umschreiben. Die bekannteste Metapher beschreibt einen Scheinwerfer der Aufmerksamkeit (*spotlight of attention*), ein Begriff der von Eriksen [EH73, EY85, EJ86] geprägt wurde. Der Scheinwerfer (die Übersetzung gibt nicht exakt die Bedeutung von *spotlight* wieder - gemeint ist ein Strahler, der etwa auf der Bühne eine einzelne Person hervorhebt) beschreibt einen räumlich zusammenhängenden, homogenen Bereich, der beleuchtet, also mit Aufmerksamkeit versehen wird. Allein in diesem Bereich sind komplexe Aufgaben wie zum Beispiel die Objekterkennung lösbar. Der verbleibende, dunkle Bereich wird weitgehend ignoriert. Der Scheinwerfer kann willentlich auf unterschiedliche Bereiche gerichtet werden, um die dort befindlichen Informationen verarbeiten zu können.

Hauptargument für die Verwendung von Aufmerksamkeit ist die angenommene Kapazitätsbeschränkung des Gehirns. Es ist einfach nicht möglich, alle Objekte einer Szene parallel auf einmal zu erkennen. Obwohl dies sicher nicht der einzige Grund für die Verwendung von Aufmerksamkeit ist - speziell die Hemmung störender Ablenker und die Isolierung eines Objektes zur Speicherung im Gedächtnis oder der Spezifikation von Motorprogrammen sei genannt, gibt es keine Zweifel an der Existenz derartiger Kapazitätsgrenzen [Pas98]. Wo genau diese Grenzen liegen, ist Gegenstand der Diskussion, die unter dem Stichwort „frühe vs. späte Selektion" zusammengefasst wird. Eine Möglichkeit im Umgang mit Kapazitätsengpässen ist die Serialisierung von Operationen. Indem ein Element nach dem anderen ausgewählt wird, verhindert man die Überschreitung von Kapazitätsgrenzen, die die gleichzeitige, parallele Anwendung von Operationen auf viele Elemente verhindern. Diese Strategie der Aufmerksamkeit findet man bei vielen Lebensformen bis hinunter zur Drosophila [HW84]. Man kann sich die Arbeitsweise von Aufmerksamkeit aber auch anders vorstellen, nämlich als kapazitätsbegrenzte parallele Abarbeitung oder als Filter. Diese Alternativen werden im Laufe des Kapitels weiter erläutert.

3.1.1 Grundsätzliche Unterscheidungen

In einem ersten Zugang zum Thema Aufmerksamkeit sollen Differenzierungen in der Arbeitsweise und dem Effekt von Aufmerksamkeit betrachtet werden.

Verdeckt und offen

Aufmerksamkeit lässt sich einfach dahingehend klassifizieren, ob sie durch motorische Aktionen erreicht wird, also durch Augenbewegungen, Orientierung von Kopf oder Körper. In diesem Fall, in dem die Zuwendung von außen beobachtbar ist, spricht man von offener Aufmerksamkeit (*overt attention*), speziell die Ausrichtung der Augen gilt dabei als interessant. Ist eine solche Beobachtung nicht möglich, wird also der Effekt von Aufmerksamkeit durch die interne Verarbeitung erreicht, bezeichnet man dies als verdeckte Aufmerksamkeit (*covert attention*). Der Schwerpunkt der psychophysischen Untersuchungen zur Aufmerksamkeit liegt im Bereich der verdeckten Aufmerksamkeit, der hier zuerst beleuchtet werden soll. Die Unterscheidung von offener und verdeckter Aufmerksamkeit und die Kenntnis der Möglichkeit, verdeckte Aufmerksamkeit unabhängig von der Blickrichtung zu steuern, wird bereits von Helmholtz [Hel96] beschrieben.

Datengetrieben und modellgetrieben

Man unterscheidet zwei Einflüsse auf die Zuweisung von Aufmerksamkeit. Zum einen kann Aufmerksamkeit willentlich gesteuert oder zumindest durch vorhandenes Wissen beeinflusst werden, andererseits durch wahrgenommene Reize bestimmt werden. Bei der als modellgetrieben, zielgetrieben oder top-down bezeichneten datengetriebenen Aufmerksamkeit spielen Erwartungen („Das Geräusch sagt mir, dass dort gleich ein Auto um die Ecke biegt."), interner Zustand („Ich bin hungrig. Wo finde ich etwas zu essen?") und Wissen („Ampeln haben eine wichtige Bedeutung.") eine Rolle. Determinieren dagegen Reizeigenschaften die Steuerung von Aufmerksamkeit, spricht man von datengetriebener oder bottom-up Aufmerksamkeit. Diese kann man sich leicht veranschaulichen, z.B. durch plötzliche Bewegungen im peripheren Gesichtsfeld, ein Blinken oder eine einzelne farbige Markierung in sonst kontrastarmem Umfeld. Im allgemeinen werden beide Aspekte gemeinsam in der Zuweisung von Aufmerksamkeit eine Rolle spielen. Die datengetriebene Aufmerksamkeit ist besser untersucht, da die visuellen Reize leichter experimentell zu kontrollieren sind, als Erwartungen, Zustand und Wissen.

Präattentiv und attentiv

Die Zuweisung von Aufmerksamkeit hat eine Unterteilung der visuellen Verarbeitung in zwei Teile zur Folge. Prozesse, die von Aufmerksamkeit unbeeinflusst bleiben und eine gleichmäßige Verarbeitung der Reize beinhalten, also vor der Zuweisung von Aufmerksamkeit stattfinden, bezeichnet man als präattentiv. Demgegenüber bezeichnet die attentive Verarbeitung alle Aspekte, die nur auf selektierten Reizen operieren oder in anderer Weise von der Zuweisung von Aufmerksamkeit beeinflusst werden. Zur präattentiven Verarbeitung gehören einfachere Operationen, wohingegen komplexere Verarbeitungen der attentiven Stufe zugeordnet werden. An welcher Stelle die genaue Grenze zwischen beiden Stufen liegt, wird unter den Schlagworten „Frühe Selektion" und „Späte Selektion" ausführlich diskutiert. Die Abgrenzung konnte bisher nicht eindeutig bestimmt werden.

Parallel und seriell

Speziell bei der visuellen Suche stellt man sich oft die Frage, inwieweit serielle und parallele Prozesse beteiligt sind. Diese Unterscheidung wird häufig gleichgesetzt mit paralleler Verarbeitung für den präattentiven Bereich und anschließender serieller attentiver Verarbeitung. In der visuellen Suche würde dies einer parallelen Integration von Informationen für die Salienzbestimmung und einer seriellen Stufe zur Identifikation der Elemente entsprechen. Wie jedoch Moore und Wolfe [MW01] argumentieren, ist es nicht notwendig, einen Widerspruch in seriellen und parallelen Prozessen zu sehen. Beide Aspekte können in einem System integriert sein, das nach außen je nach Bedingung seriell oder parallel arbeitend erscheint.

Unterdrückung und Anregung

Die Verarbeitung der Reize könnte durch Aufmerksamkeit prinzipiell auf zwei Weisen beeinflusst werden. Einerseits könnten die attendierten Reize durch Aufmerksamkeit verstärkt, angeregt oder hervorgehoben werden, andererseits könnte eine Hemmung oder Unterdrückung für die nicht attendierten Reize stattfinden. Es stellt sich die Frage, ob Aufmerksamkeit primär durch die Hervorhebung der ausgewählten Elemente oder die Hemmung nicht-ausgewählter Elemente operiert. Antworten geben neurowissenschaftliche Untersuchungen, die in Abschnitt 3.2.2 diskutiert werden.

3.2 Empirische Befunde zur visuellen Aufmerksamkeit

3.2.1 Ergebnisse der experimentellen Psychophysik

Visuelle Suche

Als zentrales Paradigma der psychophysischen Untersuchungen visueller Aufmerksamkeit muss die Visuelle Suche gelten. Hierbei muss eine Versuchsperson möglichst schnell angeben, ob in einem Display ein vorher spezifizierter Zielreiz (das *target*) anwesend war oder nicht. Variiert wird die Anzahl von Ablenkern, auch als Distraktoren bezeichnet. Gemessen wird neben der Fehlerrate vor allem die Reaktionszeit für das Finden und Erkennen des Targets. Steigt sie mit der Anzahl von Distraktoren an, z.B. in linearer Weise, geht man von einer seriellen Suche aus, bei der die Elemente des Displays in mehr oder weniger zufälliger Reihenfolge nacheinander ausgewählt und überprüft werden. Findet man keine deutliche Abhängigkeit der Reaktionszeit von der Anzahl der Distraktoren, spricht man hingegen von einer parallelen Suche, die das Target findet, ohne die Elemente erst durchsuchen zu müssen. Dies wird auch als *pop-out* bezeichnet.

Typische Experimente, in denen man eine parallele Suche findet, sind solche, in denen sich das Target anhand eines einzigen Merkmals von den Distraktoren unterscheiden lässt. Man spricht von einer Merkmalssuche, wenn z.B. ein rotes Target unter grünen Distraktoren zu entdecken ist. Ist dagegen die Verknüpfung mehrerer Merkmale nötig, den Zielreiz zu separieren, spricht man von der sogenannten Konjunktionssuche. In diesem Fall, wenn etwa ein roter Kreis von roten Quadraten und grünen Kreisen zu trennen ist, findet man oft stark ansteigende Reaktionszeiten für zusätzliche Distraktoren.

Die grundsätzliche Dichotomie von paralleler und serieller Suche wird heutzutage jedoch stark angezweifelt. Es gibt Hinweise, dass die Steigungen in den Reaktionszeiten eher ein Kontinuum dar-

stellen, wie die umfangreiche Metauntersuchung von Wolfe [Wol98] ergab. Auch kennt man mittlerweile sowohl verhältnismäßig langsame Merkmalssuchen, wie auch schnelle Konjunktionssuchen [TK94, TS90]. Aktuelle Modelle [Tre93, WCF89, Wol94, WG96], siehe Abschnitt 3.3.1, gehen daher von einem kontinuierlichen Grad an Schwierigkeit bei der Suche aus, der den Anstieg der Reaktionszeiten determiniert.

Erstaunliche Ergebnisse fanden sich bei der Untersuchung der Rolle des Gedächtnisses bei der Suche. Die üblichen Modellierungen gehen von einer dauerhaften Markierung der bereits mit Aufmerksamkeit versehenen Orte oder Elemente aus, um ein erneutes Durchsuchen zu vermeiden. Dagegen konnten Horowitz und Wolfe [HW98] zeigen, dass das Neuplatzieren aller Elemente im Abstand von 111 ms die Effizienz der Suche nach einem „T" unter „L" nicht beeinträchtigt. Hierzu ist noch kein aktuelles Modell vorhanden, das die Daten ausreichend erklärt.

Vertauscht man die Rolle von Zielreiz und Ablenker finden sich Asymmetrien in den Reaktionszeiten [Coh93]. Weitere Untersuchungen befassen sich mit der Frage, wie die Reaktionen bei abwesendem Zielreiz und speziell deren Reaktionszeiten zustande kommen [CW96]. Schließlich beziehen einige Forscher jetzt stärker die Rolle von Blickbewegungen bei der Visuellen Suche mit ein, siehe dazu die Diskussion unter 3.4.3.

Der gründliche Review von Wolfe [Wol96] zeigt, dass viele Untersuchungen sich auf das Paradigma der Visuellen Suche berufen. Mehrere Theorien widmen sich explizit den Ergebnissen und der Modellierung der Aufmerksamkeitsprozesse während der Visuellen Suche - speziell die Feature-Integration-Theorie von Treisman und das Guided-Search-Modell von Wolfe (siehe dazu Kap. 3.3.2). Sie üben auch einen starken Einfluss auf die Modellierung von Aufmerksamkeit im Computer-Sehen aus.

Weitere Experimentalparadigmen

Neben der Visuellen Suche stellen auch die Cueing-Experimente von Posner [PSD80, Pos80] klassische Experimente zur Aufmerksamkeit dar. Hier wurde vor der eigentlichen Entdeckungsaufgabe ein Hinweisreiz dargeboten, der einen Ort bezeichnete. Untersucht wurde nun die Abhängigkeit der Verarbeitung von der Gültigkeit des Hinweisreizes. Man fand, dass die Verarbeitung sowohl durch gültige Hinweise beschleunigt als auch durch fehlerhafte Hinweise verlangsamt wurde. Interpretationen dieses Effektes wiesen auf die Möglichkeit hin, beschränkte Ressourcen auf einen räumlichen Bereich beschränken zu können, was zur Beschleunigung der Verarbeitung führte.

Auf die Grenzen räumlicher Selektion gehen Experimente zur Flankerkompatibilität ein, die auf Eriksen zurückgehen [EH73, EY85, EJ86]. Bei diesen Experimenten war den Versuchspersonen der Ort, an dem der zu bewertende Zielreiz erscheinen würde, bekannt. Es gab mehrere mögliche Zielreize, denen zwei verschiedene Reaktionen zugeordnet wurden. Zum Beispiel wurde eine Reaktionstaste den Vokalen und eine andere Reaktionstaste den Konsonanten zugeordnet. Räumlich benachbart zum Zielreiz wurden die Flanker dargeboten, die nun in ihrer Kompatibilität zum Zielreiz variiert wurden. Die Flanker konnten also derselben Reaktion oder einer anderen Reaktion zugeordnet sein wie der Zielreiz. Man fand eine Verlangsamung der Reaktion bei inkompatiblen Flankern (im Beispiel also ein Vokal als Zielreiz und Konsonanten als Distraktoren oder Ablenker) im Vergleich zu kompatiblen Flankern (ein Vokal als Zielreiz und andere Vokale als Distraktoren. Dieser Effekt war also nicht mit

einer visuellen Ähnlichkeit oder gar der Identität konfundiert, vielmehr war es eine Variation auf semantischer Ebene.

So wurde eindrucksvoll die Erkennung und Verarbeitung von Informationen demonstriert, die nicht aufgabenrelevant waren. Durch Erhöhung des Abstandes von Zielreiz und Distraktoren ließ sich der Effekt eliminieren. Die erste Interpretation des Effektes definierte daher den Bereich, in dem er auftrat, als minimale Größe des „Scheinwerfers der Aufmerksamkeit". Der Kompatibilitätseffekt verschwindet jedoch nur dann, wenn der Abstand der Ablenker zum fixierten Zielreiz nicht durch eine Vergrößerung der Ablenker kompensiert wurde, die eine Erkennung trotz der außerhalb der Fovea reduzierten Auflösung der Retina erlaubt [Ege77]. Insgesamt handelt es sich um einen überaus stabilen Effekt, dessen Grenzen und Parameter von Miller [Mil91] untersucht wurden.

Baylis und Driver [BD92] demonstrierten, dass der Effekt des Abstandes auf die Flankerkompatibilität sich durch Variationen in der Ähnlichkeit sogar überschreiben ließ. Dazu wurde bei mehreren Distraktoren in unterschiedlichem Abstand die Ähnlichkeit zum Zielreiz hinsichtlich Farbe oder Bewegung variiert und die Effekte der verschiedenen Distraktoren analysiert. Dabei zeigte sich, dass ähnliche Distraktoren auch bei größerem Abstand einen stärkeren Kompatibilitätseffekt ausübten als unähnliche, aber nähere Distraktoren. In einem Beispiel kann dies also bedeuten, dass bei einem roten Zielreiz die direkt benachbarten grünen Flanker weniger Einfluss auf die Reaktion ausüben als die weiter entfernten roten Flanker. Interpretiert wurde dieser Effekt durch eine Gruppierung von Zielreiz und Ablenkern, die zwar auch durch die Nähe beeinflusst wird, aber eben auch durch die Ähnlichkeit.

Selektive Aufmerksamkeit spielt auch dann eine Rolle, wenn es um die Beachtung verschiedener Aspekte desselben Objektes geht, so dass eine räumliche Trennung nicht möglich ist. Der sogenannte Stroop-Effekt [Str35] bezeichnet ein experimentelles Paradigma, das dies verdeutlicht. Die Versuchsperson hat dabei die Aufgabe, die Farbe zu benennen, in der ein Wort dargestellt wird. Hierbei zeigen sich typischerweise Kompatibilitätseffekte, die sich in einer höheren Fehlerrate bzw. einer verlangsamten Reaktion für inkompatible Reize ausdrücken. Es dauert also länger, auf das in grün geschriebene Wort „Rot" mit der Antwort „Grün" zu reagieren, als wenn es sich in einer neutralen Bedingung um ein Wort gehandelt hätte, das selbst keine Farbe benennt.

Als *attentional blink* wird ein Effekt bezeichnet, auf den Raymond et al. [RSA92] verweisen. Bei der sogenannten RSVP (*rapid serial visual presentation*), die von Sperling [Spe60, SBSJ71] häufig zur Untersuchung des Zusammenhangs von Kurzzeitgedächtnis und Aufmerksamkeit genutzt wurde, werden an derselben Position in schneller (ca. 100 ms) Folge Reize präsentiert. Zur Untersuchung des *attentional blink* wurden auf diese Weise Ziffern dargeboten. Zweimal kommt ein Buchstabe vor, der von der Versuchsperson zu identifizieren ist. Der Effekt des *attentional blink* beschreibt nun die Schwierigkeit in der Identifikation des zweiten Targets, besonders dann, wenn es ca. 200 bis 500 ms nach dem ersten Target präsentiert wird. Der erste Buchstabe wird üblicherweise fehlerfrei erkannt. Sofern sich Maskierungseffekte ausschließen lassen, wird der Effekt als Abschottung des Systems gegen neue Reize bei Fokussierung der Aufmerksamkeit auf einen Reiz interpretiert.

Davon zu unterscheiden ist die *change blindness*, die anzeigt, dass Veränderungen im Bild ohne Zuweisung von fokaler Aufmerksamkeit oft unbemerkt bleiben [SL97, ROC97, ODCR00]. Dazu ist es allerdings notwendig, einen gewissen zeitlichen Abstand (ISI, Inter Stimulus Interval) zwischen dem ursprünglichen und dem veränderten Display zu lassen, der mindestens 50 ms beträgt. Andern-

falls wird die Veränderung als Blinken empfunden und leitet die Aufmerksamkeit an den Ort, an dem die Veränderung stattfand. Der Effekt ist nicht alleine auf Veränderungen beschränkt, die eine Identifikation voraussetzen würden, sondern gilt auch für das Hinzufügen und Entfernen von Objekten. Rensink [Ren02] findet Korrespondenzen zu Pylyshyn's FINST-Theorie oder den *object files* von Treisman bei Überwachungsaufgaben, die sich auf maximal vier bis fünf Elemente beschränken.

Auf der Suche nach einem zentralen Engpass (*central bottleneck*) oder dem Kapazitätslimit in der Verbindung von Wahrnehmung von Handlung werden Experimente zur Doppelaufgabeninterferenz durchgeführt. In ihnen wird untersucht, unter welchen Bedingungen zwei in starker zeitlicher Nähe angesiedelte Aufgaben einander stören. Meist benutzte Technik zur Bestimmung des Ausmaßes an Interferenz ist die *psychological refractory period*. Hier werden zwei Reize S1 und S2 zeitversetzt um ein Intervall SOA (*stimulus onset asynchrony*) dargeboten, auf die jeweils unabhängig schnell reagiert werden muss (Reaktionen R1 und R2). Gemessen wird jeweils die Zeit zwischen Reiz und Reaktion (RT1 und RT2). Während RT1 meist kaum vom SOA beeinflusst wird, findet man in bestimmten Zeitbereichen eine sehr starke Verlängerung von RT2 bei Verkürzung des SOA. Der Effekt tritt selbst bei einfachen Aufgaben auf und ist stabil über Eingabe- und Reaktionsmodalitäten und auch bei Verwendung unterschiedlicher Modalitäten (visuelle und auditive Präsentation, Reaktion als Tastendruck und Sprache) [Pas93]. Trotzdem ist die Gesamtzeit meist kürzer als die Summe der beiden einzelnen Reaktionen, was leicht in Form eines zentralen Flaschenhalses interpretierbar ist, der wohl hauptsächlich in der Vorbereitung der Reaktion liegt, weniger im perzeptiven Teil [Pas98].

Verstraten, Intriligator und Kollegen [VCL00, IC01] widmen sich der temporalen und räumlichen Auflösung von Aufmerksamkeit. Mit verschiedenen experimentellen Paradigmen wurde nahegelegt, dass Aufmerksamkeit zeitlich auf einen Wechsel von 4 bis 8 Hz beschränkt ist. Für längerfristige Skalen weist Enns [Enn90] hinsichtlich der Bedeutung von Aufmerksamkeit darauf hin, dass Aufmerksamkeit nicht nur das momentane Verhalten steuert, sondern durch die Unterdrückung von Wahrnehmung und Speicherung bestimmter Reize auch die gesamte Entwicklung, das Lernen und die Erinnerung beeinflusst.

3.2.2 Die neuronale Basis von Aufmerksamkeit

Während sich die Psychophysik mit von außen beobachtbaren Reaktionen der ganzen Person befasst, interessieren sich Neurowissenschaftler für die Umsetzung dieser und anderer Prozesse auf der Ebene von Neuronenverbänden oder einzelner Neuronen. Man unterscheidet Einzelzellableitungen, die die genaue Aktivation einzelner Neuronen wiedergeben von Verfahren, die die gemittelte Aktivation ganzer Hirnareale oder größerer Zellgruppen messen. Letztere werden vor allem in den letzten Jahren durch den Fortschritt im Bereich bildgebender Verfahren vermehrt eingesetzt. Die hier beschriebene Darstellung nimmt starke Vereinfachungen vor. Neben den dargestellten Verbindungen kennt man mittlerweile sehr viele weitere Verbindungen zwischen visuellen Arealen sowie Verbindungen, die den beschriebenen Verarbeitungspfaden gerade entgegenlaufen. Auch sind den beschriebenen Arealen der visuellen Verarbeitung (s. Abb. 3.1) meist weitere Aufgaben und Unterteilungen zuzuschreiben.

Um die Hirnbereiche einordnen zu können, in denen Aufmerksamkeit wirkt, ist die klassische Trennung der visuellen Verarbeitung in einen ventralen „Was"-Pfad (über die visuellen Areale V1, V2, V3, V4 nach IT), dem die Bereiche Identitätsinformationen und Objekterkennung zugeordnet sind und einen dorsalen „Wo"-Pfad, der für Verfolgung, Lokalisation und räumliche Interaktion zuständig

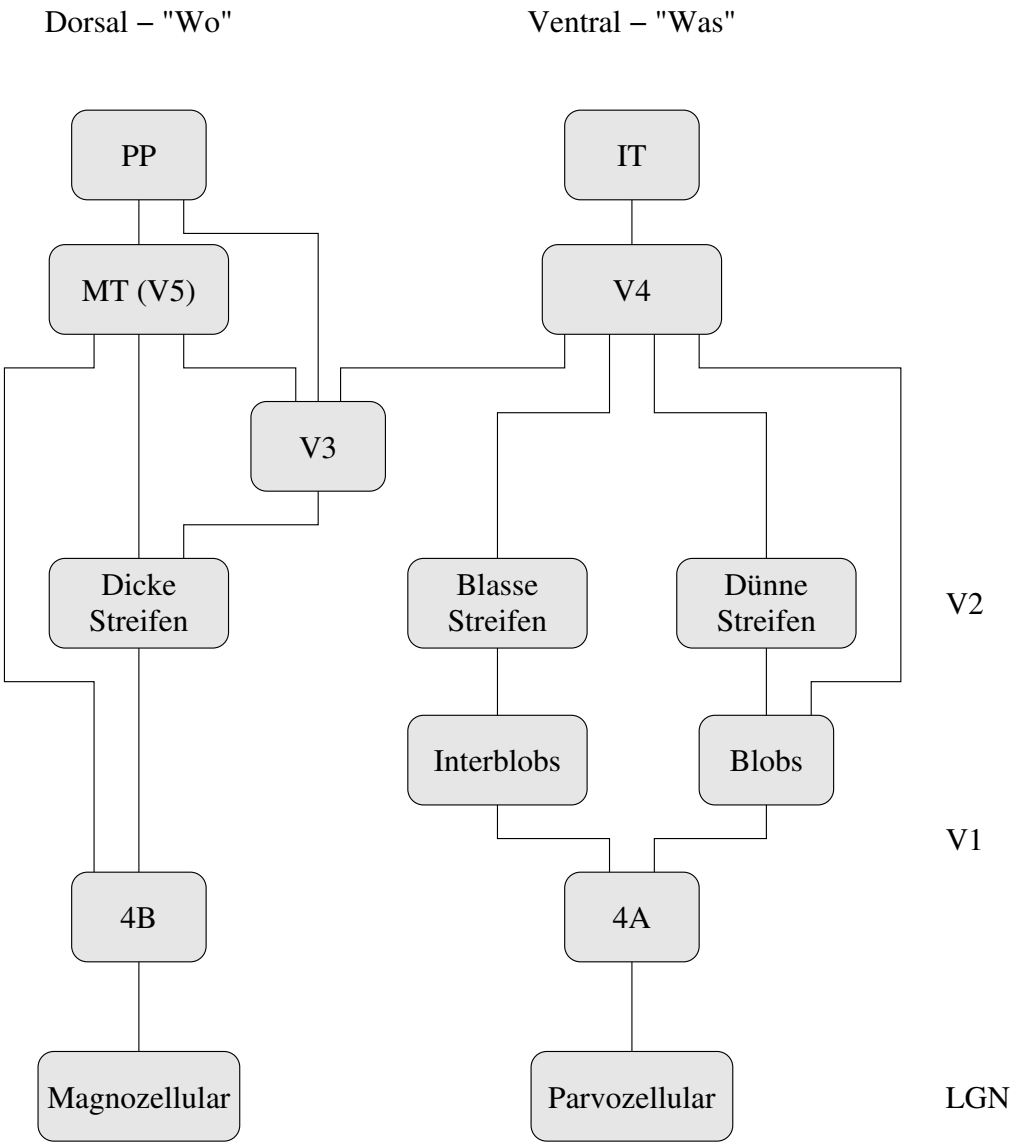


Abbildung 3.1: Verbindungen und Datenströme der frühen visuellen Areale (nach Bollmann [Bol00]).

ist, nützlich (s. Abb. 3.1). Sie geht auf Ungerleider und Mishkin zurück [UM82]. Speziell entlang des ventralen Pfades kann man beobachten, dass die sogenannten „rezeptiven Felder", also die retinalen Bereiche, in denen Stimuli zur Reaktion eines Neurons führen können, immer größer werden. Dafür findet in derselben Richtung aber auch eine immer weitergehende Spezialisierung auf die Art des Reizes statt, der eine Reaktion auslöst. Die Verarbeitung führt also von einer lokalen Bestimmung einfacher Merkmale hin zu einer globaleren Detektion komplexer Formen oder Attribute.

Die klassische Studie von Moran und Desimone [MD85] zeigt, dass Aufmerksamkeit bis hinunter auf den Level einzelner Neurone und ihrer rezeptiver Felder operiert. In den Arealen V4 und IT mit ihren verhältnismäßig großen rezeptiven Feldern findet eine Unterdrückung nicht-attendierter Reize statt, die sich in früheren Arealen (V1 und V2) mit den kleineren rezeptiven Feldern nicht zeigt. Diese Unterdrückung ist im Areal V4 aber davon abhängig, ob sich im rezeptiven Feld ebenfalls ein Zielreiz befindet. Im Areal IT findet sie vermutlich aufgrund der Größe der rezeptiven Felder grundsätzlich statt. Diese Studien beziehen jedoch ausschließlich die durchschnittliche Feuerrate der Neurone mit ein, nicht aber die zeitliche Verteilung, der in letzter Zeit immer mehr Bedeutung zugesprochen wird.

So wird die Synchronisation des Feuerns zwischen Neuronen, das als *temporal tagging* bezeichnet wird, als Alternative zur seriellen Lösung des Bindungsproblems angesehen [NKR93, SG95]. Die Präsenz von Merkmalen wird dabei durch die Feuerrate kodiert, während die Mikrostruktur des Feuerns als Tag dient, also als Indikator der Zugehörigkeit zu einem Objekt. Luck und Beach [LB98] zeigen allerdings, dass das nicht alleine zur Lösung des Bindungsproblems reichen kann, das in realistischen Szenen mit vielen beieinanderliegenden Objekten zu komplex ist. Dazu werden nach ihrer Auffassung zusätzlich Aufmerksamkeitsmechanismen benötigt, die durch Unterdrückung nicht-relevanter räumlicher Bereiche operieren.

Kastner und Ungerleider [KU00] beschreiben den Wettbewerb der visuellen Reize um eine neuronale Repräsentation im visuellen Kortex, für den es sowohl auf der Ebene der Einzelzellableitung als auch anhand bildgebender Verfahren Evidenz gibt. Dieser Wettbewerb kann datengetrieben und modellgetrieben beeinflusst werden, wobei der top-down-Einfluss sehr unterschiedlich stattfinden kann. Sie stellten sowohl eine Anregung der Neuronenaktivität, als auch eine Filterung der Reize, eine Erhöhung der Ruheaktivität und eine Erhöhung der Sensitivität fest. Somit kann datengetriebene Aufmerksamkeit auch ohne Präsenz visueller Reize nachgewiesen werden. Als Resultat des Wettbewerbs erhält der Sieger Zugang zum Gedächtnis für Speicherung und Abruf sowie zu Motorprogrammen.

Mit Hilfe der Anwendung funktioneller Magnetresonanztomographie (fMRI) während der Durchführung von Flankerkompatibilitätsaufgaben konnten Casey et al. [CTW+00] eine Trennung von Bereichen des Antwortkonfliktes gegenüber der räumlichen Aufmerksamkeit vornehmen. Durch Kombination der Lokalisation mittels PET (Positronen-Emissions-Tomographie) und fMRI mit der zeitlichen Auflösung ereigniskorrelierter Potenziale konnten Hillyard und Anllo-Vento [HAV98] demonstrieren, dass räumliche Aufmerksamkeit auch auf neuronaler Ebene früher und anders lokalisiert werden kann als merkmalsbasierte Aufmerksamkeit. Anhand von PET-Daten in Kombination mit ereigniskorrelierten Potenzialen zeigen Mangun et al. [MHS+01] räumlich-attentive Modulationen der Eingabeverarbeitung bei sehr einfachen Reizen, bei denen höhere Prozesse ausgeschlossen werden konnten. Die Modulationen traten selbst dann auf, wenn keine Distraktoren dargeboten wurden.

Im Gegensatz dazu konnten Luck und Ford [LF98] zeigen, dass die typischen Merkmale der Zuweisung von Aufmerksamkeit in ereigniskorrelierten Potenzialen genau dann auftreten, wenn ein

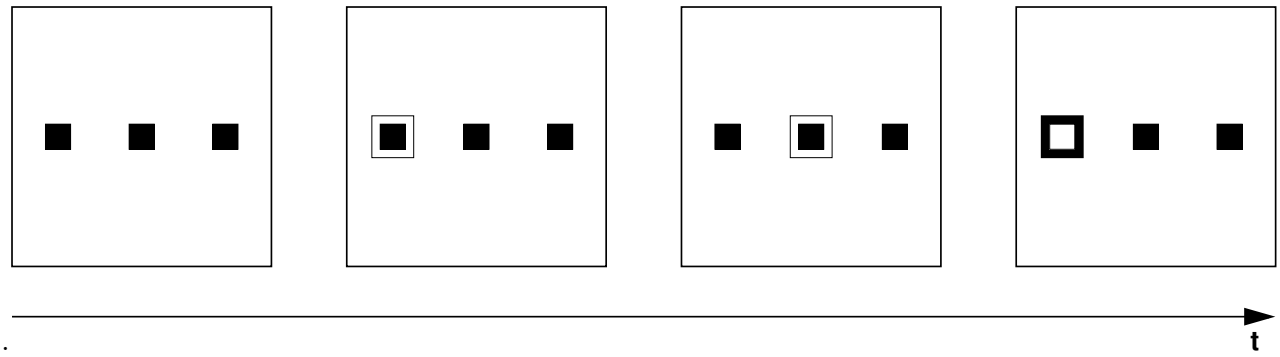


Abbildung 3.2: Schematische Darstellung der IOR-Experimente nach Posner und Cohen [PC84]. In zeitlicher Abfolge von links nach rechts folgen initiales Display, Cueing eines Objektes, Cueing des neutralen Objektes und zu detektierende Veränderung eines Objektes (in diesem Fall für das mit Cue versehene Objekt).

Mechanismus zur Trennung sonst mehrdeutiger neuronaler Kodierungen benötigt wird. Dies bestätigt die Rolle der Aufmerksamkeit in der Unterdrückung von Ablenkern und der korrekten Bindung von Merkmalen zu Objekten.

Sehr starke Effekte von Salienz fanden Gottlieb et al. [GKG98] im lateralen interparietalen Bereich (LIP) des posterioren Parietalkortex (PPC) bei Einzelzellableitungen an Affen. Im Normalfall finden sich hier nur sehr schwache Reaktionen auf Reize im rezeptiven Feld der Zellen. Reaktionen wurden nur dann erzielt, wenn die Salienz des Reizes manipuliert wurde, entweder durch einen plötzlichen Onset oder indem der relevante Reiz verhaltensrelevant wurde. Diese Aktivation war unabhängig von der Planung von Motorprogrammen. Die Repräsentation in diesem Bereich entspricht damit einer *master map of attention* oder *saliency map*.

Während es also schon viele einzelne Ergebnisse zur neuronalen Grundlage von Aufmerksamkeit gibt, fehlen derzeit noch Theorien, die diese zu einem einheitlich zusammenhängenden Bild zusammenfassen.

3.2.3 Inhibition of return

Neben der Zuweisung von Aufmerksamkeit kann man auch den Prozess des Loslösens der Aufmerksamkeit untersuchen, der einer neuen Zuweisung vorausgehen muss. Ein wichtiger Effekt wird hier als *Inhibition of return* (IOR) bezeichnet und beschreibt die Hemmung einer Aufmerksamkeitszuwendung an zuvor mit Aufmerksamkeit versehene Ziele. Der ursprünglich von Posner und Cohen [PC84] beschriebene Effekt zeigt sich, wenn eine Bewegung der Aufmerksamkeit zu einem Ort versucht wird, der bereits kurz zuvor mit Aufmerksamkeit versehen wurde. Die Dauer dieser Inhibition wird mit etwa 1,5 bis 2 Sekunden angegeben. Der Effekt stellt einen Aspekt des Kurzzeitgedächtnisses dar. Seine Bedeutung liegt darin, die Suche nach Objekten oder die Exploration einer Szene effizienter zu gestalten, indem bereits verarbeitete Bereiche für eine gewisse Zeit aus der Suche ausgeblendet werden und die Aufmerksamkeit auf kürzlich nicht beachtete Bereiche konzentriert werden kann.

Die experimentelle Evidenz geht auf Experimente zurück, deren Ablauf in Abb. 3.2 dargestellt wird. Ein Display aus drei horizontalen Elementen wird dargeboten, von denen das mittlere laut Instruktion von der Versuchsperson fixiert werden soll. Eines der äußeren Elemente wird durch Einblenden eines Quadrates hervorgehoben, wodurch die Aufmerksamkeit auf dieses Element gezogen

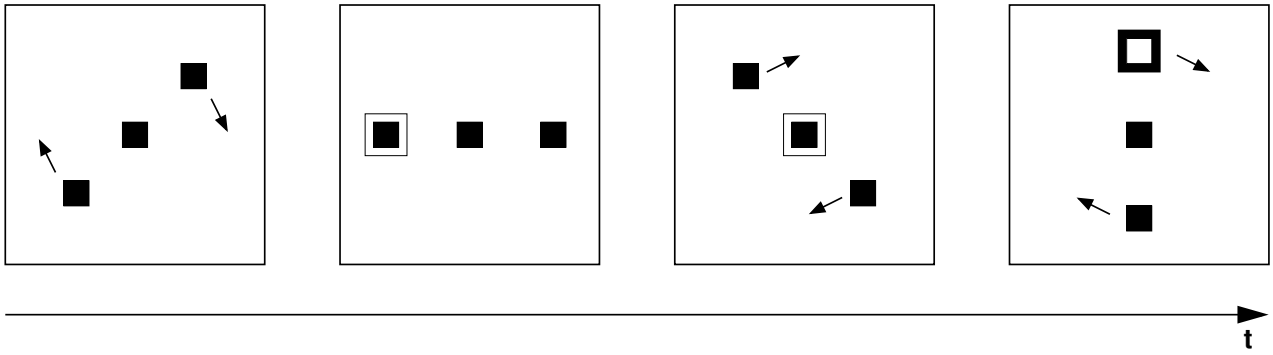


Abbildung 3.3: Experiment zur Bindung des IOR an Ort oder Objekte nach Tipper, Driver und Weaver [TDW91]. Entspricht in der zeitlichen Abfolge und Struktur Abb. 3.2, wobei sich die äußeren Objekte jedoch in einer gedachten Kreisbahn um das zentrale Objekt bewegen.

wird. Durch ebensolches Hervorheben des mittleren Elementes wird die Aufmerksamkeit zurück auf dieses neutrale Element gerichtet. Die Versuchsperson hat eine Diskriminationsaufgabe zu lösen; sie muss möglichst schnell entscheiden, welches der beiden äußeren Elemente sich verändert. Untersucht wird, wie sich die Reaktionen auf Veränderungen in Abhängigkeit vom Cue, also vom Hervorheben der Elemente, unterscheiden. Im Gegensatz zu den klassischen Cueing-Experimenten (s. Kap. 3.2.1) findet sich hier ein stark hemmender Effekt für das zuvor hervorgehobene Element. Dieser Effekt wird damit erklärt, dass die Aufmerksamkeit beim IOR-Experiment nach dem Cue wieder wegbewegt werden muss. Mit dem Zurückkehren zu einem zuvor verlassenen Ort sei ein zusätzlicher Aufwand assoziiert, dessen Ursache in einer Hemmung dieses Ortes liegt.

Die übliche Modellierung besteht in einer Hemmungskarte (*inhibition map*), in der der aktuell selektierte Ort mit einer hohen Aktivation versehen wird, die im Laufe der Zeit nachlässt. Diese Hemmungskarte wirkt inhibitiv auf die *Master map of attention* und bewirkt somit eine Unterdrückung der zwar auffälligen, aber bereits kürzlich selektierten Bereiche.

Das Modell bezieht allerdings keine dynamischen Veränderungen der Umgebung mit ein. Was passiert mit der Inhibition, wenn sich das Objekt vom inhibierten Ort weg bewegt? Diese Frage stellten sich Tipper et al. [TDW91] und variierten das ursprüngliche Experiment entsprechend so, dass eine Bindung der Inhibition an das Objekt und eine Bindung an den Ort genau entgegengesetzte Vorhersagen erlaubten. Wie in Abb. 3.3 dargestellt, wurden dazu die äußeren Objekte auf einem imaginären Kreis bewegt, so dass Cue und Detektionsaufgabe zwar an dasselbe Objekt gebunden waren, jedoch an unterschiedlichen Orten auftraten. In der entscheidenden Bedingung konnten sie sich sogar genau gegenüberliegen, so dass objektbasierte und raumbasierte Theorien exakt entgegengesetzte Vorhersagen treffen würden. Die Resultate demonstrierten eindeutig die Bindung der Inhibition an das selektierte Objekt, nicht an den Ort und schließen damit die Inhibitionskarte als einzig zutreffende Modellierung der Inhibition aus.

Abb. 3.4 illustriert zusätzlich den Effekt einer angenommenen Inhibitionskarte im Ablauf beider Experimente.

Spätere Untersuchungen wiesen darauf hin, dass es durchaus beide Effekte gibt [TW98b], die sich addieren oder gegenseitig vermindern können. So sind die relativ starken Effekte in den klassischen Experimenten von Posner wohl auf die Konfundierung objektbasierter und räumlicher Hemmung zurückzuführen. Zusammenfassend ist festzustellen, dass es eine Bindung der Inhibition an bewegte

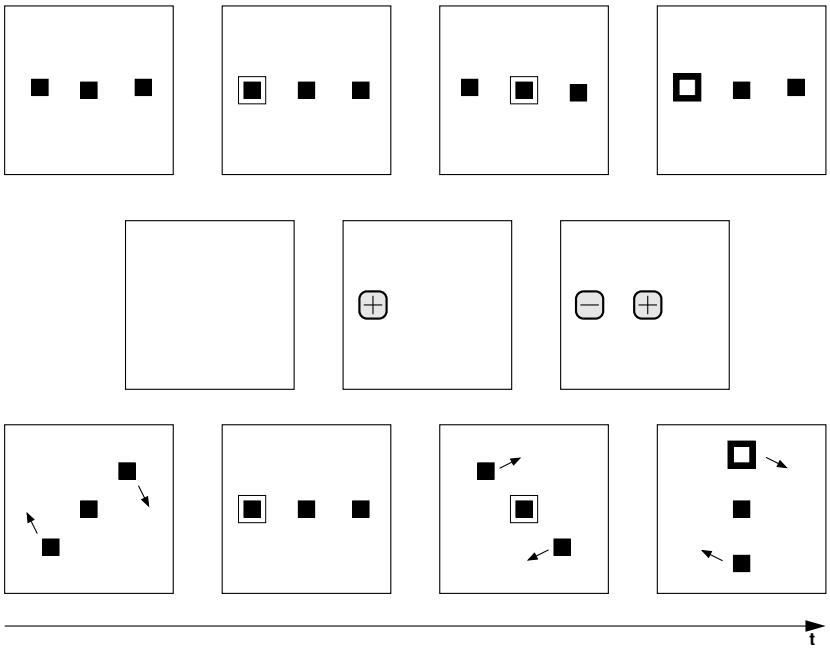


Abbildung 3.4: Effekt der Inhibitionskarte in den vorgestellten Experimenten(erste Reihe nach Posner, letzte Reihe nach Tipper et al.). In der Mitte ist der Zustand einer Auffälligkeitskarte dargestellt, die positive Einflüsse (datengetrieben durch Cue und modellgetrieben durch Aufgabenstellung) sowie negative Einflüsse einer statischen Inhibitionskarte enthält. Der Zustand der Karte ist jeweils zwischen den Veränderungen der Reize angegeben. Entscheidend ist der negative Anteil vor dem letzten Reiz, der zwar im oberen Experiment die Ergebnisse erklären kann, nicht jedoch im unteren Fall.

Objekte gibt, auch während diese Objekte nicht durch fokale Aufmerksamkeit selektiert werden.

3.2.4 Aufmerksamkeit und Tiefe

Räumliche Tiefe spielt in der Untersuchung visueller Aufmerksamkeit eine doppelte Rolle: einmal gilt es als salienzanzeigendes Merkmal wie Orientierung oder Farbe, dann aber auch als räumliche Dimension der Umgebung wie die Position im 2D-Raum. Es mag gerade die typische Darbietung der Reize in Experimenten zur Aufmerksamkeit auf einem Computermonitor sein, die zur Modellbildung anhand zweidimensionaler Reize und entsprechender kortikaler Karten führte. So wurde Tiefe zuerst primär die Rolle des Merkmals zugewiesen.

Dass es jedoch zumindest eine ungewöhnliche Rolle als Merkmal spielt, demonstrierten Nakayama et al. [NS86]. Auf ihre Ergebnisse wird sich bis heute vielfach berufen. Sie führten Experimente zur visuellen Suche (s. Kap. 3.2.1) durch, bei denen sie die übliche serielle Suche bei Konjunktionen zweier Merkmale replizieren konnten, Konjunktionen von Tiefe und einem anderen Merkmal jedoch zu einer parallelen Suche führten. Die Autoren nahmen daher an, dass sich die Aufmerksamkeit auf vorab spezifizierte Tiefenebenen beschränken lässt. Später wiesen He und Nakayama [HN95] jedoch nach, dass diese Interpretation der Daten unzutreffend war. Vielmehr stellen dreidimensionale Oberflächen, die sich nicht notwendigerweise in einer Tiefenebene befinden müssen, die relevante Selektionseinheit dar. So wäre die Zuweisung von Aufmerksamkeit innerhalb einer Ebene einfacher als eine Verteilung über zwei Ebenen.

Jedoch konnten Viswanathan und Mingolla [VM99] nachweisen, dass eine Verfolgung mehrerer

Objekte (siehe dazu auch den folgenden Abschnitt) über mehrere Tiefenebenen einfacher ist als die entsprechende Aufgabe in einer Ebene. Eine Aufteilung der Elemente in mehrere Farben erzeugte keine vergleichbare Vereinfachung.

McSorley und Findlay [MF01] weisen darauf hin, dass die absoluten Zeiten für die Konjunktion von Tiefe und anderen Merkmalen in [NS86] sehr hoch sind. Sie fanden wenig effiziente Sakkaden zu Zielreizen, die als Konjunktion aus Tiefe und Orientierung definiert waren. Ein Vergleich zu anderen Konjunktionen (ohne Tiefe) wurde nicht durchgeführt. Auch Theeuwes et al. [TAK98] sehen in der Tiefe einfach ein weiteres Merkmal, das keine der 2-D-Position vergleichbare Sonderrolle spielt.

Blaser und Domini [BD02] demonstrieren wiederum spezielle Nacheffekte in der Kombination von Tiefe und Features, die als Hinweise auf die Verwendung dieser Konjunktion zur Bildung von Oberflächen an einer frühen Stelle der visuellen Verarbeitung, einzuordnen in der präattentiven Stufe, gedeutet wird und die spezielle Rolle von Tiefe weiter belegen. Man also zusammengefasst davon ausgehen, dass auch die dritte räumliche Dimension eine besondere Rolle gegenüber den normalen Merkmalen spielt, die jedoch nicht den beiden Dimensionen der retinalen Koordinaten gleichkommt.

3.2.5 Aufmerksamkeit und Dynamik

Aufmerksamkeit wird als ein dynamischer Prozess beschrieben, der gerade die Verarbeitung relevanter Bestandteile vor den weniger relevanten beinhaltet und so eine zeitliche Reihenfolge festlegt. Die Untersuchung und leider auch die Modellierung beschränken sich häufig auf die Verarbeitung statischer Eingabereize.

Diskutiert wird in der Literatur, in welcher Art sich denn Aufmerksamkeit dynamisch verhält, speziell, wie sich der Fokus der Aufmerksamkeit bewegt. Eriksen und Murphy [EM87] untersuchten die Hauptthesen: eine kontinuierliche Bewegung des Fokus, die den zwischen Start und Ziel liegenden Bereich „mitnimmt" im Unterschied zu einem diskreten Wechsel von einem Ort zu einem anderen, ohne den dazwischenliegenden Bereich mit Aufmerksamkeit zu versehen. Weiterhin wurde untersucht, ob die Zeit für einen Aufmerksamkeitswechsel proportional zur Entfernung ist oder eine konstante Zeitspanne benötigt. Sie kamen zu dem Ergebnis, dass es für jede der Alternativen Evidenz gibt und die Frage als ungeklärt zu gelten hat.

Wichtige Experimente zur Zuweisung von Aufmerksamkeit in dynamischen Szenen stammen von Pylyshyn und Mitarbeitern [PS88, PBF+94, Pyl98]. Im Paradigma des *multi object tracking* (siehe auch Abb. 3.5) bestehen die Displays aus einer Anzahl identischer Elemente, von denen einige verfolgt werden sollen. Bevor jedoch die Bewegung einsetzt, werden sie statisch dargeboten. Die Zielreize werden hervorgehoben, indem sie umrandet werden oder aufblinken. Danach setzt die Verfolgungsphase ein, in der sich die Reize unabhängig voneinander in wechselnde Richtungen bewegen. Schließlich wird die Bewegung angehalten und ein einzelnes Element hervorgehoben. Die Versuchsperson muss nun entscheiden, ob dieses Element zu den Zielreizen gehört. Die Ergebnisse zeigen, dass diese Aufgabe mit hoher Effizienz und Genauigkeit gelöst werden kann, sofern die Zahl der zu verfolgenden Zielreize bei maximal vier oder fünf liegt. Die Leistung ist weitgehend unabhängig von der Anzahl der Distraktoren.

Daher muss man tatsächlich von einer parallelen Verfolgung mehrerer Elemente ausgehen. Ein schneller Wechsel fokaler Aufmerksamkeit zwischen den Objekten mit Speicherung der Positionen ist keine brauchbare Alternativerklärung, weil sich die dazu notwendige Geschwindigkeit für die

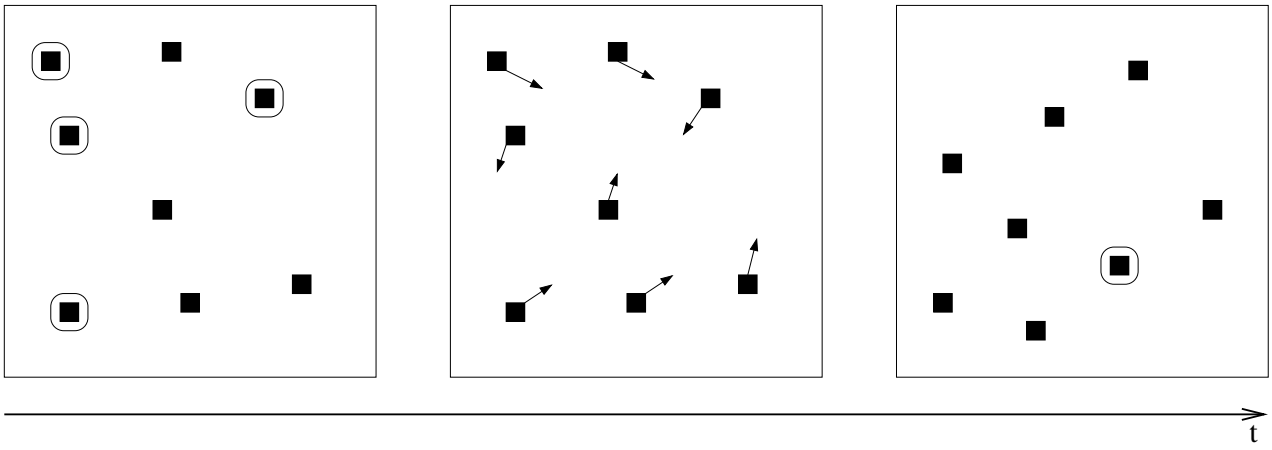


Abbildung 3.5: Schematische Darstellung des Ablaufes eines Experimentes zum *multi object tracking* nach Pylyshyn und Storm[PS88]. Nachdem im statischen Display (links) die Zielelemente hervorgehoben werden, bewegen sich alle Elemente unvorhersagbar (Mitte). Anhand des abschließenden statischen Displays (rechts) ist zu entscheiden, ob das hervorgehobene Element zu den Zielelementen gehört.

Aufmerksamkeitswechsel weit außerhalb der üblichen Schätzungen befindet. Das auch aus diesen Untersuchungen hervorgegangene FINST-Modell wird in Kapitel 3.5.2 genauer vorgestellt.

In Experimenten dieser Art wurde eine zeitweise Überlappung der bewegenden Elemente grundsätzlich vermieden, da sie eine zuverlässige Zuordnung der Elemente verhindern würde. Viswanathan und Mingolla [VM98, VM99] stellten sich nun die Frage, welche Eigenschaften der Elemente das Tracking unter derart erschwerten Bedingungen ermöglichen würden. Die Elemente wurden in ihren Experimenten anhand von Farbe oder Tiefe unterschieden.. Die Tiefe konnte einerseits durch Disparität definiert sein, andererseits durch schattierte Elemente, bei denen in der Überlappung das vordere weiter vollständig sichtbar war, das hintere jedoch verdeckt. Wichtig zu beachten ist, dass sich die Elemente auch in der Tiefe bewegen, die initial wahrgenommene Tiefe also kein Merkmal ist, anhand dessen die Aufgabe ohne dauerhaftes Tracking lösbar wird. Jeder der beiden Tiefenhinweise reicht eindeutig aus, um die Aufgabe zu lösen; Farbe trug nicht zum Tracking bei.

Obwohl sich mehrere bewegte Objekte ohne fokale Aufmerksamkeit verfolgen lassen, wird Aufmerksamkeit benötigt, um selbst einfache Bewegungsmuster erkennen zu können, wie Cavanagh et al. [CLT01] demonstrierten.

3.2.6 Aufmerksamkeit als Schnittstelle von Perzeption und Aktion

Das Vorhandensein von Kapazitätsengpässen in der Verarbeitung visueller Informationen ist keineswegs das einzige Argument für die Selektivität durch Aufmerksamkeit. Vielmehr verlangt das Vorhandensein vielfältiger sensorischer Informationen eine Selektion, wenn es um die Spezifikation und Ausführung von Aktionen geht. Hier können Informationen, die nicht zum relevanten Objekt, auf das sich die Aktion bezieht, zu Fehlern in der ausgeführten Aktion führen.

Man stelle sich ein jagendes Tier vor, das eine Herde oder einen Schwarm von Beutetieren vor sich hat. Jedes einzelne Beutetier weist alle Merkmale auf, die für das jagende Tier von Bedeutung sind. Trotzdem ist es wichtig, zur Spezifikation der motorischen Programme ein einzelnes Tier auszuwählen,

um genau dieses fangen zu können. Orte und Bewegungsparameter weiterer Beutetiere würden die Steuerung behindern. Ähnliches gilt für das Erinnern: wollen wir uns ein bestimmtes Gesicht merken, sollten auf keinen Fall Merkmale momentan naheliegender Gesichter in das Gedächtnis mit eingehen. Diese Beispiele sollen nahelegen, wie wichtig das Unterdrücken von Distraktoren, von ablenkenden Sensordaten für die Ausführung mentaler und motorischer Aktionen ist. Hier wird auch deutlich, wann Probleme entstehen, wenn die Zuordnung von Aufmerksamkeit fehlerhaft war: man denke an das Reagieren auf das Grün der Ampel, das nur den Abbiegern galt oder das taktil gesteuerte Herausgreifen eines einzelnen Stiftes aus einer Sammlung.

Diese Aufgabe der Aufmerksamkeit wird besonders von den Protagonisten der sogenannten späten Selektion ([Bun90, Dun80], s. 3.3.1) hervorgehoben. Sie gehen davon aus, dass Aufmerksamkeit in einer späten Stufe der Verarbeitung operiert. Die Theorien gehen von der weitgehenden Verarbeitung der Reize ohne Aufmerksamkeit aus, die dann als bereits erkannte Objekte selektiert werden. Das Stichwort „*Selection for action*" wurde von Allport [All90] geprägt, der als Grund für die Notwendigkeit von Aufmerksamkeit eben gerade die korrekte Bindung von sensorischen Daten an auszuführende Handlungen ansah.

Einen wesentlichen Einfluss der Handlung auf die Wahrnehmung konnten Tipper et al. [TLB92] in Experimenten zur dreidimensionalen Repräsentation zeigen. Dort determinierte die durchzuführende Aktion die räumliche Repräsentation. Eine besondere Aktion stellt das Abspeichern von Informationen im Gedächtnis dar. Mittels Doppelaufgaben konnten Joliceour und Dell'Acqua [JD99] zeigen, dass gerade das Abspeichern auch kleiner Informationseinheiten im visuellen Arbeitsgedächtnis Aufmerksamkeit benötigt und streng seriell abläuft.

Um die Bedeutung der Handlungssteuerung anhand der Wahrnehmung deutlich zu machen, sollte in diesem Zusammenhang die Theorie des *ecological vision* nach Gibson [Gib79] beachtet werden. Gibson betont die Situierung der Wahrnehmung in Umgebung und Zweckbindung. Wahrnehmung ist für ihn eine Bestimmung von Handlungsmöglichkeiten. So bedeutet die Wahrnehmung einer stabilen, ausgedehnten und flachen Oberfläche die Möglichkeit, sich auf ihr zu bewegen. Dies nimmt Koffka's [Kof35] Idee des Aufforderungscharakters auf, in dem jedes Objekt zu einer Handlung aufruft (eine Frucht sagt „Iss mich"). Diese Theorie wurde im technischen Bereich von den Protagonisten des Aktiven Sehens (Abschnitt 4.2) wieder aufgegriffen.

3.3 Modellierungen natürlicher visueller Aufmerksamkeit

3.3.1 Theoretische Beschreibungen

Einen aktuellen Review zu Modellierungen visueller Aufmerksamkeit findet man bei Itti und Koch [IK01a]. Die Autoren identifizieren fünf zentrale Punkte in der Modellierung visueller Aufmerksamkeit:

- Die lokale Salienz ist kontextabhängig.
- Eine zentrale topographische Karte akkumuliert die lokale datengetriebene Salienzinformation.
- *Inhibition of return* stellt einen zentralen Prozess dar.

- Es gibt einen starken Zusammenhang zwischen Augenbewegungen und verdeckter Aufmerksamkeit.
- Prozesse wie Objekterkennung üben einen starken Einfluss auf die Zuweisung von Aufmerksamkeit aus.

Eine klassische Streitfrage in der Modellierung visueller Aufmerksamkeit liegt in der Bestimmung des Ortes, an dem die Selektion stattfindet. Hierbei stellt sich die Frage, welche Prozesse präattentiv stattfinden und welche Prozesse fokale Aufmerksamkeit brauchen? Die gegensätzlichen Standpunkte der frühen und späten Selektion stellen dabei Extreme dar, zwischen denen viele Kompromisse vorgeschlagen wurden.

Verfechter der frühen Selektion wie Broadbent [Bro58], Treisman [TG80], Eriksen [EY85] und Wolfe [WCF89] beschreiben Aufmerksamkeit als primär räumliche Auswahl eines Bereiches. Nur dieser Bereich kann dann komplexen Prozessen wie einer Erkennung, Klassifikation oder dem Speichern im Gedächtnis zugeführt werden. Als wichtigster Grund für diese Selektion werden Kapazitätsengpässe in der Verarbeitung angegeben.

Theorien der späten Selektion wie von Deutsch und Deutsch [DD63], Duncan [Dun80], Bundesen [Bun90] sowie Shiffrin und Schneider [SS77] siedeln die Selektion erst nach einer weitgehenden Verarbeitung der Reize an. So wird auf höheren Repräsentationen als der rein räumlichen ausgewählt. Hier wird stärker der Handlungsaspekt betont, der eine Auswahl der dafür relevanten Informationen voraussetzt. Der Disput zwischen beiden Theorien dauert an. Die reine Form der späten Selektion konnte zwar experimentell weitgehend widerlegt werden, die rein räumliche Sichtweise der frühen Selektion aber auch. So zeigen aktuelle Theorien meist Aspekte beider Vorstellungen, etwa in der Form, dass die Selektion von der Last des Systems und der Komplexität der visuellen Reize abhängt.

Neben Modellen, die den Effekt der Aufmerksamkeit als reine Selektion in der Form eines Filters modellieren [Bro58], existieren auch solche, die in der Zuordnung von Kapazitäten oder Ressourcen die Arbeitsweise von Aufmerksamkeit sehen, z.B. das *zoom lens*-Modell[EJ86]. Während die Modellierungen sich durchaus unterscheiden, gibt es bisher keine wirklich eindeutige Möglichkeit, eines der beiden Modelle zu widerlegen. Die vorliegenden Ergebnisse sind im Wesentlichen mit beiden konsistent. Langfristig werden am ehesten Ergebnisse der neurowissenschaftlichen Untersuchungen zur Unterscheidung beider Formen beitragen. Auch hier gilt natürlich, dass beide sich nicht grundsätzlich ausschließen, sondern im Prinzip auch innerhalb eines Modelles koexistieren können.

Theeuwes [The93] konzipierte ein Modell visueller Suche, das sich auf räumliche Aufmerksamkeit, frühe Selektion und eine stark eingeschränkte Möglichkeit zu top-down-Einflüssen stützt. Speziell wird angenommen, dass alles, das einmal in den Fokus der Aufmerksamkeit fällt, nur noch datengetrieben beeinflusst wird.

Sperling und Weichselgartner [SW95] bieten eine abstrakte mathematische Beschreibung von Aufmerksamkeit als Folge von diskreten Zuständen (Episoden), deren Übergänge durch eine räumlich-zeitlich separable Funktion definiert ist. Eine einzelne Episode lässt sich dabei als räumliche Funktion beschreiben. Zu den Vorhersagen gehört, dass der Fokus der Aufmerksamkeit sich nicht kontinuierlich bewegt, sondern diskret. Besonderer Wert wird auf die Modellierung der Vorbereitung von Aufmerksamkeit gelegt, wie sie in Cueing-Experimenten gemessen wird.

Chelazzi [Che99] versucht in seinem Modell visueller Suche, die gesammelte Evidenz aus Psychophysik, Neurowissenschaften und Läsionsstudien zu integrieren. Er widmet sich primär der Frage,

inwieweit serielle Prozesse in der Suche eine Rolle spielen. In diesem Modell wird die Suche in komplexen Aufgaben, z.B. der Konjunktionssuche, nicht durch ein serielles Absuchen als Target in Frage kommender Elemente verstanden, sondern als paralleler Wettbewerb zwischen diesen Kandidaten. Dieser dauert umso länger, je schwieriger die Unterscheidung zwischen Zielreiz und Ablenkern ist.

3.3.2 Psychophysische Modelle

Den modernen Modellierungen von Aufmerksamkeit gehen Modelle voraus, die auf einer weitaus geringeren empirischen Datenbasis basieren und sich insofern nur in einigen Kernaspekten in den aktuelleren Arbeiten wiederfinden. Dazu gehören strikt serielle Modelle wie das von Neisser [Nei67] und Vorläufer des Spotlightmodells von Broadbent [Bro58].

Als klassisches Modell visueller Aufmerksamkeit, das mittlerweile jedoch viele Modifikationen erfahren hat, kann die *Feature Integration Theory* von Treisman und Gelade [TG80] gelten. Sie beruht stark auf den Experimenten zur Visuellen Suche (s. 3.2.1). Der Grundgedanke ist dabei, dass einfache Merkmale präattentiv bestimmt werden können, die räumliche Konjunktion von Merkmalen jedoch ohne fokale Aufmerksamkeit nicht herstellbar ist. Es wird dazu von sogenannten *feature maps* (Merkmalskarten) ausgegangen, in denen lokal die Präsenz bestimmter Basismerkmale kodiert werden. Zu diesen Basismerkmalen gehören demnach Farbe, Orientierung und Größe. Die Merkmalskarten sind unabhängig voneinander organisiert, so dass zwar die gleichzeitige Präsenz zweier Merkmale leicht bestimmt werden kann. Die Überprüfung, ob die Merkmale am selben Ort vorkommen, setzt jedoch die Zuweisung fokaler Aufmerksamkeit voraus. Diese fokale Aufmerksamkeit ist immer nötig, um die am selben Ort befindlichen Merkmale zu einem Objekt zu verknüpfen. Durch starke Belastung des Systems entstehen als Fehler die sogenannten *illusory conjunctions*, die sich dadurch ausdrücken, dass Merkmale von benachbarten Objekten miteinander verbunden wahrgenommen werden. Die Ausrichtung des Fokus der Aufmerksamkeit anhand der Master map of attention kann durch einen Hinweis auf einen Ort gelenkt werden (Cueing-Experimente). Er kann sich aber auch sich nach der Präsenz top-down aktivierter Merkmale richten (in der Visuellen Suche). Das Modell wurde inzwischen mehrmals überarbeitet und neueren Ergebnissen angepasst. So hat Treisman [Tre93, Tre98] die Selektionsmöglichkeiten nach einem objektbasierten Ort und einem handlungsrelevanten *object file* (siehe weiter unten) hinzugefügt.

Ebenfalls an Experimenten zur Visuellen Suche richtet sich das *Guided Search*-Modell von Wolfe et al. [WCF89, Wol94, WG96] in seinen verschiedenen Versionen aus. Es entspricht der Feature Integration-Theorie in der Modellierung mehrerer Merkmalskarten, auf denen bereits eine Kontrastbildung zur Hervorhebung singulärer Merkmalsausprägungen stattfindet. Diese werden in eine *master map of attention* gewichtet integriert, die die präattentiv gesammelten Hinweise zur Steuerung des Fokus der Aufmerksamkeit bereitstellt. Die Gewichte werden durch top-down-Kontrolle gesteuert, so dass Merkmale mit bekannt großer Bedeutung hervorgehoben, bzw. andere Merkmale gehemmt werden können. Die Suche erfolgt, indem der Fokus der Aufmerksamkeit sich nach dem Maximum der Mastermap ausrichtet und das dort befindliche Element mit dem Target verglichen wird. Hat man das Target gefunden, erfolgt die entsprechende Reaktion. Andernfalls wird die Position in einer sogenannten *inhibition map* markiert und die Suche richtet sich auf das Maximum der Differenz aus *master map* und *inhibition map*. Findet sich nach gewisser Zeit kein Zielreiz, wird die Suche erfolglos mit der Reaktion „abwesend" abgebrochen.

Die *Feature Integration Theory* und das *Guided Search*-Modell ähneln einander hinsichtlich der Verwendung mehrerer parallel berechneter Merkmalskarten und der Orientierung an der Visuellen Suche. Beide haben sich in ihrer langen Entwicklung gegenseitig stark beeinflusst. Während jedoch die Feature Integration Theory in ihrer eigentlichen Form deutlich unterschiedliche Prozesse für Merkmals- und Konjunktionssuchen vorsieht, operiert im Guided Search-Modell derselbe Mechanismus mit unterschiedlichen Signal-Rausch-Verhältnissen von Zielreizen und Ablenkern.

Die oft verwendete Metapher des Scheinwerfers der Aufmerksamkeit bezeichnet ein Modell, das Eriksen [EH73, EY85, EJ86] aufgrund von Experimenten zu Flankerkompatibilitätseffekten konzipierte. Dieser Scheinwerfer kann kontinuierlich geschwenkt werden, so dass sich der mit Aufmerksamkeit versehene Bereich bewegt und dabei andere Bereiche überstreicht. In gewissen Grenzen ist es laut der Erweiterung als *zoom lens model* möglich, die Größe des Scheinwerfers zu verändern. Dabei wird die Auflösung in diesem Bereich gleichzeitig so verändert, dass die Informationsmenge konstant bleibt. Eine weitere Variante dieses Modells geht von einer Variation in der Auflösung aus, die vom Zentrum des Scheinwerfers nach außen hin abnimmt. Insgesamt ist dieses Modell der verdeckten Aufmerksamkeit recht stark an der offenen Aufmerksamkeit durch Blickbewegungen orientiert.

Moore und Wolfe stellen mit dem *assembly line*-Modell [MW01] eine Alternative zu Modellierungen der seriellen und parallelen Anteile in der Verarbeitung vor. Die Analogie sieht ein Fließband vor, das zwar in einem bestimmten Takt Ergebnisse produziert (also etwa alle 10 Minuten ein Auto), bei dem aber die Durchlaufzeit für ein einzelnes Objekt sehr viel höher ist als diese Taktzeit (im Beispiel könnte es also einen Tag dauern, ein Auto herzustellen). Das Modell kann dazu dienen, verschiedene Effekte, wie den *attentional blink* [RSA92, Ray01] und die *change blindness* [ROC97] zu erklären.

Object files

Eine weitere wichtige Datenstruktur in der Modellierung visueller Aufmerksamkeit neben den Kartenrepräsentationen für Merkmale, Salienz und Inhibition stellen die sogenannten *(preattentive) object files* (OF) dar. Sie wurden von Kahneman und Treisman [KT84, Tre91] als Lösung für die Probleme der Bindung, der Objektkonstanz und der objektbasierten Selektion eingeführt. In vielen Situationen ist die Identität eines Objektes nicht von Anfang an bekannt. Obwohl es seine Größe und Position ändert, besteht kein Zweifel, dass es sich die ganze Zeit um ein und dasselbe Objekt handelt. Informationen, die sich auf dieses Objekt beziehen, können ihm zugeordnet werden, ohne dass seine Identität bereits bekannt ist. Diese Zuordnung soll durch die Verknüpfung der Objekte mit *object files* erreicht werden, die sensorischen Daten zu einem Objekt integrieren. *Object files* werden über ihre momentane Position adressiert und nicht über ihre Identität.

Experimentelle Unterstützung für die Verwendung von Objectfiles demonstrieren Kahneman, Treisman und Gibbs [KTG92]. Der Erkennung eines Objektes durch die Versuchsperson geht in den Experimenten ein weiteres Display voraus. Durch verschiedene experimentelle Manipulationen wird in einigen Durchläufen eine Verknüpfung zu einem bereits dargebotenen Objekt erzeugt, die im Falle der Identität die Erkennung des Objektes beschleunigt. Die Displays konnten dabei statisch, dynamisch und mit sehr unterschiedlichen Zeitabständen präsentiert werden. Diese Objektkonstanz verlangt eine erkennungsunabhängige, präattentive Datenstruktur mit Verweis auf ein (bewegliches) Objekt. Ein existentes Objectfile beschleunigt die Bearbeitung von Objekten und wird mit Vergleich der Objekte mit existierenden Objectfiles erklärt.

Interessanterweise bezieht sich Treisman [Tre98] zur Motivation der Datenstruktur explizit auf die Experimente von Pylyshyn zum *multi object tracking* als Evidenz zur Bindung und Objektkonstanz durch Objectfiles. Der deutliche Unterschied zwischen den FINST-Indizes von Pylyshyn und den Objectfiles von Treisman liegen primär im Informationsgehalt. Während FINST alleine als Verweise auf Orte dienen, enthalten die Objectfiles bereits Merkmalsinformationen. Daher bezeichnen Kahneman et al. [KTG92] FINST als mögliche initiale Phase eines Objectfile, bei der noch keine Merkmalsinformationen verfügbar sind.

Der Frage, welche Informationen in einem Objectfile enthalten sind, widmeten sich Wolfe und Bennett [WB97] mit Experimenten zur visuellen Suche nach Objekten, die aus mehreren Teilen unterschiedlicher Eigenschaften bestanden. Es ließ sich zuerst nachweisen, dass bereits präattentiv eine Einteilung der Szene in visuelle Objekte stattfindet. Diesen Objekten lassen sich primitive Merkmale zuordnen, die auch eine „parallele" visuelle Suche ermöglichen. Die Merkmale werden dabei jedoch nicht im Sinne einer Konjunktion verknüpft, denn obwohl zwar bekannt ist, welche Merkmale zum Objekt gehören, steht nicht fest, ob sie am selben Ort vorhanden sind. Entscheidend jedoch ist, dass die Form des Objektes (*shape* im Gegensatz zu *form*) nicht zu diesen Eigenschaften gehört. Dieser Aspekt wurde durch umfangreiche Experimente verifiziert.

Mit einem Fokus auf das visuell-räumliche Arbeitsgedächtnis stellte Schneider [Sch99] ein Modell visueller Aufmerksamkeit vor. Es sieht eine erste Stufe der Bildung einzelner visuell-räumlicher Einheiten vor, aus denen zu jedem Zeitpunkt jeweils eines selektiert wird, das der zweiten Stufe zugeführt wird. Diese zweite Stufe ist zuständig für Objekterkennung, räumliche Spezifikation für Motorkommandos und das Erzeugen von Objectfiles. Damit ist das Anlegen von Objectfiles im Gegensatz zur Vorstellung von Wolfe [WB97] und Treisman [Tre91] hier ein attentiver Prozess. Zusammen mit dem gerade aktiven gibt es zu jeder Zeit maximal vier Objectfiles.

Raymond [Ray01] demonstrierte kürzlich, dass sich der Effekt des *attentional blink* bereits auf Objektebene abspielt. Es handelt sich also um keinen rein perzeptuellen Effekt handelt. Dies wurde durch das Verschwinden des Effektes bei Darbietung desselben Objektes in unterschiedlicher Darstellungsform demonstriert. Somit wird ein Zusammenhang mit dem Anlegen neuer Objectfiles begründet, der einen Flaschenhals in der Verarbeitung darstellt.

3.3.3 Konnektionistische Modelle

Viele Modelle versuchen die Umsetzung von Aufmerksamkeit auf einer Basis zu modellieren, die dem natürlichen Vorbild entspricht und verwenden Implementationen Künstlicher Neuronaler Netze.

Das Modell von Mozer und Sitton [MS96] besteht aus einer einfachen Objekterkennung, die - im Beispiel - auf Buchstaben trainiert wird und einer *attention map*, deren Aktivation den Zugang der Eingabe zur Objekterkennung reguliert. Das Modell ist darauf ausgerichtet, einzelne psychophysische Befunde, wie Precueing, Crosstalk von Distraktoren sowie den Unterschied zwischen Merkmalssuche und Konjunktionssuche mit sehr einfachen Stimuli zu reproduzieren. Interessant ist, dass anhand des Modells keine absolute Filterung der Eingabe vorgenommen wird, sondern die nicht attendierten Bildbereiche zu einem reduzierten Anteil an der Verarbeitung teilhaben.

Ahmad's VISIT-Modell [AO91, Ahm91] besteht aus einer Reihe von Neuronalen Netzen für unterschiedliche Aufgaben. Dazu gehören ein *Priority Network* zur Bestimmung des Ortes des FOA, ein *Gating Network* zur Ausführung der räumlichen Selektion mit einem runden FOA und ein *Control*

Network als Arbeitsspeicher und zur Beeinflussung des Datenflusses zwischen *Gating Network* und *Priority Network*. Es wurde zur Modellierung der Visuellen Suche und der Berechnung räumlicher Relationen eingesetzt.

Von Hassoumi et al. [HCT] stammt das *Competitive Search* Modell, das sich ebenfalls zum Ziel setzt, die Prozesse bei der Visuellen Suche neuronal abzubilden. Es besteht aus einem zweischichtigen Netzwerkmodell. Die erste Schicht nimmt Eingaben von jeweils einer Merkmalskarte (nicht Bestandteil des Modells) entgegen. Die Eingaben werden an eine zweite Schicht weitergegeben, die zusätzlich zu einer globalen Inhibition inhibitorisch auf die erste Schicht zurückprojiziert. Die Simulationen psychophysischer Experimente enden allerdings bei der Bestimmung von Aktivationswerten. Es findet keine tatsächliche Modellierung des Effektes der attentiven Selektion statt.

Im Unterschied dazu beruht das SERR-Modell (*SEarch via Recursive Rejection*) von Humphreys und Müller [HM93] auf der Gruppierung einfacher Reize, deren Konkurrenz untereinander den wesentlichen Aufmerksamkeitseffekt ausmacht. Diesen Gruppen wird als Ganzes Aufmerksamkeit zugewiesen. Sie können auch als Ganzes inhibiert werden, um die Suche unter den übrigen Gruppen fortzusetzen.

Ein Modell attentiver Selektion anhand oszillatorischer Korrelation, wie sie in Abschnitt 3.2.2 unter dem Stichwort *temporal tagging* beschrieben wurde, stammt von Wang [Wan99]. Es bietet eine Alternative zu klassischen WTA-Modellen, indem auf eine globale Konnektivität verzichtet wird, was in dem Vorteil der Bewahrung räumlicher Relationen resultiert. Durch Modifikation der Parameter kann eine Selektion mehrerer Objekte erreicht werden, die sich zeitlich ablösen. Dabei findet eine implizite einfache Segmentierung der Objekte statt. Eine reale Anwendung des Modells wird jedoch nicht vorgestellt.

Das FeatureGate Modell von Cave [Cav99] besteht aus einer Hierarchie neuronaler Schichten, die die lokale Präsenz von Merkmalen kodieren. Sie reichen von einer vollständigen Repräsentation der Eingabe an der Basis bis hin zu einer nicht-räumlichen, sondern ausschließlich Merkmale anzeigenden Schicht. Innerhalb einer lokalen Nachbarschaft setzten sich Singletons beim Gating durch und werden nach oben weitergereicht. Top-down werden bestimmte Gates geschlossen, deren Merkmale nicht dem Ziel entsprechen. Die Suche wird ähnlich wie in Guided Search über die Inhibition bereits selektierter Elemente vollzogen. Außer der verbreiteten Modellierung von Cueing und visueller Suche gehört FeatureGate zu den wenigen Modellen, die sich auch dem Flankerkompatibilitätseffekt widmen. Es beruft sich dabei jedoch auf die Begrenzung des Effektes durch die Entfernung zwischen Zielreiz und Distraktoren, die so nicht mehr als zutreffend gilt.

Weitere konnektionistische Modelle, die jedoch ihren Schwerpunkt im Computer Vision als in der reinen Modellierung natürlicher visueller Aufmerksamkeit haben, sind in Abschnitt 4.1.1 beschrieben.

3.4 Blickbewegungen

3.4.1 Sakkadische Suppression

Angesichts der hohen Geschwindigkeit, mit der Sakkaden ausgeführt werden, stellt die Wahrnehmung während einer solchen Sakkade ein Problem dar. Die Szene müsste durch die schnelle Umgebung unscharf (sozusagen verschmiert) wirken. Das visuelle System reagiert darauf mit der sogenannten sakkadischen Suppression, die die Wahrnehmung in dieser Zeit unterdrückt. Reize, die während der

Sakkade dargeboten werden, werden nicht weiter verarbeitet [BMR94], Veränderungen der Szene, die während der Sakkade auftreten, werden kaum bemerkt. Dies trägt zu den Kosten bei, die mit einer Sakkade assoziiert sind. Weitere Kosten liegen im Energieaufwand für die Bewegung und darin, dass Bereiche, über die bereits Informationen gesammelt wurden, nicht mehr im Gesichtsfeld liegen. Diese Kosten spielen eine entscheidende Rolle in der Entscheidung, ob eine verdeckte Aufmerksamkeitsverschiebung oder eine offene Blickbewegung durchgeführt werden soll.

Allerdings gibt es Hinweise, dass die Suppression zwar das Bewusstwerden von Informationen, die während der Sakkade aufgenommen werden, verhindert, diese Informationen jedoch unter Umständen dennoch verhaltensrelevant werden [MAJ00]. Schließlich kann die sakkadische Suppression auch als Maskierung der vorherigen Information durch die neuen Reize interpretiert werden [CW78] oder zumindest die fehlende bewusste Wahrnehmung zum Teil dadurch erklärt werden.

3.4.2 Transsakkadisches Gedächtnis

Ein Problem für das visuelle System ist die Aufrechterhaltung eines stabilen Weltbildes über Sakkaden hinweg, da diese eine erhebliche Veränderung des retinalen Bildes und letztlich des wahrgenommenen Szenenausschnitts bewirken. Als Gedächtnisstruktur hierzu kommt das VSTM in Frage, da es im Gegensatz zum ikonischen Gedächtnis nicht anhand retinaler Koordinaten organisiert ist (siehe Abschnitt 2.1.8). Hierzu haben Mitchell und Zipser [MZ01] ein konnektionistisches Modell konzipiert, das die notwendige Speicherung von Orten zum Wiederbesuchen von mittlerweile aus dem Blick geratenen Bereichen erlaubt.

Bei der Untersuchung von Sakkaden während Experimenten zur Visuellen Suche stellten Findlay et al. [FBG01] fest, dass die Programmierung der Sakkade im Wesentlichen von den bei der aktuellen Fixation aufgenommenen Informationen abhängt, was eine interessante Korrespondenz zum Ergebnis von Horwitz und Wolfe darstellt, dass die verdeckte Aufmerksamkeit bei der Visuellen Suche „kein Gedächtnis hat" [HW98]. Allerdings weisen einige sehr schnelle Sakkaden darauf hin, dass unter Umständen doch ein transsakkadisches Gedächtnis genutzt wird, das aber für diesen Kontext auch in der Speicherung einer vorherigen Programmierung mehrerer Sakkaden bestehen kann. Diese Programmierung mehrerer Sakkaden wird auch von McPeek et al. [MSN00] beschrieben, was auf einen zwar attentiven aber nicht singulären Mechanismus verweist.

Rayner [Ray98] fand für Blickbewegungen beim Lesen eine Verlangsamung, wenn die peripheren Informationen nicht dauerhaft präsent waren, was als *preview advantage* interpretiert wird, der eine Verwendung vorher präsentierter Informationen für die Programmierung der Blicksprünge impliziert. Unerwarteterweise stellten Gysen et al. [GVG02] fest, dass eine Verschiebung während der Sakkade leichter für sich bewegende Objekte als für statische Objekte detektiert wird, wobei es keiner Landmarken für die bewegten Objekte bedarf.

3.4.3 Zusammenhang von offener und verdeckter Aufmerksamkeit

Offene und verdeckte Aufmerksamkeit stellen zwar unterschiedliche Mechanismen dar, sie sind jedoch nicht voneinander unabhängig. Der Zusammenhang zeigt sich unter anderem darin, wie ähnlich und zum Teil überdeckend die Bereiche im Hirn sind, die für beide zuständig sind, wie Corbetta [Cor90] nachwies. Schon frühzeitig wurde von Klein [Kle80] die Hypothese formuliert, dass die Bewegung des Fokus der Aufmerksamkeit an einen Punkt Voraussetzung für einen entsprechenden Blicksprung

sei. Erst in den letzten Jahren fand sich dafür jedoch substanzielle experimentelle Basis. Kowler et al. [KADB95] fanden einerseits eine Beschleunigung der Ausführung von Sakkaden zum Fokus der Aufmerksamkeit, konnten andererseits keine Dissoziation von Fokus und Sakkadenziel erreichen. Ditterich et al. [DES00] konnten nachweisen, dass die visuellen Reize im Bereich des Fokus der Aufmerksamkeit zur Spezifikation des Blicksprunges dienen. Auch die Berechnung einer Rückmeldung zur eventuellen Korrektur der Sakkade basiert auf den Informationen innerhalb des vorherigen Fokus der Aufmerksamkeit.

Für Sakkaden wiesen Hooge und Frens [HF00] kürzlich einen der Inhibition of return vergleichbaren Effekt der Hemmung von Blicksprüngen zu kurz zuvor besuchten Orten nach, den sie als *Inhibition of Saccade Return (ISR)* bezeichneten. Allerdings stellten Melcher und Kowler [MK01] bei der Messung von Sakkaden bei Personen, die über mehrere Sekunden eine Szene memorieren sollten, keine Abhängigkeit der Blicksprünge von den bereits besuchten Objekten fest. Vielmehr wirkte die Auswahl zufällig, mit der Ausnahme einer Tendenz zu kleineren Distanzen.

Pomplun et al. [PRSW00] vergleichen ihr Modell für Blickbewegungen bei Visueller Suche erfolgreich mit empirischen Daten. In ihrem Modell werden Blickbewegungen von Aktivationen, die Salienzen wiedergeben und einer von der Aufgabenschwierigkeit abhängige Bereichsgröße beeinflusst. In einer Strategie, die der von Guided Search ähnelt, werden die so gebildeten Bereiche anhand ihrer Auffälligkeit in einen Scanpath geordnet und fovealisiert. Im Gegensatz zu den vorherigen Daten liegt den Sakkaden hier eine Strategie zugrunde, die Planung und Gedächtnis impliziert.

3.5 Offene Fragen

Die Untersuchung natürlicher visueller Aufmerksamkeit ist jedoch ein sehr aktives Gebiet, das keineswegs als abgeschlossen gelten kann. Auch wenn es also Übereinstimmungen bezüglich verschiedener grundlegender Aspekte gibt, sind viele Fragen noch zu beantworten. Über den vorher dargestellten Teil hinaus befassen sich aktuelle Untersuchungen zur Aufmerksamkeit unter anderem mit den folgenden Fragen.

3.5.1 Einheit der Selektion - Raum oder Objekt

Die zuvor beschriebenen Modelle der Aufmerksamkeit verstehen visuelle Aufmerksamkeit als räumlich verteilt. Sie nehmen also an, dass ein bestimmter Teil des retinalen Bildes selektiert wird. Jedoch kommen auch andere Einheiten in Frage, die die vorgenommene Selektion beschreiben können, speziell Merkmale und Objekte.

Die sogenannte featurebasierte Selektion nach Shih und Sperling [SS96] drückt eine Einschränkung der Verarbeitung auf bestimmte Merkmale aus. Diese wird meist als datengetrieben angesehen, so dass durch Weltwissen nicht beliebige Merkmale, sondern primär bestimmte Merkmale, etwa eine Farbe oder eine Orientierung verarbeitet wird. Die experimentelle Evidenz dazu wird jedoch meist so interpretiert, dass das ausgewählte Merkmal dazu dient, den räumlichen Bereich zu definieren, der dann selektiert wird. Die Evidenz für eine rein merkmalsbasierte Selektion ist eher spärlich.

Evidenz für objektbasierte Selektion fand sich dagegen für Aufgaben, bei denen zwei Aspekte beurteilt werden sollten, die sich bei gleicher räumlicher Distanz entweder auf demselben Objekt befanden oder auf zwei getrennten Objekten befanden. Baylis und Driver [BD93] konnten zeigen, dass

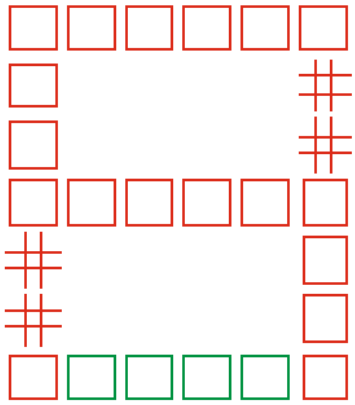


Abbildung 3.6: Reiz, wie er von Hübner und Backer [HB99] verwendet wurde, um zwei Objekte am selben Ort darzubieten. Entlang der Farbdimension ergibt sich ein A, während hinsichtlich der Form ein S erscheint. In diesem Beispiel haben beide Buchstaben 18 Elemente gemein, während nur jeweils vier Elemente zu nur einem der beiden Buchstaben gehören.

die Aufgabe für ein einzelnes Objekt leichter zu lösen war. Kramer und Jacobson [KJ91] verwendeten eine Variante des Flankerkompatibilitätseffektes und zeigten, dass dieser durch die Einteilung des Displays in einzelne oder mehrere Objekte stark modifiziert wurde. Auch ließ sich zeigen, dass sich der sogenannte halbseitige Neglect (eine krankhafte Störung der Aufmerksamkeit, bei der grundsätzlich eine Seite vernachlässigt wird, obwohl sie wahrnehmbar ist) nicht alleine auf den Raum, sondern auch auf Objekte bezieht [BT94].

Vecera und Farah [VF94] konnten die Existenz von objekt- und raumbasierter Aufmerksamkeit von experimentellen Variationen abhängig machen. Sie folgerten daraus, dass die Aufgabe den Selektionsmechanismus bzw. die Einheit der Selektion determiniert. Einen anderen Kompromiss zur Lösung des Disputes um räumliche oder objektbasierte Selektion schlugen Kramer et al. [KWW97] vor. Danach könnte man die Resultate von Vecera und Farah als objektbeeinflusste räumliche Selektion ansehen. In dieser Vorstellung bleibt zwar der Ort die Einheit der Selektion; der genaue Raum, der selektiert wird, muss jedoch keineswegs eine homogene Fläche sein, sondern wird durch Gruppierung und Objekteigenschaften beeinflusst (*grouped array hypothesis*).

Das einfache Scheinwerfermodell kann jedenfalls diese Daten nicht erklären, jedoch wurde auch nicht gezeigt, dass es raumunabhängige Selektion gibt. In eine ähnliche Richtung weisen auch Ergebnisse von Kim und Cave [KC01], die zeigen konnten, dass der Gruppierung einiger Reize eine entsprechende räumliche Aufmerksamkeitsverteilung folgte.

Allerdings wiesen bereits Baylis und Driver [BD93] auf das Problem hin, dass sich Ort und Raum nicht perfekt trennen lassen, da der Raum auch die Objekte definiert. Hübner und Backer [HB99] stellten jedoch Reize vor, die die Konfundierung von Objekt und Ort aufhoben. Sie verwendeten dazu Reize, die sich aus einer Anzahl kleinerer Elemente zusammensetzten, wobei diese in Form einer 8 angeordnet waren. Die Reize definierten sich nun durch die Eigenschaften der kleinen Elemente, die hinsichtlich Farbe und Form variiert wurden. Durch unterschiedliche Variationen entlang beider Dimensionen konnten innerhalb der dargestellten 8 zwei verschiedene Reize eingebettet sein, die zum allergrößten Teil aus denselben Elementen zusammengesetzt waren (s. Abb. 3.6).

Den Versuchspersonen (für die jeweils mehrere Buchstaben unterschiedlichen Reaktionen zugeordnet waren) wurde vor jedem Durchgang angegeben, auf welche Dimension (Farbe oder Form) sie achten sollten. Es ergab sich ein Konsistenzeffekt, der dem in Flankerkompatibilitätseffekten entsprach. Somit wurde nachgewiesen, dass beide Elemente identifiziert wurden und eine Selektion erst auf der hohen Ebene der Objektidentitäten stattfand. Jegliche Selektion an früherer Stelle hätte nämlich zur Folge haben müssen, dass nur noch der relevante Buchstabe zu erkennen war, weil dem Distraktor ein großer Teil an Elementen gefehlt hätte. Damit wurde die Existenz einer rein merkmalsbasierten Objektselektion, die nicht von räumlichen Aspekten beeinflusst wird, nachgewiesen.

Eine andere Methode zur Präsentation zweier Objekte am selben Ort verwendeten Blaser et al. [BPH00]. Sie überlagerten zwei Gaborpatches unterschiedlicher Orientierung, Frequenz und Sättigung, die sich im Unterschied zum zuvor genannten Experiment hinsichtlich ihrer Merkmale veränderten. Die Versuchspersonen waren in der Lage, die Objekte getrennt voneinander zu „verfolgen". Bei Beurteilungen zeigten sich Vorteile für Beurteilungen desselben Objektes gegenüber Beurteilungen unterschiedlicher Objekte, was eine objektbasierte Selektion demonstriert, die nicht mit dem Ort konfundiert ist.

Zusammen mit der von Tipper und Weaver [TW98b] gesammelten Evidenz zu *inhibition of return* und halbseitigem Neglect kann man zusammenfassend sagen, dass Selektion nicht nur anhand einer Einheit (Raum) möglich ist, sondern auch an späteren Verarbeitungsstufen, an denen bereits objektbezogene Informationen zur Verfügung stehen, stattfinden kann.

Ein Modell, das raum- und objektbasierte Selektion integriert, stellt Logan [Log96] vor. Es setzt sich aus einem System zur Gruppierung nach Nähe und einem Selektionsmechanismus zusammen, wobei die Selektion aus dem Ergebnis der Gruppierung selektiert. Indem er sich mit der Form des FOA und der Selektion innerhalb des FOA befasst, kommt Logan zu einer Integration von objekt- und raumbasierter Selektion.

Luck und Vogel [LV97] erkannten auch in der Untersuchung des visuellen Arbeitsgedächtnisses eine objektbasierte Struktur. Bei der Memorierung einfacher Features ergab sich eine Kapazitätsgrenze, die bei vier Elementen liegt und unabhängig von gleichzeitiger Beanspruchung des sprachlichen Arbeitsgedächtnisses und der Darbietungszeit ist. Diese Grenze von vier bezieht sich aber nur auf die Anzahl der Objekte, die es zu memorieren gilt. Sie gilt aber nicht für die Anzahl der Merkmale. Dies wurde anhand einer Aufgabe nachgewiesen, bei der vier Objekte jeweils vier relevante Eigenschaften aufwiesen, so dass insgesamt 16 Merkmale eine Rolle spielten. Hier zeigte sich eine mit der einfachen Bedingung vergleichbare Leistung der Versuchspersonen. Es kann davon ausgegangen werden, dass zwar nicht mehr als vier Objekte im visuellen Arbeitsgedächtnis gespeichert werden können, diese jedoch mit jeweils mindestens vier Merkmalen zur Verfügung stehen. Kritik an der objektbasierten Deutung der Ergebnisse kommt jedoch von Wheeler und Treisman [WT02], Xu [Xu02] und Saiki [Sai03].

3.5.2 Jenseits des „Spotlights"

Gegenüber der üblichen Vorstellung von einem einzigen Fokus der Aufmerksamkeit (sei er raum- oder objektbasiert), dessen Ort sich verändert, gibt es auch Beschreibungen, die von der Möglichkeit mehrerer solcher Foki oder einer kontinuierlichen räumlichen Verteilung ausgehen. Zugunsten einer kontinuierlichen Ressourcenverteilung oder einer kleinen Anzahl von gleichzeitig selektierbaren Orten

oder Objekten wird die Hypothese einer rein seriellen Verarbeitung am Fokus der Aufmerksamkeit aufgegeben.

Kramer und Hahn [KH95] beschreiben etwa Experimente, bei denen durch Cues zwei Positionen markiert wurden. An diesen Positionen mussten später dort erscheinende Zeichen, die Zielreize, verglichen werden. Zwischen den Zielreizen wurden jedoch gleichzeitig Ablenker dargestellt. Die Ablenker hatten dann einen Effekt, wenn sie per Onset dargeboten wurden, jedoch nicht, wenn die Darstellung per Offset erfolgte. Diese experimentelle Manipulation soll den Effekt des *attentional capture* [FRW94] vermeiden, der die automatische Anziehung von Aufmerksamkeit durch Onset, also durch plötzliches Erscheinen des Reizes bezeichnet.

Kramer und Hahn folgern, dass es mehrere (mindestens zwei) Foki der Aufmerksamkeit geben muss, die unter Umständen unabhängig voneinander räumlich positioniert werden können. Dazu werden sehr vorsichtige Annahmen über die mögliche Geschwindigkeit des Fokus der Aufmerksamkeit verwendet. Diese leiten sich aus Experimenten zur Visuellen Suche und zusätzlichen Experimente, die eine Bewegung des FOA voraussetzten, ab. Mit diesen Annahmen lassen sich die Leistungen in den Experimenten nur erklären, indem man annimmt, dass mehrere Foki zur Verfügung stehen, die individuell ausgerichtet werden können.

Als Ergänzung zu solchen Experimenten, die das Aufteilen des Fokus für eine sehr kurze Zeit belegen, konnten Müller et al. [MMGH03] anhand von elektrophysiologischen Daten das Vorhandensein von mindestens zwei getrennten Foki der Aufmerksamkeit für mehrere Sekunden zeigen.

Johnston et al. [JMR96] stellen als Modifikation des Fokus der Aufmerksamkeit ein zweiteiliges Selektionsmodell vor, das jedoch denselben Selektionsmechanismus an zwei unterschiedlichen Stufen der Verarbeitung ansiedelt. Zum einen in der Selektion der Eingabestimuli und zum anderen als „zentrale Aufmerksamkeit", die eher handlungsbezogen operiert.

Das bekannteste Modell in diesem Bereich stammt von Pylyshyn und wird als FINST-Modell (*FINgers of INSTantiations*) bezeichnet [PS88, PBF+94, BP97c, SP00]. Es beschreibt eine kleine Anzahl (etwa vier oder fünf) von Zeigern auf bewegliche Objekte, durch die diese Objekte bevorzugt verarbeitet werden. Der Zeiger ist dabei durchaus als Zeiger im Sinne der Computerdatenstruktur zu verstehen [PBF+94]. Eine wichtige Eigenschaft der Zeiger ist, dass sie am Element, auf das sie sich beziehen, „haften" (*sticky*), dem Element also prinzipiell bei Bewegung folgen. Die Notwendigkeit für einen solchen Mechanismus wird durch relevante Aufgaben, die ein visuelles System zu lösen hat, begründet. Zu diesen gehört die Berechnung einfacher räumlicher Relationen wie Kollinearität oder Enthaltensein.

Die experimentelle Evidenz für die FINST-Hypothese umfasst mehrere Experimentalparadigmen. Als erstes wäre die Dichotomie von *subitizing* gegenüber *counting* zu nennen [TP94], also dem Unterschied beim Zählen einer kleinen Anzahl von Elementen gegenüber einer größeren Zahl. Der erste Unterschied zeigt sich in der Dauer und der Fehlerrate. Letztere ist bei bis zu vier Elementen vernachlässigbar und steigt dann an. Der Zeitbedarf liegt bei bis zu vier bis fünf Elementen auf konstant niedrigem Niveau und steigt erst von da an ungefähr linear, was einen grundsätzlichen Wechsel in der Verarbeitung weniger Reize gegenüber vielen Reizen anzeigt, nicht eine rein quantitative Veränderung. Das *subitizing* funktioniert im übrigen genau so lange, wie zur Aufnahme der Elemente keine fokale Aufmerksamkeit benötigt wird, sie also sich zum Beispiel nicht durch eine Konjunktion von Merkmalen definieren. Für den Unterschied von *counting* und *subitizing* konnten Sathian et al.

[SSP+99] einen signifikanten Unterschied hinsichtlich der neuronalen Aktivation nachweisen.

Auf dieselbe Grenze von vier bis fünf Elementen weisen auch die in Kap. 3.2.5 beschriebenen Experimente zum *multi element tracking* [PS88, SP00] hin, die eine gleichzeitige Verfolgung einmal hervorgehobener Elemente unter lauter identischen Elementen beinhalten und bis zu dieser Anzahl von Elementen von den Versuchspersonen sehr gut gelöst werden können, aber nicht darüber hinaus. Dabei ist zu beachten, dass die Objekte sich alle unabhängig voneinander bewegten, sich allerdings nicht berühren durften. Alternativerklärungen, die auf ein schnelles serielles Überprüfen der hervorgehobenen Elemente hinauslaufen, konnten durch Analyse der dazu notwendigen Geschwindigkeit des Fokus der Aufmerksamkeit ausgeschlossen werden.

Die Bedingung, dass sich Objekte nicht verdecken dürfen, wurde in [SP99] insofern eingeschränkt, als (nicht sichtbare) Verdecker eingeführt wurden, hinter denen die zu verfolgenden Objekte von Zeit zu Zeit verschwanden. Die bekannte Verfolgungsleistung wurde jedoch solange beibehalten, wie es visuelle Hinweise gab, die die Verdeckung plausibel machten, d.h., solange die Objekte kontinuierlich verschwanden und wieder auftauchten. Wurden die Objekte jedoch (entlang identischer Pfade) plötzlich entfernt und wieder eingesetzt, ergab sich eine drastisch verschlechterte Verfolgungsleistung.

Experimente zur Integration mehrerer Cues von Burkell und Pylyshyn [BP93, BP97c] zeigten, dass es möglich ist, auf mehrere einmalig hervorgehobene Elemente direkt zuzugreifen, als ob sie die einzigen präsenten Elemente wären. Die Zahl zur Verfügung stehender Elemente musste bei über drei liegen.

Eine weitere Reihe von Experimenten beruht auf der Line-Motion-Illusion, vorgestellt von Hikosaka et al. [HMS93]. Die Illusion zeigt sich in Experimenten, in denen nach Fixation eines Punktes zuerst ein Cue dargeboten wird, der die Aufmerksamkeit anzieht. Nach einer gewissen Zeit (ISI, *inter stimulus interval*) wird eine Linie so dargeboten, dass einer ihrer Endpunkte dem Ort des Cues entspricht. Die Versuchspersonen berichten durchweg von dem Eindruck, dass die Linie von diesem Ort aus „gezeichnet" wurde. Diesen Effekt erzielten Fisher et al. [FSP93] auch mit einer größeren Anzahl von Cues. In weiteren Experimenten [SFP98], in denen eine Auswahl der Cues zusätzlich hervorgehoben wurde, konnten alternative Erklärungen zur Annahme mehrerer visueller Indizes ausgeschlossen werden.

Bemerkenswerterweise hat die umfangreiche Evidenz für dieses Modell nicht Einzug in sonstige Modellierungen visueller Aufmerksamkeit gehalten. Die Gruppe um Pylyshyn ihrerseits verwendet das Modell als Grundlage für weitergehende Theorien visueller Aufmerksamkeit im Zusammenhang mit anderen kognitiven Fähigkeiten [Pyl99, Pyl00].

Neben den beschriebenen Vorstellungen finden sich schließlich noch solche Modelle, die eine mehr oder weniger kontinuierliche Verteilung von Ressourcen annehmen. LaBerge et al. [LCWB97, LB89] etwa beschreiben Aufmerksamkeit als Effekt einer Aktivitätsverteilung. Unterschieden wird zwischen einer dauerhaften, räumlich verbreiteten Verteilung vorbereitender Aufmerksamkeit und dem kurzzeitigen Öffnen eines „Kanals" selektiver Aufmerksamkeit, wobei die Dauer bis zum Öffnen eines Kanals nicht von räumlichen Relationen, sondern alleine vom Ausmaß der dort vorhandenen Aktivität abhängt. Dieses Modell bewährt sich im Vergleich mit dem klassischen singulären Fokus bei Aufgaben, die mehrere Aufmerksamkeitsverschiebungen aufweisen.

Heslenfeld et al. [HKKM97] stellen ein Modell vor, das die Aspekte mehrerer neurobiologisch motivierter Modelle räumlicher Aufmerksamkeit zu integrieren versucht. Aufmerksamkeit entspricht

dort einer Verteilung von Energie auf unterschiedliche Merkmalsrepräsentationen. Attendierte Reize erfahren also in zumindest einer der Repräsentationen eine erhöhte Aktivation, die sich für räumliche Aufmerksamkeit auch in den Experimentaldaten zeigt. Die Modellvorhersagen treffen jedoch für nicht-räumliche Aufmerksamkeit nicht zu.

Kapitel 4

Computerimplementationen visueller Aufmerksamkeit

Computermodelle visueller Aufmerksamkeit zeichnen sich durch unterschiedliche Techniken und Ziele aus. Der Modellierung durch Neuronale Netze stehen klassische Filteroperationen aus der Bildverarbeitung gegenüber. Dem Anspruch, ein Modell von Aufmerksamkeit zu implementieren, steht eine konkret zu lösende Aufgabe für ein Vision-System gegenüber. Im Vergleich zu den vorher diskutierten Modellen natürlicher Aufmerksamkeit weisen diese Modelle eine Konkretisierung auf. Angenommene Operationen müssen durchführbar sein und man kann mit dem System praktische Experimente auf realen Daten ausführen. Auf der anderen Seite verliert man natürlich auch einiges an Freiheit. Konkrete Hardware und Laufzeitbeschränkungen machen bestimmte Lösungen weniger praktikabel und führen zu Kompromissen hinsichtlich der Korrespondenz zur natürlichen Aufmerksamkeit. Es wird im folgenden zwischen Modellen verdeckter Aufmerksamkeit und Systemen des Aktiven Sehens unterschieden, auch wenn diese Unterscheidung nicht immer eindeutig vorzunehmen ist.

4.1 Computermodelle verdeckter Aufmerksamkeit

Das erste wichtige Aufmerksamkeitsmodell wurde Mitte der 80er Jahre von Koch und Ullman vorgestellt [KU85] und weist bereits viele Bestandteile auf, die sich auch noch in aktuellen Modellen finden. So wird präattentiv eine Anzahl von Merkmalskarten oder Featuremaps berechnet, auf denen inhibitorische Verbindungen für die Verstärkung von Kontrasten sorgen. Diese werden in einer Salienzkarte integriert, auf der ein WTA-Prozess operiert, um so den Szenenausschnitt zu bestimmen, der für fokale Aufmerksamkeit ausgewählt wird. Dieser Ort wird nach der attentiven Verarbeitung in einer Inhibitionskarte markiert, die wiederum hemmend auf die Salienzkarte zurückwirkt.

Das Modell orientiert sich stark an der Modellierung menschlicher visueller Aufmerksamkeit, speziell der Feature-Integration-Theory [TG80]. Abb. 4.1 zeigt die Architektur des Modells von Koch und Ullman. Es handelt sich hier um ein Modell der computational intelligence, das zum Zeitpunkt seiner initialen Veröffentlichung noch nicht implementiert war. Implementationen wurden erst später vorgestellt.

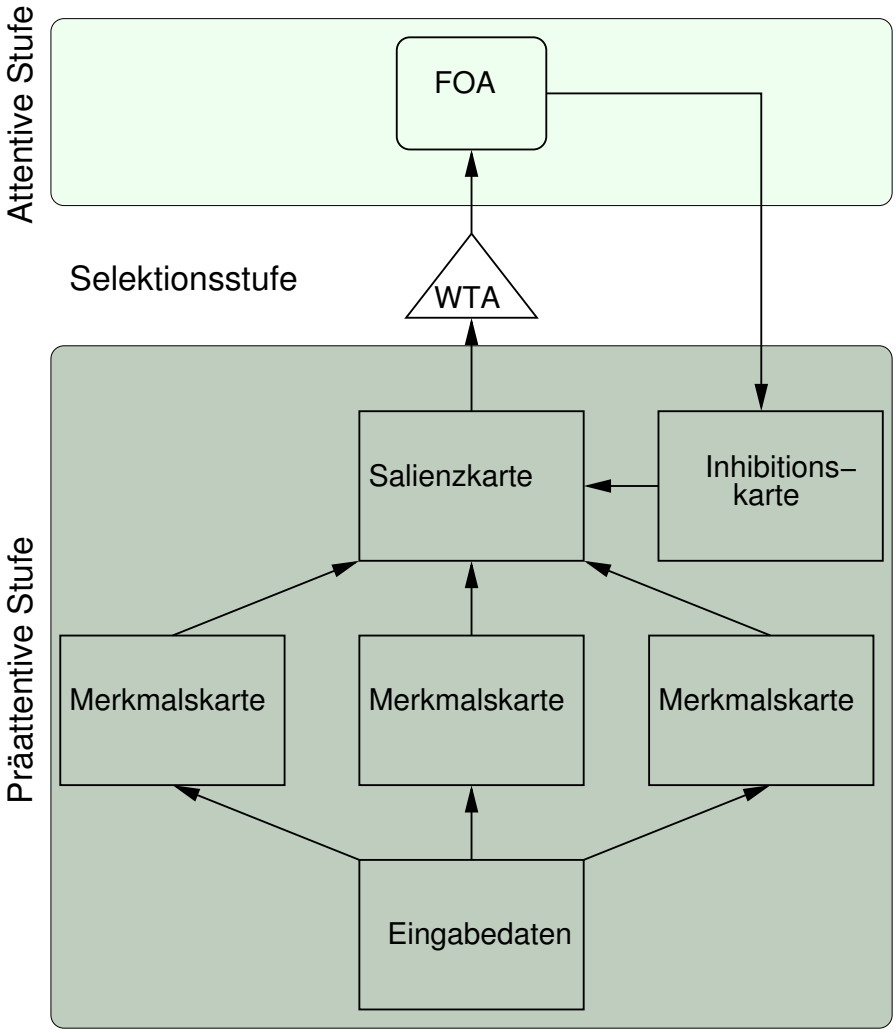


Abbildung 4.1: Architektur des klassischen Aufmerksamkeitsmodells von Koch und Ullman [KU85]

4.1.1 Konnektionistische Modelle

Transformation des selektierten Bereiches

Ein Schwerpunkt der sogenannten konnektionistischen Modelle, die vor allem auf neuronalen Repräsentationen und Verarbeitungen basieren, liegt auf der Transformation eines Bildbereiches in einen Referenzrahmen als zentrale Repräsentation, um eine translations- und größeninvariante Darstellung des ausgewählten Bereiches zu erhalten. Tsotsos und Mitarbeiter [CT92, Tso93, TCW+95] verwenden dazu einen sogenannten *inhibitory beam*, der als WTA-Prozess in einer Multiskalenrepräsentation von der höchsten Ebene in die niedrigeren Ebenen fortgesetzt wird. Dabei wird in den unteren Schichten immer nur noch ein lokaler WTA-Wettbewerb unter denjenigen Einheiten durchgeführt, die mit dem Gewinner der oberen Schicht verbunden sind. Eine Umsetzung des Modells von Tsotsos in einem Robotikkontext zeigen Livingstone und Spacek [LS96]. Hierbei wird auf verhaltensorientierter Basis Suche und Verfolgung von Objekten anhand von Auffälligkeitsmechanismen durchgeführt.

Olshausen et al. [OEA93, OAV95] verwenden sogenannte *routing circuits*, die durch Kontrollneuronen den Datenfluss in höhere Regionen so steuern, dass räumliche Relationen in der Repräsentation erhalten bleiben. Somit kann eine positionsinvariante und in der späteren Version sogar größeninvariante Repräsentation der Selektion erreicht werden.

Im sogenannten SCAN-Modell von Postma [Pos94] werden zwei unterschiedliche Netzwerke verwendet. In der ersten Stufe findet ein Routingnetzwerk Verwendung, das auf mehreren Schichten jeweils aus einer Anzahl sogenannter *gating lattices* besteht, die einen WTA-Prozess ausführen. So wird schließlich ein einzelnes Muster ausgewählt, das an ein Klassifikationsnetzwerk weitergeleitet wird. Datengetrieben kann durch Angabe von erwarteten Mustern der Routingprozess beeinflusst werden. Ansonsten besteht Ähnlichkeit zu den Modellen von Olshausen und Tsotsos [OAV95, Tso93].

Da alleine „aufmerksam zu sein" keine Aufgabe ist, für die üblicherweise ein Bildverarbeitungssystem erstellt wird, finden sich viele Systeme, die Aufmerksamkeit als eine Komponente eines größeren Systems modellieren und weitere Aufgaben wie Tracking oder Objekterkennung integrieren. Unter Verwendung eines Fuzzy ART-Netzwerks zeigen Pessoa et al. [PE99] eine mögliche Integration von Objekterkennung und Aufmerksamkeit, wobei datengetriebene und modellgetriebene Einflüsse auf die Generierung eines Scanpaths umgesetzt werden.

Die Modelle von Deco und Hamker [Dec00, Dec01, Ham00] zeichnen sich dadurch aus, dass versucht wird, die Lokalisation und Erkennung von Objekten in einem einzigen neuronalen Prozess zu vereinen. Beide Probleme werden dabei als komplementäre Suche angesehen, bei der Identität oder Ort bekannt sind und die jeweils andere Information gesucht wird.

Ein Modell, das Aufmerksamkeit mit einer Lernkomponente und einer Verhaltenskomponente verknüpft, die jeweils auf unterschiedlichen Typen Neuronaler Netze beruhen, stellen Tani et al. [TYN97] vor. Die Hauptaufgabe der visuellen Aufmerksamkeit liegt bei ihnen in der Umschaltung zwischen zwei Verhaltensmodulen.

Einen anderen Schwerpunkt setzt das *dynamic relevance*-Modell von Baluja und Pomerleau [BP97a]. Hier wird versucht, durch ein Neuronales Netz eine Vorhersage für die Bildeingabe zu erreichen. Abweichungen von dieser Vorhersage werden als auffällig bewertet und entsprechend mit Aufmerksamkeit versehen. Das Modell fand Anwendung im Robotiksystem ALVINN [BP97b] und weiteren Anwendungskontexten [Bal98]. Lee und Lee [LL00] stellen ein Modell vor, das anhand von

Top-down-Informationen eine attentive Selektion der verwendeten Features durchführt.

Eine echtzeitfähige Umsetzung des FeatureGate-Modells von Cave [Cav99] (siehe Abschnitt 3.3.3) demonstrieren Stasse et al. [SKC00]. Das Modell verdankt seine Geschwindigkeit einerseits dem parallelen Einsatz mehrerer Rechner, andererseits der Verwendung einer log-polaren Repräsentation, die die ortsvariante Auflösung der Retina und die damit verbundene Datenreduktion ausnutzt.

Zuordnungen zwischen Modulen eines technischen Systems und Hirnarealen, die vergleichbare Aufgaben lösen sollen, werden oft formuliert. Hier ist jedoch das Wissen über die Struktur der Hirnareale noch so neu und im Fluss, dass das Ziehen derartiger Parallelen verfrüht und noch nicht ausreichend begründet erscheint. Weitere konnektionistische Komponenten sind in einigen der in Abschnitt 4.2.2 vorgestellten Aktiven Sehsysteme enthalten.

4.1.2 Filtermodelle

Im Gegensatz zu den konnektionistischen Modellen steht für Filtermodelle die Bestimmung der interessanten Bildbereiche durch Implementation entsprechender Merkmale im Vordergrund.

Milanese [Mil93] unterscheidet mehrere Kartenrepräsentation in seinem Modell. Merkmalskarten geben die Präsenz einfacher Eigenschaften, in diesem Fall Orientierung, Krümmung, Kanten und Farbkontrast wieder. Deren Daten werden in *conspicuity maps*, die durch weitere Filteroperationen aus den Merkmalskarten entstehen, hinsichtlich ihrer Auffälligkeit bewertet. Sie werden in der zentralen *saliency map* integriert, die die Auffälligkeit jeden Ortes beschreibt. Die *saliency map* entsteht durch ein Relaxationsverfahren, das die Herausbildung kompakter und homogener Regionen, fördert. Aus ihr entsteht durch Binarisierung schließlich die *attention map*. In einer erweiterten Version gehen zusätzlich top-down-Informationen ein, die die Suche nach interessanten Bildbereichen beeinflussen [MWG+94].

In der Gruppe um Koch werden, ausgehend vom klassischen Modell [KU85], mehrere Entwicklungen fortgeführt. Itti et al. [IK00, Itt00, IK01b] versuchen dabei, allgemeine Strategien zur Integration vieler Merkmalskarten zu entwickeln und damit ein dem natürlichen Vorbild entsprechendes Suchverhalten zu erzielen. Das Modell setzt vor allem auf Bildpyramiden und *center surround* Mechanismen zur Bewertung des Kontrastes. Als Besonderheit werden bereits innerhalb der Merkmale Kontraste gebildet und innerhalb iterativer Verfahren zur Maximumssuche bereits Inhibition verwendet. Miau et al. [MPI01, MI01] untersuchten besonders das Zusammenspiel von Objekterkennungsmechanismen mit visueller Aufmerksamkeit. Insbesondere sollen dazu modellgetriebene Einflüsse in die Modellierung eingehen.

Kürzlich wurde dieses Modell ausführlich mit dem Verhalten menschlicher visueller Aufmerksamkeit verglichen. Dazu haben Parkhurst et al. [PLN02] Bilder verschiedener Art (Fraktale, Innenraum, Gebäude, Natur), die in einer Blickbewegungsstudie verwendet wurden, dem Modell dargeboten. Sie untersuchten, inwieweit die Positionen der Fixationen mit Orten hoher Auffälligkeit übereinstimmten. Sie fanden für alle Versuchspersonen und jeden Typ von Bildern einen signifikanten Zusammenhang. Für die initiale Fixation war der Zusammenhang am stärksten, blieb jedoch auch darüber hinaus signifikant. Auch war die Stärke des Zusammenhangs abhängig vom Bildtyp. Der starke Einfluss für Fraktale und der schwache für Innenräume wurde durch den zusätzlichen Einfluss modellgetriebener Aufmerksamkeit erklärt. Auch die Bedeutung der einzelnen Merkmale (Farbe, Intensität, Orientierung) ließ sich für die Bildtypen unterscheiden. Einen Teil der fehlenden Korrespondenz konnten die

Autoren über eine zusätzliche Gewichtung der Zentralität der Positionen erreichen, nachdem zu beobachten war, dass die menschlichen Fixationen einen starken Trend zur Bildmitte zeigten. Insgesamt kann dies auch als deutlicher Verweis auf die Bedeutung datengetriebener Aufmerksamkeit bei der Beobachtung natürlicher Szenen gewertet werden.

4.1.3 Weitere Modelle

Den Vorschlag, die Aufteilung in präattentive und attentive Verarbeitung in die technische Bildverarbeitung zu übernehmen, wurde von Leavers [Lea94] unterstützt. Als präattentiv werden von ihm Mechanismen angesehen, die rein datengetrieben arbeiten und unabhängig von Systemzustand und Eingabedaten eine konstante Zeit brauchen. Es findet dort eine starke Orientierung an Ergebnissen der Psychophysik statt.

Von Colombo et al. [CRD94, CRD96] wird die Verwendung eines log-polaren Sensors demonstriert, der von vornherein eine ortsvariante Auflösung liefert und so durch Fixation eines Bildbereiches eine entsprechende Auswahl bestimmt. Die weitere Verarbeitung teilt sich in einen parallelen Zweig, der für das ganze Bild Features in Bildpyramiden organisiert. Auf ihnen operiert ein WTA-Prozess, während seriell zur Erkennung Objektteile mit gespeicherten Vorlagen verglichen werden.

Balkenius und Hulth [BH99] betonen die Sichtweise des *selection for action*, also der Auswahl von Sensordaten zur Spezifikation von Handlungen im Gegensatz der Selektion zur Reduktion des Berechnungsaufwandes. Das Aufmerksamkeitsmodell ist in ein System zur Steuerung eines mobilen Roboters integriert und zeigt als eines der wenigen sowohl verdeckte als auch offene Aufmerksamkeit. Die Autoren weisen jedoch selbst darauf hin, dass dem Aufmerksamkeitsmechanismus wichtige Aspekte wie *inhibition of return* fehlen.

Yeshurun [Yes97] argumentiert für die Verwendung von Aufmerksamkeit in Systemen des Computer-Sehens und stellt als einen Baustein einen generellen Symmetriedetektor vor.

Eines der wenigen Modelle, die explizit eine objektorientierte Form der Aufmerksamkeit (s. 3.5.1) umsetzen, wurde von Fellenz [FH96, Fel97] vorgestellt. Als erster Schritt fand dort eine Segmentierung durch einen Relaxationsprozess und eine Diffusion statt, die Kandidaten für die attentive Selektion lieferte. Die Salienz wurde jedoch unabhängig von dieser Segmentierung berechnet und geht in einen üblichen WTA-Prozess mit Markierung in einer Inhibitionskarte ein. Hier wurde objektorientierte Aufmerksamkeit so interpretiert, dass die Selektion aus einer Anzahl fertig segmentierter Objekte passiert. Sowohl die räumlichen Aspekte der Selektion als auch die Tatsache, dass eine vollständige Gruppierung zu einem Objekt wohl die Zuweisung fokaler Aufmerksamkeit voraussetzt, werden weitgehend ignoriert.

Ein neueres Modell objektbasierter visueller Aufmerksamkeit kommt von Sun und Fisher [SF03]. Es beruht auf der Zuweisung von datengetriebener und modellgetriebener Salienz an gruppierte Objekte, die dann untereinander um Aufmerksamkeit konkurrieren. Leider fehlt der Implementation bisher die eigentliche Bestimmung der Gruppierung.

Im Kontext einer Erkennung von menschlichen Gesten setzen Fislage et al. [FRR99] einen Mechanismus visueller Aufmerksamkeit ein, der im wesentlichen den Standardmodellen entspricht. Allerdings weist er spezialisierte Merkmalskarten und einen passenden Mechanismus zur Integration dieser Merkmalskarten auf.

Anders als konventionelle Modelle weist Jägersand [Jäg95] räumliche Aufmerksamkeit nicht nur einem Ort, sondern auch einer Auflösungsstufe in einem *scale space*-orientierten Modell visueller Aufmerksamkeit zu. Die Selektion orientiert sich dabei an informationstheoretischen Maßen.

4.1.4 Berücksichtigung von Dynamik

Auch wenn die vorgestellten Systeme prinzipiell zur Verwendung in dynamischen Umgebungen gedacht sind, erfolgt keine explizite Modellierung der Dynamik innerhalb der Aufmerksamkeitssteuerung. Die im folgenden genannten Systeme stellen die relevanten Ausnahmen dar.

Maki et al. [MNE96] setzen Tiefen- und Bewegungsmerkmale ein, um Masken zu berechnen. Diese dienen zur Ausblendung des Hintergrundes oder eines auffälligen Objektes. Als Bewegung wird die horizontale Komponente des optischen Flusses ausgewertet. Auf diese Weise kann das System eine Verfolgung ausführen (*pursuit mode*), indem die Bereiche ausgewählt werden, die dem aktuell selektierten Objekte bezüglich Tiefe und Bewegungsrichtung entsprechen. Sobald jedoch ein anderes Objekt dem Beobachter näher liegt, wird die gebildete Maske negiert, so dass ein Wechsel der Aufmerksamkeit (*saccade mode*) auf das näher liegende Objekt ermöglicht wird.

Maki et al. [MNE00] stellen eine Erweiterung des Systems vor, die unter anderem eine zusätzliche Verhaltensweise zulässt. Da die Selektion eines Objektes meist mit einer Aufgabe verknüpft ist, kann die Aufmerksamkeit nach Ausführung einer solchen Aufgabe auf das nächste Objekt gelenkt werden, wobei derartige Aufgaben jedoch nicht Bestandteil des vorgestellten Modells sind. Die Weiterentwicklung benutzt außerdem eine Prädiktion bezüglich der zu erwartenden Merkmalsausprägungen des selektierten Objektes und setzt nicht mehr auf deren Konstanz.

Auf ähnliche Weise benutzen Barile et al. [BBC+97] eine Zielmaske, die in diesem Fall initial durch Farbinformationen berechnet wird, um das zu verfolgende Objekt auszuwählen. Die Verfolgung findet nun allerdings ohne Farbinformation durch Bewegung eines Systems zweier Monochromkameras statt. Weitere Verhaltensweisen werden nicht bereitgestellt.

Die Gruppe von Dickmanns [Dic92, Dic98, DW99] stellt Ansätze unter dem Label des „Dynamic Vision" vor. Diese zeichnen sich durch die Einbeziehung von Eigenbewegung und dynamischer Umgebung bei einer Prädiktion des nächsten Zustandes aus. Bemerkenswert ist die temporale Schichtung von Aktionsmodellen. In einigen Systemen werden gleich mehrere Kamerasysteme miteinander kombiniert, um zum Beispiel gleichzeitig hochaufgelöste Bilder und breite Überblicke durch unterschiedliche Zoomeinstellungen zur Verfügung zu haben. Anwendungen liegen im Bereich der autonomen sichtgestützten Steuerung von Fahrzeugen.

Das Modell von Takacs und Wechsler [TW98a] modelliert eine Alternative zur ortsvarianten Verarbeitung der Retina. Anstelle der unterschiedlichen Auflösung wird in diesem Modell Pixeln im Zentrum ein höheres Gewicht in der folgenden Verarbeitung zugewiesen. So wird bei homogener Bildauflösung eine inhomogene Zuweisung von Ressourcen an Bildbereiche erzielt. Als salienzanzeigendes Feature wird über mehrere Skalen hinweg ein Maß für den Informationsgehalt eines Bereiches ermittelt. Da bereits deren Berechnung von der aktuellen Position des Fokus der Aufmerksamkeit abhängig ist, muss ein Kurzzeitgedächtnis realisiert werden, das neben dem Alter der extrahierten Information auch eine Gewichtung entsprechend der Nähe zum Fokus der Aufmerksamkeit berücksichtigt. Das Problem der möglichen Veränderung im Bild wird zwar angesprochen, aber nicht gelöst. Es wird hier in einer Anwendung zur Gesichtserkennung demonstriert, dass Aufmerksamkeitsme-

chanismen selbst bei eher klassischen Bildverarbeitungsanwendungen, die sich nicht durch komplexe dynamische Umgebungen auszeichnen, vorteilhaft verwendet werden können.

Das Modell von Horswill und Barnhart[HB96] erhebt zwar nicht den Anspruch, ein vollständiges Modell visueller Aufmerksamkeit zu sein. Es ist jedoch relevant, da es die Modellierung Visueller Suche mit Verfolgung und Segmentierung vereint und so sowohl den Aspekt der Selektion relevanter Regionen, als auch eine Verallgemeinerung der Selektion auf dynamische Szenen beschreibt. Das System betrachtet Pixel als Elemente eines n-dimensionalen Raumes, der sich aus den beiden räumlichen Dimensionen und den Werten von präattentiven Merkmalen ergibt. Segmente ergeben sich als Cluster in diesem Raum, die durch einen einfachen *k-means* Clustering-Algorithmus berechnet werden. Eine Vereinfachung ergibt sich durch die Verwendung einiger Heuristiken. Schließlich zeichnet sich das Modell durch eine Echtzeitimplementierung aus.

Die von Bollmann et al. [BHM97] dargestellte Version von NAVIS (siehe dazu auch 4.2.2) verfügt zusätzlich zu der normalen Aufmerksamkeitssteuerung und Objekterkennung über einen dauerhaft aktiven und dementsprechend schnellen Mechanismus zur Detektion von Bewegungen. Reagiert der Detektor, erfolgt ein Umschalten von der normalen Steuerung durch das Aufmerksamkeitsmodell auf ein Trackingsystem, das die Kontrolle wieder zurückgibt, sobald die Bewegung nicht mehr präsent ist und die Zuweisung von Aufmerksamkeit stören könnte.

4.1.5 Berücksichtigung von räumlicher Tiefe

Es existieren wenige technische Aufmerksamkeitsmodelle, die räumliche Tiefe verwenden. Wie im vorigen Kapitel geschildert, verwendet Maki [MNE96, MNE00] nicht nur Bewegung, sondern auch Tiefe als Merkmal zur Bestimmung der Objektmasken. Maki [MUE96, Mak96] spezifiziert das benutzte Verfahren zur Disparitätsbestimmung genauer. Es handelt sich um einen phasenbasierten Ansatz. Die Objektmaske und die errechneten Disparitätswerte werden dabei zur Vorhersage der nächsten Disparitäten eingesetzt, um den Suchraum entsprechend einzuschränken. Es wird auch ein Gütemaß berechnet, das die Zuverlässigkeit der Disparitätsinformationen beschreibt. Diese Güte wird benutzt, um eine gewichtete Propagation der Tiefeninformation durchzuführen.

Ratan [Rat95] verwendet in seinem integrierten Modell attentiver Objekterkennung Farbe und Tiefe als Hinweise zur Segmentierung von Objekten. Dabei werden für beide Stereobilder zunächst per Farbe und Kantendetektion Regionen gebildet, für die dann die entsprechenden Stereokorrespondenzen gesucht werden. Der Selektionsprozess richtet sich alleine nach vorab spezifizierten Eigenschaften des Zielobjektes. Dieses wird durch die Kameras fixiert und mit der Objektdatenbank verglichen.

Ahrns und Neumann [AN99, Ahr00] zeigen ein Modell zur monokularen, ortsvarianten Berechnung von Tiefeninformationen, das einen Aufmerksamkeitsmechanismus zur Selektion des nächsten Objektes enthält. Diese Auswahl findet zur Steuerung eines Roboterarmes statt, der visuell gesteuert das Objekt greifen soll und dafür entsprechend Tiefeninformationen berechnet. Diese Berechnung findet monokular durch Bewegung des Sensors und phasenbasierte Tiefenberechnung statt. Das Aufmerksamkeitsmodell selbst „weiß" hier jedoch nichts von der dreidimensionalen Raumstruktur.

Ein weiteres Aufmerksamkeitsmodell, das Tiefeninformation verwendet, stellen Ouerhani und Hügli vor [OH00]. Es handelt sich dabei weitgehend um ein Standardmodell nach Koch und Ullman [KU85] mit paralleler Merkmalsberechnung, Integration in eine Salienzkarte, Maximumsselektion und Markierung in einer statischen Inhibitionskarte. Die Tiefe wird als Feature verwendet, wobei die Sali-

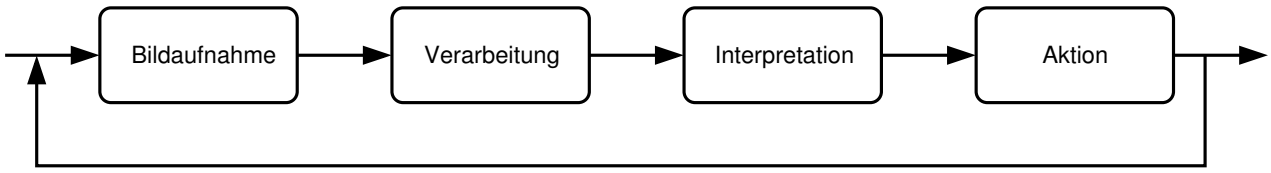


Abbildung 4.2: Perzeptions-Aktions-Zyklus beim Aktiven Sehen

enz jeden Ortes mit seiner Nähe wächst. Es findet jedoch keine Berechnung der Tiefeninformationen statt. Vielmehr ist das Modell auf einen direkten Tiefensensor angewiesen.

Im Kontext Mobiler Roboter ist das System von Frintrop et al. [FRNS03] angesiedelt, das einen dreidimensionalen Laser-Range-Finder zur gleichzeitigen Bestimmung von Tiefen- und Helligkeitsdaten verwendet. Diese gehen in ein an das Modell von Itti et al. [IKN98] angelehntes System ein, das die Modalitäten zu einem gemeinsamen Salienzmaß integriert.

Obwohl keine Berechnung der Tiefe vorgenommen wird, ist auch das Modell von Dickinson et al. [DCTO97] hier zu nennen. Es operiert zwar auf 2D-Bildern und bestimmt flächenbasierte Merkmale, verwendet jedoch für die integrierte Objekterkennung 3D-Modelle der zu erkennenden Objekte und steuert die Sensoren modellgetrieben so an, dass Uneindeutigkeiten in der Erkennung minimiert werden können. Insofern richtet sich die Aufmerksamkeit nach 3D-Informationen aus.

4.2 Aufmerksamkeit als Bestandteil Aktiver Sehsysteme

Aus der Sicht der visuellen Aufmerksamkeit stellt Aktives Sehen das technische Pendant zur offenen Aufmerksamkeit dar. Durch Bewegungen der Kameras ist es hier möglich, interessante Bildbereiche zu fokussieren. Der Bezug zur visuellen Aufmerksamkeit wird nur von einigen Systemen explizit untersucht. Die Motivation ist oft eher technischer Natur. Der Schwerpunkt liegt in der Betonung des Sehens als aktivem Prozess im Gegensatz zu der passiven Sichtweise auf die Verarbeitung gegebener Kameradaten. Das Sehen wird als Handlung begriffen, die zu Wissen über die Umgebung führt. Somit befinden sich Aktive Sehsysteme dauerhaft in einem Perzeptions-Aktions-Zyklus, wie er in Abb. 4.2 dargestellt ist.

4.2.1 Paradigmen des Aktiven Sehens

Man unterscheidet heute mehrere Ansätze des Aktiven Sehens, die einander keineswegs widersprechen, sondern unterschiedliche Aspekte betonen. Das Aktive Sehen fungiert dabei gleichzeitig als Überbegriff für die beschriebenen Herangehensweisen wie auch als einzelnes dieser Verfahren. Einen Überblick darüber geben Mertsching und Schmalz [MS99].

Active Vision: Als Reaktion auf die Probleme des Marr'schen Paradigmas wurde der aktive Aspekt der Wahrnehmung betont, der nicht aus den zur Verfügung stehenden Daten eine vollständige 3D-Rekonstruktion der Szene abzuleiten versucht. Vielmehr wird durch die Ansteuerung der Sensoren eine möglichst geeignete Datenbasis zur Lösung des aktuellen Problems erzielt. Aloimonos [AWB87, FA95] muss als herausragender Vertreter dieser Schule betrachtet werden. Er

betont vor allem die Reduktion in der Komplexität der zu lösenden Aufgaben durch Vermeidung ungeeigneter Blickpositionen mit mehrdeutigen Informationen.

Purposive Vision: Einen weiteren Schwerpunkt setzt Aloimonos [Alo93], indem er die aktuelle Aufgabe, das Ziel des Systems, in den Vordergrund stellt. Diese erlaubt eine Spezialisierung des Systems in Hinblick auf die Aufgabe. Sollte die Aufgabe Vereinfachungen erlauben, die sonst nicht möglich sind, sollten diese einbezogen werden. Noch stärker als im reinen *active vision* folgt Aloimonos hier der Theorie des *ecological vision* nach Gibson [Gib79]. Ein System, das diesen Ansätzen folgt, stellen Firby et al. [FKPS95] vor.

Active Perception: Schon 1985 wurde von Bajcsy [Baj85] das Problem des Bildverstehens von der Verarbeitung der Signale verlagert hin zur Kontrolle der Datenakquisition. Sie betont die Rückkopplung von gemessenen Parametern in den Aufnahmeprozess, um dem Ziel des *„We don't just see, we look"* ([Baj88, S. 996]) näher zu kommen. Eine mobiles System, für das das eigene Überleben im Vordergrund steht, dem alle anderen Ziele nachgeordnet sind, sucht die dafür benötigten Informationen aktiv. Die Selektivität wird stark betont. Die Selektivität bezieht sich nicht nur auf die verarbeiteten Reize. Es geht gleichzeitig um eine Selektion der verwendeten Berechnungen, deren Aufwand mit dem erwarteten Informationsgewinn abgeglichen wird.

Qualitative Vision: Unter dem Stichwort *qualitative vision* betont Aloimonos [Alo94], dass es im Aktiven Sehen weniger um die Vermessung der Umgebung durch visuelle Sensoren geht. Vielmehr sollen die relevanten qualitativen Aspekte der Szene bestimmt werden. Nicht die Repräsentation steht im Vordergrund, sie ist nur Werkzeug zum Lösen eines Problems. Dabei sind durchaus auch dreidimensionale Repräsentationen zugelassen, sofern sie dem Problem angemessen sind.

Animate Vision: Ballard [Bal91] sieht in der menschlichen visuellen Verarbeitung eine wichtige Quelle für die Konstruktion von Sehsystemen. Dazu gehört, dass ein Sehsystem aus einer kleinen Anzahl elementarer Kompetenzen oder Verhaltensweisen in Modulform zusammengesetzt ist. Auf eine interne Repräsentation wird soweit irgend möglich verzichtet. Stattdessen wird die Umgebung als externe Repräsentation aufgefasst. Das Modell ist zu unterscheiden von *animat vision* [Ter97]

Active Cooperation: Bajcsy [Baj95] sieht in der Modellierung der Interaktion eigenständiger aktiver Prozesse zur kooperativen Lösung von Sehaufgaben einen entscheidenden Schritt zur Lösung komplexer Aufgaben des Aktiven Sehens oder allgemeiner autonomer aktiver Systeme. Die Kooperation findet auf so verschiedenen Ebenen wie Wahrnehmung, Handlung, Bewegung und der Verwendung gemeinsamer Ressourcen statt. Während sich die Agenten über die gemeinsamen Ziele einig sein müssen, dürfen durchaus Differenzen dahingehend bestehen, wie diese Ziele zu erreichen sind.

Dass für viele konkrete Probleme keine vollständige Rekonstruktion im Sinne Marr's nötig ist, zeigt sich am Beispiel der Kollisionsvermeidung. Anstelle einer 3D-Repräsentation der Szene reicht die direkte Bestimmung der Kollisionszeit aus dem optischen Fluss aus. Dieses „Ausreichen" ist nun aber kein Begnügen mit einer einfacheren Lösung, sondern erzielt robustere Resultate, da die vollständige

Rekonstruktion generell ein schlecht gestelltes Problem ist. Ihre Lösung ist zusätzlich empfindlich gegen kleine Störungen der Eingabe.

Bereiche, in denen die Vorteile des aktiven Ansatzes offensichtlich sind, betreffen die Trennung von Vordergrund und Hintergrund, die sich ebenso wie die Auflösung von Mehrdeutigkeiten und die Behandlung von Okklusionen durch Bewegung des Sensors stark vereinfachen lässt.

Zu betonen ist die Übereinstimmung zwischen Aufmerksamkeitssteuerung und Aktivem Sehen hinsichtlich der Selektivität des Vorgehens. Beide heben die Tatsache hervor, dass nicht alle Teile des verfügbaren sensorischen Inputs zu jedem Zeitpunkt die gleiche Wichtigkeit haben und daher gleich behandelt werden sollten. Es geht jeweils darum, die momentan relevanten Informationen zu finden, sei es durch Veränderung der Aufnahmegeometrie oder interne Selektion. Diese relevante Information ist dann einer aufgabenabhängigen komplexeren Verarbeitung zu unterziehen.

4.2.2 Modelle

Es wurde in der relativ kurzen Zeit seit der Existenz des Paradigmas eine Vielzahl von Modellen zum Aktiven Sehen dervorgestellt, so dass hier vor allem besonders einflussreiche Arbeiten und solche mit einem Bezug zur vorgestellten Arbeit beschrieben werden.

Zwei Umsetzungen des Aufmerksamkeitsmodells FeatureGate [Cav99] auf Aktive Sehsysteme stellen Driscoll et al. [DPC98] und Stasse et al. [SKC00] vor. Das erste Modell konzentriert sich stärker auf die Ansteuerung des Kamerakopfes, das zweite eher auf das Erreichen von Echtzeit in der Modellierung des Systems. Rivlin und Rotstein [RR00] konzentrieren sich in ihrem Modell auf die Umsetzung der wesentlichen natürlichen Augenbewegungen und unterscheiden langsame Folgebewegungen und Sakkaden. Ihr Modell bezieht dafür eine foveale Region zur Berechnung ein, deren Größenbestimmung sich als einfaches Optimierungsproblem ergibt.

Von der Universität Paderborn stammen Aktive Sehsysteme [HBD99], die im Bereich der autonome Demontage von Autos eingesetzt werden. Grundkonzepte sind die Eckenverbreiterung zur Objektrepräsentation, eine ansichtenbasierte Erkennung, die Aufteilung in charakteristische Details und deren räumliche Relation. Die verwendete Stereoberechnung [SBDH00] soll dem System eine sehr exakte Ansteuerung eines Roboterarms ermöglichen.

Neben den bereits diskutierten Arbeiten von Maki [MNE96, MNE00] gibt es noch mehrere Modelle von der CVAP-Gruppe aus Stockholm. Hier sind es vor allem Eklundh, Granlund und Westelius, die Systeme des Aktiven Sehens entwickelt haben [GKWW94]. Brunnström et al. [BEU94] verknüpfen Aktives Sehen mit einer Objekterkennung, indem wiederholt Eckpunkte, Kreuzungspunkte lokalisiert und fixiert werden. Sie werden klassifiziert, so dass aus dieser Folge von Fixationen die Beschreibung eines Objektes resultiert. In einer neueren Arbeit konzentrieren sich Björkman und Eklundh [BE01] auf die Detektion von Bewegung zur Ausrichtung von Kameras auf solcherart als interessant bewertete Gebiete. Dazu ist es zuerst nötig, die Eigenbewegung zu schätzen, um den verbleibenden Anteil der Objektbewegung richtig einschätzen zu können.

McLauchlan, Murray et al. [MM95, MBM+95] rekonstruieren die Bewegung von Bildsegmenten, um eine Sakkade zum Ziel und langsame Folgebewegungen auszuführen, die eine Vorhersage der Segmentposition miteinbeziehen. Die aktuellen Schätzungen werden kontinuierlich aktualisiert und verbessert.

Den Schwerpunkt auf die Konstruktion eines Systems, das zur Untersuchung derartiger Mechanismen die Realzeitumsetzung von Blickbewegungen und der Kontrolle durch einen Aufmerksamkeitsmechanismus erlaubt, setzen Yamamoto et al. [YYL96]. Es beruht auf einem ortsvarianten Aufmerksamkeitsmechanismus, der im wesentlichen den üblichen Filtermodellen (s. 4.1.2) entspricht. Dieser wird um einen räumlichen Bildspeicher ergänzt, der die berechneten Merkmalskarten integriert. Von Interesse ist vor allem die vorgeschlagene Evaluation der Steuerung, auf die in Abschnitt 9 genauer eingegangen wird.

Ein weiteres Filtermodell visueller Aufmerksamkeit, das in ein Aktives Sehsystem integriert wurde, präsentieren Giefing, Janßen und Mallot [GJM92]. Über die übliche parallele Merkmalsberechnung mit Maximumsbestimmung und Inhibitionskarte zur Hemmung kürzlich besuchter Positionen hinaus verfügt das System über top-down-Einflüsse. Diese entstehen, indem anhand des ersten Aufmerksamkeitspunktes eine Hypothese über das gefundene Objekt gebildet wird. Für jedes Objekt sind Teilmuster in räumlichen Relationen vorhanden. Die zur momentanen Hypothese gehörigen Musterpositionen werden in einer *interest map* markiert und beeinflussen so die folgenden Blickbewegungen. Als Merkmale wird hier neben Linienelementen, Krümmungen und Schnittpunkten auch die zeitliche Ableitung des Signals zur Lokalisation von starken Bildveränderungen verwendet. Für alle Merkmale erfolgt eine ortsabhängige Gewichtung, die die peripheren Bereiche bevorzugt.

Reece und Shafer [RS95] demonstrieren an einem System zur Steuerung eines Autos (Ulysses 1-3) die erreichbare Beschleunigung durch selektive Wahrnehmung. Drei Schritte werden zur Verbesserung vollzogen:

- Datengetriebene Einschränkung der räumlichen Suche.
- Situationsabhängige Einschränkung der Suche nach Informationen, die für die aktuelle Entscheidung benötigt werden.
- Eine Modellierung des Veraltens von Informationen zur Neuakquisition.

Jeder Schritt bringt mehrere Größenordnungen an Beschleunigung mit sich, so dass am Ende eine ursprünglich unlösbare Aufgabe gelöst werden kann.

Die Serialisierung der Informationsgewinnung durch ein Aktives Sehen explorieren Arbel und Ferrie [AF01b, AF01a]. Dabei richtet sich die Aktivität des Systems nach einem Maß, das den Informationsgewinn des Systems beschreibt. Das Maß bezieht die Möglichkeiten zur Auflösung von Mehrdeutigkeiten mit ein. Callari und Ferrie [CF01] ergänzen das Vorgehen um die Einbeziehung bestehenden Wissen über wahrscheinliche Objekte in der Szene in die Blickplanung. Heinze und Groß [HG01] beziehen in die Blickplanung mehrere Vorhersagen über die Umgebung und die Konsequenzen eigenen Handelns mit ein, wodurch ein konnektionistisches Modell entsteht, das Perzeption und Aktion miteinander integriert.

Die zugrunde liegenden Repräsentationen von Aktiven Sehsystemen betrachten Rao und Ballard[RB95]. Objekterkennung und Objektsuche werden durch zwei Gedächtnisstrukturen modelliert, die unterschiedlich adressiert werden. Es findet eine Trennung von attentiver und präattentiver Verarbeitung statt. Als Datenstruktur werden hochdimensionale Merkmalsräume (DoG-Filter unterschiedlicher Orientierung und Skalen) verwendet. Wasson [Was99] hingegen sieht eine Sammlung attendierter Bereiche als geeignete Repräsentation der Umgebung für ein Aktives Sehsystem auf einem mobilen Roboter. Die Bereiche werden durch Stereokorrelation räumlich lokalisiert.

Paulus et al. [PDR+00, PAH+00] betonen die breite Suche nach einem Objekt und seine genaue Fokussierung zur Verifikation der Hypothese. Dazu wird ein Zusammenspiel von datengetriebenen und modellgetriebenen Einflüssen in einem objektorientierten Rahmenwerk vorgeschlagen. Der Aufbau von Aktiven Sehsystemen stellt schließlich eine Herausforderung an Technik, Ansteuerung, Kalibrierung und Softwarearchitektur dar. Diese Themen werden für ein umfangreiches System von Crowley und Christensen [CC95] analysiert. Der Interaktion und Kommunikation von Modulen Aktiver Sehsysteme widmen sich Crowley et al. [CBBS94]. Wie sich durch die Verwendung mehrerer Module ein redundantes und dadurch leistungsfähigeres Sehsystem ergeben kann, demonstrieren Fayman et al. [FPCR96].

Die ursprüngliche Version der Aufmerksamkeitssteuerung von NAVIS (Neural Active VIsion System) [MBHS99, Bol00], dem Aktiven Sehsystem der AG IMA an der Universität Hamburg, integriert datengetriebene und modellgetriebene Einflüsse sowohl statischer als auch dynamischer Art. Sie ist Teil eines Aktiven Sehsystems mit Objekterkennung. NAVIS weist damit die Integration eines komplexen biologienahen Aufmerksamkeitsmodells in ein umfangreiches Aktives Sehsystem auf.

Es stehen mehrere Verhaltensmodelle zur Verfügung, die jeweils datengetriebene und modellgetriebene Aspekte aufweisen. Speziell verwendet die Objekterkennung sogenannte Aufmerksamkeitspunkte, Maxima der datengetriebenen Salienzberechnung, schon beim Lernen der Objektmodelle. Die Aufmerksamkeitspunkte und eine Repräsentation der Umgebung werden zusammen mit dem zugehörigen Merkmal abgespeichert. Außerdem wird die räumliche Relation der Aufmerksamkeitspunkte zueinander gespeichert. Beim Erkennen der Objekte findet nun ein entsprechender Abgleich der Umgebung statt, um eine Hypothese zu bilden. Anschließend gehen die aus der Hypothese resultierenden erwarteten Positionen der weiteren Aufmerksamkeitspunkte in die Steuerung der Aufmerksamkeit mit ein. So kann die Hypothese durch Ansteuern dieser Punkte verifiziert oder falsifiziert werden.

Das in dieser Arbeit vorgestellte Aufmerksamkeitsmodell entstand im Kontext von NAVIS und baut auf einigen Modulen auf, die im Rahmen von NAVIS anderweitig Verwendung fanden. Zu einem späteren Zeitpunkt ist eine Integration dieses Aufmerksamkeitsmodells mit weiteren Bestandteilen von NAVIS intendiert.

Teil II

Modellierung visueller Aufmerksamkeit

Kapitel 5

Die Berechnung lokaler Salienz

Der erste Schritt in der Zuordnung datengetriebener Aufmerksamkeit besteht in der Bestimmung der lokalen Auffälligkeit. Um die Selektion eines interessanten Objektes oder Bereiches durchzuführen, braucht man ein Maß dafür, was interessant ist. Dieses Maß muss für die Bereiche berechnet werden, für die eine fokale Selektion in Frage kommt, also nicht nur am momentan selektierten Ort. Eine derartige Berechnung ist somit für jedes Modell unerlässlich. Sie gehört eindeutig zur präattentiven Stufe, da sie der Zuordnung von Aufmerksamkeit eben gerade voraus geht und auch eine Basisinformation zur Verlagerung von Aufmerksamkeit darstellt.

5.1 Ziel

Die Berechnung lokaler Salienz ist ein zentraler Bestandteil von Modellen visueller Aufmerksamkeit. Sie dient der datengetriebenen Bestimmung von Orten mit hoher Relevanz für das System. In der reinen Modellierung von Aufmerksamkeit unabhängig von spezifischen Aufgaben und Umgebungen sollen möglichst allgemeine Merkmale Verwendung finden. In der Literatur Aktiver Sehsysteme finden sich vielfach einfache Filteroperationen aus der klassischen frühen Bildverarbeitung. Wolfe und Gancarz [WG96] weisen jedoch darauf hin, dass der Mensch offensichtlich wesentlich komplexere Merkmale ausnutzt, die bereits objektbezogene Eigenschaften miteinbeziehen. Eine Annahme, die experimentell durch Malinowski und Hübner [MH01] gestützt wird.

Die Berechnung der Merkmale erfolgt präattentiv, d.h. vor der Zuweisung von Aufmerksamkeit. Die Merkmale werden somit für den gesamten Bildbereich bestimmt. Das Ergebnis wird üblicherweise - und so auch in dieser Arbeit - in jeweils einer Merkmalskarte abgelegt, die die Salienzwerte anhand des betrachteten Merkmals für jeden betrachteten Ort wiedergeben.

Die Merkmale können entweder problemspezifisch sein oder aber eine allgemeine Ausrichtung haben. Problemspezifische Merkmale bieten sich immer dann an, wenn Vorwissen über die Art der Umgebung oder die zu lösende Aufgabe zur Verfügung steht. Beispiele für solche Merkmale wären Detektoren für Gesichter oder allgemeiner Hautfarben, für bekannte räumliche Aspekte (künstliche Marker) oder für interessante Formen (Verkehrszeichen). Für das allgemeine Aufmerksamkeitssystem muss auf solche Merkmale verzichtet werden, sie sollten aber bei jeder konkreten Anwendung eingesetzt werden. Somit bleiben für dieses System allgemeine Merkmale, deren Auswahl und Umsetzung sich an folgenden Eigenschaften orientiert:

- **Informativität:** Das Merkmal sollte eine relevante Szeneneigenschaft wiedergeben.
- **Objektzusammenhang:** Wenn möglich sollte sich das Merkmal bereits auf Objekte oder Regionen als Objektkandidaten beziehen.
- **Stabilität:** Das Merkmal sollte sich durch einfache Veränderungen der Szene (Größe, Translation, Rotation, Rauschen, Beleuchtung) nur in angemessener Weise beeinflussen lassen.
- **Ähnlichkeit zum menschlichen Vorbild:** Das Merkmal sollte Eigenschaften entsprechen, die durch den Menschen präattentiv detektierbar sind.
- **Komplementarität:** Die Merkmale sollten sich so ergänzen, dass möglichst viele unterschiedliche Bildinformationen ausgewertet werden.
- **Einfachheit:** Da die Auswertung präattentiv parallel für das gesamte Bild stattfindet, spielen Rechenzeiterwägungen eine wichtige Rolle.

Im Sinne der Komplementarität werden durch die vorgestellten Merkmale folgende Bildeigenschaften ausgewertet: Grauwert, Farbe und Stereoinformationen. Zu jedem Merkmal gibt es individuelle Beispiele, bei denen dieses Merkmal besonders deutlich wird sowie ein durchgehendes allgemeines Beispiel. Zusätzlich wird die Stabilität jedes Merkmals durch gezielte Variationen der Szene überprüft.

5.2 Grauwertbasierte Merkmale

5.2.1 Einführung

Die klassische Bildverarbeitung operiert auf einer zweidimensionalen Intensitätsrepräsentation, die die im Bild enthaltenen Informationen auf die Helligkeit reduziert. Auch für den Menschen ist bekannt, dass Helligkeitsunterschiede mit einer höheren Auflösung wahrgenommen werden als Farbunterschiede [Wan95]. Es ist also naheliegend, zuerst diese Informationen auszunutzen und sich dabei an klassischen Bildverarbeitungsoperationen zu orientieren.

In einem Intensitätsbild kann man dann zwischen Flächen oder Kanten als einfachsten Informationsträgern unterscheiden. Weitere Basen wären Texturen oder höhere Merkmale wie Ecken. Kanten weisen eine hohe Informationsdichte auf und die Mehrheit der Neurone im primären visuellen Kortex des Menschen sind sensitiv für orientierte Kanten. Demgegenüber entsprechen Flächen eher den Oberflächen von Objekten und sind daher von Bedeutung. Diesem Zusammenhang soll im Modell durch zwei Merkmale Rechnung getragen werden, die auf den unterschiedlichen Informationen beruhen und durch ihre Komplementarität möglichst unterschiedliche Bildinformationen ausnutzen.

Die beiden Merkmale beruhen auf denen von Bollmann für NAVIS entwickelten Merkmalen [Bol00, MBHS99]. Auf die in der ursprünglichen Arbeit vorgenommene Einteilung in Merkmalskarten, die Informationen der Umgebung wiedergeben und Auffälligkeitskarten, die eine Bewertung der Information als salient enthalten, wird in dieser Arbeit weniger eingegangen, da sie im Rahmen der neuen Architektur keine bedeutende Rolle spielt. Nichtsdestotrotz bietet auch das vorgestellte Modell die Möglichkeit, Informationen weiterzuverwenden, die als Zwischenergebnisse der Merkmalsberechnungen anfallen.

5.2.2 Symmetrie

Die Symmetrie stellt für den Menschen ein präattentives Merkmal dar, das sowohl natürliche Objekte wie Tiere oder Pflanzen als auch viele künstliche Objekte auszeichnet. Dagegen finden sich in zufälligen Zusammenstellungen von Objekten oder deren Teilen kaum Symmetrien. Das legt die Verwendung von Symmetrie als aufmerksamkeitsleitendem Merkmal nahe. Bereits frühzeitig hat man nachgewiesen, dass auch Menschen Symmetrien bei der Wahrnehmung von Objekten auswerten. Die klassischen Arbeiten von Kaufman und Richards [KR69] zeigen, dass bei Blickbewegungen Fixationen signifikant häufiger auf Symmetriezentren gerichtet werden. Schließlich zeichnet sich gerade die Kreissymmetrie durch ihre Invarianz gegen Größenänderung, Translation und Rotation aus. Als aufmerksamkeitserregendes Merkmal wird Symmetrie auch in [RWY95, YYL96, HNR98] verwendet. Das hier vorgestellte Vorgehen basiert auf dem von Bollmann [Bol00] vorgestellten Symmetriemerkmal.

Die Symmetrie stellt bereits eine ästhetische Qualität der wahrgenommenen Welt dar, deren Wahrnehmung der Erkennung und Speicherung vorausgeht. Die Detektion von Symmetrie findet laut Corbalis und Roldan sowie Royer [CR75, Roy81]in weniger als einer Sekunde statt. Dabei dominieren nach Julesz et al. [Jul71, JC79] tiefe Ortsfrequenzen die Detektion der Symmetrie. Einen Überblick über die Psychophysik der Symmetrieerkennung bietet Zabrodsky [Zab90].

Kantendetektion

Die Berechnung lokaler Kreissymmetrie geht in dieser Arbeit von der biologisch plausiblen Bestimmung lokaler Kanten durch Gaborfilter aus. Gaborfilter [Gab46] stellen einerseits eine gute Modellierung der Antwortfunktion rezeptiver Zellen im visuellen Areal V1 dar, wie Pollen und Ronner [PR83] zeigten. Andererseits sind Gaborfilter aus informationstheoretischer Sichtweise eine ideale Wahl, da sie die Unschärferelation des Orts-Bandbreitenproduktes an der unteren Grenze erfüllen. Aus diesem Grunde wurden Gaborfilter in NAVIS bereits für unterschiedlichste Zwecke eingesetzt, unter anderem für die Objekterkennung [MMS98a, MMS98b], Gruppierung [MBM02], Bestimmung von Tiefeninformationen [LMS98] und Objektverfolgung [HMS99]. Die Verwendung in diesem Kontext bedeutet also eine Steigerung der Effizienz, da die Ergebnisse der Gaborfilterung mehrfach verwendet werden können.

Ein Gaborfilter besteht aus einer komplexwertigen Schwingung mit einer Gauß-förmigen Einhüllenden und führt so eine lokale Bestimmung der Ortsfrequenz durch. Die in diesem Kontext interessanten zweidimensionalen Gaborfilter werden definiert durch :

$$g(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} \cdot \exp(-\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{2}) \cdot \exp(i\mathbf{k}^T\mathbf{x}). \tag{5.1}$$

Relevante Parameter sind die Matrix $\mathbf{A}$, die sich aus Vorzugsorientierung φ und radialer und tangentialer Filterbandbreite σ_x bzw. σ_y ergibt und der Vektor der Schwerpunktsortsfrequenzen $\mathbf{k}$:

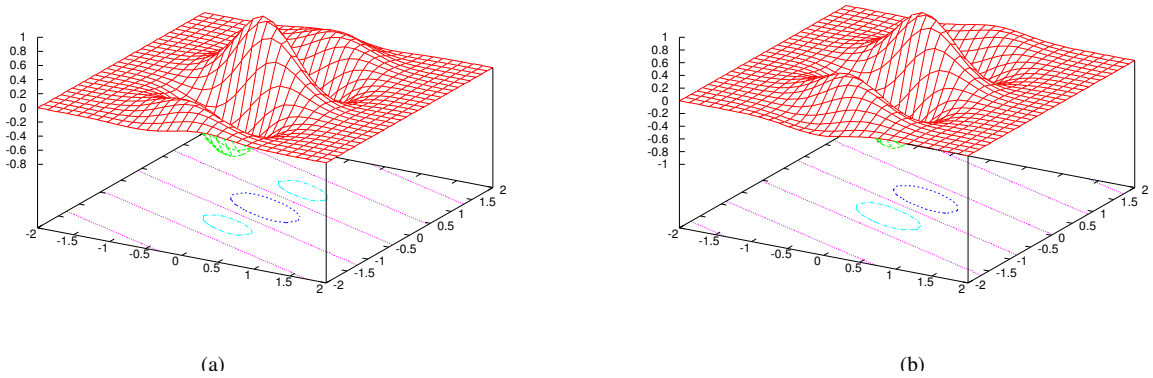
$$\mathbf{A} = \mathbf{R}\mathbf{P}\mathbf{R}^T = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} \sigma_x^{-2} & 0 \\ 0 & \sigma_y^{-2} \end{bmatrix} \begin{bmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{bmatrix}. \tag{5.2}$$

Anschaulich bedeutet dies, dass ein Gaborfilter bevorzugt Bildanteile einer bestimmten Frequenz $\mathbf{k}$ und Orientierung φ durchlässt. Wie stark die Auswahl ist, wird dabei durch die Filterbandbreiten

bestimmt.

Um einen Filtersatz zu bilden, der die Orientierungen einer bestimmten Skala abdeckt, setzt man $\mathbf{k} = [k_0 \cos\varphi,\ k_0 \sin\varphi]^T$ mit $k_0 = |\mathbf{k}|$ als Betrag der Schwerpunktsortsfrequenz, so dass die Modulationsrichtung orthogonal zur Hauptachse der Einhüllenden liegt. Für eine gleichmäßige Abdeckung der Orientierungen sei $\varphi = n\Delta\varphi_0$. Die Fouriertransformierte des Filters ergibt sich so zu:

$$G(\mathbf{k}) = \exp(-\frac{(k-k_0)^T (A^{-1})^T (k-k_0)}{2}). \tag{5.3}$$



(a) (b)

Abbildung 5.1: Gaborfilter im Ortsraum: (a) Realteil, (b) Imaginärteil.

Für einen mehrskaligen Filtersatz sind die Parameter k_0, φ_0, σ_x und σ_y anzupassen. Einen solchen Filtersatz stellen zum Beispiel Theimer und Mallot [TM94] vor, hier wird jedoch der von Trapp für NAVIS entwickelte Filtersatz [Tra96] Verwendung finden. Bei diesem wird die relative Bandbreite der Filter konstant gehalten, was zu einer logarithmischen Aufteilung des Ortsfrequenzspektrums führt. Die Überlappung der Gaborfilter unterschiedlicher Orientierungen und unterschiedlicher Skalen wird so auf einen konstanten Wert r festgelegt, der das Verhältnis zwischen σ_x und σ_y determiniert.

Der gesamte Filtersatz besteht aus 4 Skalen bei einer Winkeldifferenz φ_0 von 15° (s. Abb. 5.2), wovon aus Rechenzeiterwägungen bisher innerhalb der datengetriebenen Aufmerksamkeitssteuerung nur eine Skala verwendet wurde. Diese Einschränkung soll in dieser Arbeit überwunden werden.

Abb. 5.1 zeigt den reellen und den imaginären Anteil eines Gaborfilters im Ortsraum, sowie den Gaborfilter im Frequenzraum (dort ist er reell). In der Arbeitsgruppe IMA wurde auch eine Hardwareimplementierung der Gaborfilter vorgestellt, die sich zur Realzeitberechnung der entsprechenden Daten auf einem FPGA anbietet [VM01]. Im Rahmen dieser Arbeit wurde auf die dort entstandene optimierte C-Implementierung der Gaborfilterung zurückgegriffen[1]. Abb. 5.3 zeigt an einem Beispielbild das Ergebnis einer Gaborfilterung mit 12 Orientierungen und folgenden Parametern: $k_0 = 0.75$, $r = 0.5$.

[1] Für die zur Verfügung gestellte Implementation sei Dipl.-Phys. Nikolaus Voss gedankt.

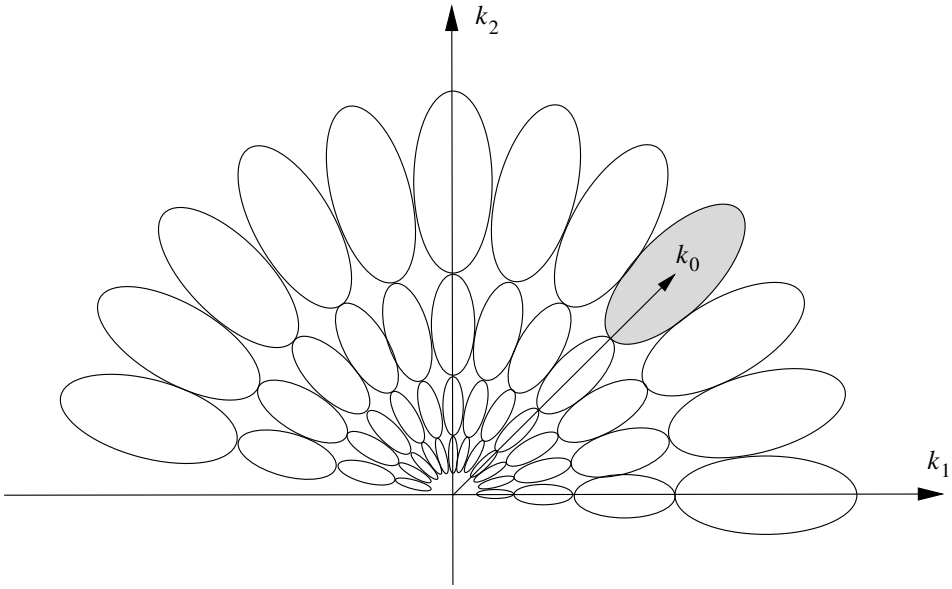


Abbildung 5.2: Filterdurchlassbereich in der Frequenzebene für den beschriebenen Gaborfiltersatz.

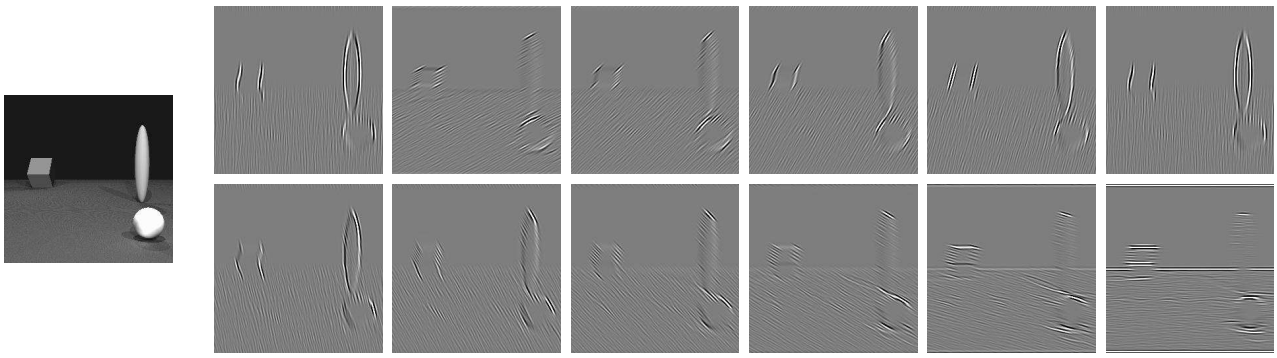


Abbildung 5.3: Ergebnis der Gaborfilterung an einem einfachen Beispielbild.

Auflösung	Radien	Breite	Radien (Originalauflösung)	Breite (Originalauflösung)
64*64	6, 9, 12, 15	3	24, 36, 48, 60	12
128*128	6, 9, 12, 15	3	12, 18, 24, 30	6
256*256	6, 9, 12	3	6, 9, 12	3

Tabelle 5.1: Verwendete Parameter für die Multiskalenversion der Symmetrieberechnung. Alle Angaben in Pixel.

Symmetrieberechnung

Die so berechneten Kanteninformationen sollen nun derart eingesetzt werden, dass sie die lokale Symmetrie einer aus ihnen definierten Struktur wiedergeben. Dazu wird für jeden Punkt die Energie der Kanten, die in einem Bereich um einen bestimmten Radius tangential zu einem Kreis mit diesem Radius liegen, aufaddiert. Dies geschieht getrennt für mehrere Radien so, dass die Bereiche um diese Radien aneinander grenzen. Von den Summen, die sich für die verschiedenen Radien ergeben, wird das Maximum ausgewählt. Das Vorgehen ist schematisch in Abb. 5.4 dargestellt.

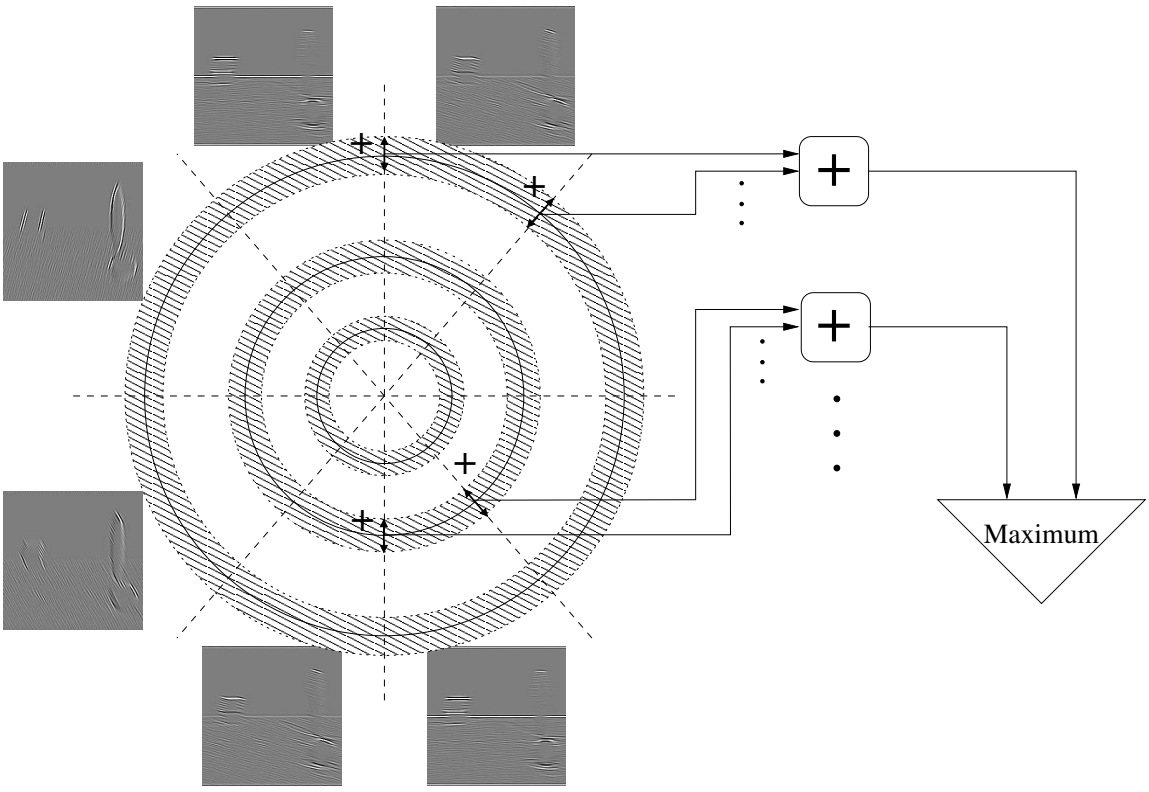


Abbildung 5.4: Verfahren zur Berechnung der lokalen Symmetrie - schematische Darstellung. Für einige Orientierungen wird ein Beispiel für die Gaborfilterantwort mit angegeben.

Eine Normalisierung der Werte ergibt sich aus den Maximalwerten für die jeweiligen Radiussummen, von denen das Maximum tatsächlich ausgewählt wird.

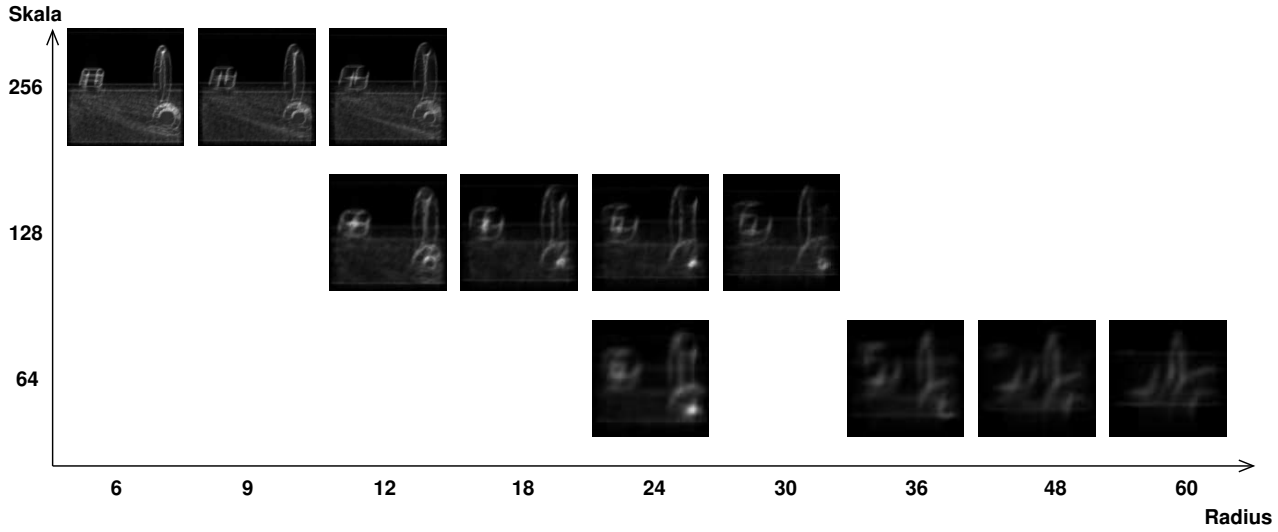


Abbildung 5.5: Symmetriesalienz für verschiedene Skalen und Radien (in Pixel) am Beispiel aus Bild 5.3. Die Radien beziehen sich auf die Originalbildauflösung von 256 Pixeln. Man beachte, dass zum Beispiel die Kugel im Vordergrund mit ihrem Bildradius von 23 Pixeln durch die Breite der Gaborfilterantworten für zwei benachbarten Radien eine starke Antwort erzeugt.

Erweiterung auf Multiskalenberechnung

In der von Bollmann [Bol00] vorgestellten Variante bleibt der Summationsbereich für wachsende Radien konstant. Auch der Frequenzbereich der Kantendetektion bleibt für alle Radien und damit für alle Größen der symmetrischen Strukturen konstant, was einer Veränderung des Verhältnisses von Größe und Kantenbreite entspricht. Die Translation eines Objektes entlang der Tiefe führt so selbst dann zu einer deutlich veränderten Reaktion des Systems, wenn die entsprechende Größenänderung im Bereich der untersuchten Radien liegt. Eine Vergrößerung der Summationsbereiche nach außen hätte Reaktionen auf Kanten derselben Breite für alle Radien zur Folge. Daher bleibt als Konsequenz die Verwendung mehrerer Skalen in der Gaborfilterung, so dass für größere Radien auch Kanten tieferer Frequenz und damit höherer Breite ausgewählt werden.

Dies wird durch einen Mehrskalenansatz umgesetzt, in dem größere Radien in entsprechend größenreduzierten Bildern untersucht werden. Dies hat gegenüber der Verwendung mehrerer Filter unterschiedlicher Frequenzbereiche auf denselben Bilddaten den Vorteil einer drastisch beschleunigten Berechnung. Die Beschleunigung betrifft dabei sowohl die Bestimmung der Filterantworten als auch die Berechnung der Symmetrieinformation.

Um zusätzlich die Detektion weiterer Verhältnisse von Kantenbreite und Radius zu ermöglichen, werden die Radien in den Skalen so gewählt, dass sich die Objektgrößen teilweise überdecken. Die Integration der Salienzwerte unterschiedlicher Radien und Skalen erfolgt weiterhin per Maximumsbildung. Die Größe der Strukturen geht zusätzlich als kleiner multiplikativer Faktor ein, um größere Strukturen zu bevorzugen. Salienzwerte unter einem gewissen Schwellwert werden unterdrückt.

Abb. 5.5 und Tab. 5.1 erläutern den Mechanismus und geben die Parameter an, die in den vorgestellten Experimenten Verwendung fanden. Es bleibt bei der Maximumsbildung über die Skalen hinweg als integrierender Operation. Das Ergebnis für ein Beispielbild ist Abb. 5.6 zu entnehmen.

Weiterer Vorteil dieses Vorgehens ist die Möglichkeit, den Aufwand für die Berechnung durch die

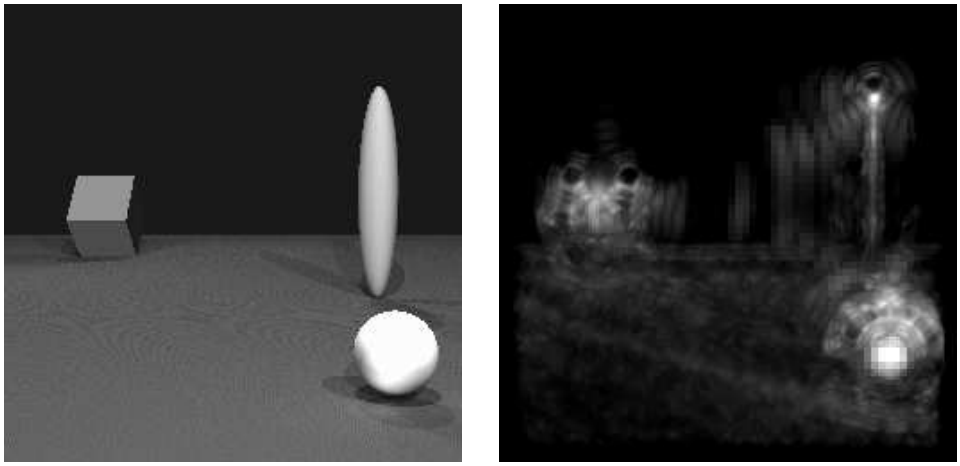


Abbildung 5.6: Ergebnis der Multiskalensymmetrieberechnung für ein Beispiel (links Eingabe, rechts Ergebnis)

Auswahl der verwendeten Skalen zu parametrisieren. Dies erlaubt es, zustandsabhängig die Ressourcen für diese Art der Salienzberechnung zu kontrollieren. So wäre denkbar, die Merkmalsberechnung nach der Systeminitialisierung oder nach Kamerabewegungen, in den Momenten also, in denen der Aufwand für die folgenden Selektionsstufen sehr hoch ist, in einer reduzierten Variante berechnen zu lassen, die eben zum Beispiel die am höchsten aufgelösten Skalen auslässt. Eine andere denkbare Ausnutzung zur Erhöhung der Gesamtleistung ist die wechselweise Berechnung der Skalen in aufeinanderfolgenden Frames, bei denen die zuvor berechneten Ergebnisse ersetzt werden.

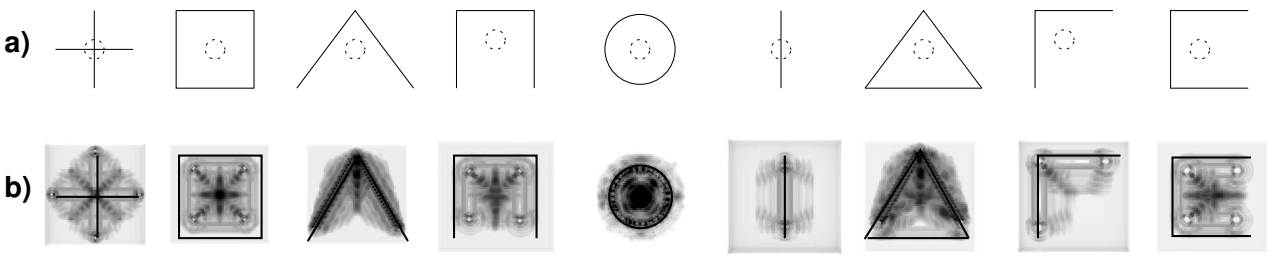


Abbildung 5.7: Initiale Fixation bei einfachen Figuren nach Kaufman und Richards [KR69] (a). Die gepunkteten Kreise geben die Bereiche an, in denen 86 % der Fixationen liegen. (b) zeigt die Salienz anhand des Symmetriemerkmales der Eingabe überlagert, wobei hier der günstigeren Darstellung wegen höhere Auffälligkeit mit dunkleren Bildwerten einhergeht.

Die Ähnlichkeit zum menschlichen Vorbild lässt sich anhand von Fixationsexperimenten analysieren, die Kaufman und Richards [KR69] durchführten. Abb. 5.7 zeigt für einige einfache Formen diejenigen Bereiche, in die 86 % der spontanen Fixation von Menschen fallen. Darunter ist die Salienzkarte hinsichtlich Symmetrie für dieselben Formen dargestellt. Es zeigt sich eine sehr gute Übereinstimmung zwischen der Merkmalsberechnung und den empirischen Daten, wenngleich das Symmetriemerkmal meist mehrere Bereiche hoher Auffälligkeit identifiziert. Das Antwortmaximum liegt jedoch mit Ausnahme des sechsten Reizes (senkrechte Linie) immer im Bereich der häufigsten Fixationen.

5.2.3 Exzentrizität

Die Orientierung von Elementen gehört zu den Eigenschaften, die häufig in Experimenten zur Visuellen Suche als präattentiv detektierbare Merkmale eingesetzt werden. Im Gegensatz zur Symmetrie, zu der dieses Merkmal möglichst komplementär angelegt ist, beruht die Exzentrizität auf Flächensegmenten, deren Ausgedehntheit bewertet wird. Während die Symmetrie als starke Eigenschaft der Grenzen zur Formung eines Objektes beiträgt, benötigt die Bestimmung nicht-symmetrischer Objekte bzw. von Objekten, deren Symmetrie durch Verdeckungen gestört ist, Homogenitätseigenschaften der Flächen.

Dazu wird in einem ersten Schritt die lokale Veränderung der Grauwerte mit einer einfachen Sobelfilterung in x- und y-Richtung bestimmt. Hohe Werte geben eine starke Veränderung der Grauwerte an, stellen also eine Verletzung der Homogenität dar. Durch ein einfaches *region growing*-Verfahren (s. Kap. 2.2.5) wird die Information ausgewertet. Die Startpunkte müssen folgendes Kriterium erfüllen, das gleichzeitig das Kriterium für das Wachstum der Segmente darstellt:

> Der Betrag des Gradienten beider Richtungen muss unter einem festgelegten Schwellwert liegen, damit der Punkt als Startpunkt ausgewählt wird bzw. dem Segment zugeordnet wird.

Im Unterschied zur von Bollmann [Bol00] vorgestellten Version des Merkmals wird der Schwellwert anhand des Histogramms der Sobelfilterantwort berechnet. Er wird so bestimmt, dass 65 % aller Pixel unterhalb dieses Wertes liegen. Die Grenze ist empirisch festgelegt worden, später wird die Abhängigkeit des Verfahrens von diesem Wert untersucht. Die Suche nach Startpunkten wird zeilenweise in Leserichtung durchgeführt, ist aber prinzipiell von der Suchreihenfolge unabhängig. Von jedem Startpunkt aus werden die benachbarten Punkte, die das Wachstumskriterium erfüllen, dem Segment hinzugefügt. Dieser Prozess wird rekursiv fortgeführt, bis keine zulässigen Nachbarn mehr existieren.

Der Schwellwert könnte für eine optimale Segmentierung lokal bestimmt werden. Da dies jedoch zeitaufwändig ist, wird ein relativ kleiner Schwellwert voreingestellt und stattdessen nach dem Regionenwachstum ein Verschmelzungsverfahren auf die vielen kleinen entstandenen Segmente angewandt. Das Vorgehen wird vor allem durch die Probleme anderer Verfahren mit größeren Oberflächen, die lokale Strukturen wie etwa Beschriftungen enthalten, motiviert. Als Kriterien für die Verschmelzung zweier Segmente gelten die Differenz der durchschnittlichen Grauwerte und die Varianz der Grauwerte beider Segmente. Konkret sind die beiden folgenden Bedingungen zu erfüllen:

1. Die absolute Differenz der durchschnittlichen Grauwerte beider Segmente darf den Schwellwert nicht überschreiten: $|\mu_A - \mu_B| < max_\mu$ mit $max_\mu = 20$.

2. Die Varianz der Grauwerte in beiden Segmenten muss in derselben Größenordnung liegen: $1/k < \sigma_A/\sigma_B < k$ mit $k = 2$.

Die Suche nach benachbarten Segmenten und die möglichen Verschmelzungen beziehen die Pixel mit ein, die in der initialen Segmentierung keinem Segment zugeordnet werden konnten, es findet also gleichzeitig eine Dilation der Segmente statt. Das Verschmelzungsverfahren wird einige Male iteriert, in den vorgestellten Experimenten wurden maximal vier Iterationen durchgeführt, ein Wert

Abbildung 5.8: Berechnung der Salienz für die Exzentrizität am Beispiel

der empirisch bestimmt wurde, ebenso wie die Schwellwerte in den Verschmelzungsbedingungen. Abb. 5.8 zeigt für ein Beispielbild die Sobelfilterung, die große Zahl initialer Segmente (hier 881), die durch Dilation, Verschmelzung und Entfernung zu kleiner Segmente deutlich reduziert wird (hier auf 89 Segmente) und schließlich die Bewertung als salient anhand der Exzentrizität.

Anschließend werden Segmente, deren Größe zu gering ist, um sie für die Aufmerksamkeitssteuerung interessant zu machen, entfernt. Für die verbliebenen Segmente wird per Hauptachsentransformation die dominante Orientierung berechnet. Hierzu ist es nötig, die Momente des Segmentes zu bestimmen, die sich im diskreten zweidimensionalen Fall mit $\overline{x}$ und $\bar{y}$ als Mittelwerten der entsprechenden Koordinaten durch

$$m_{p,q} = \sum (x - \bar{x})^p * (y - \bar{y})^q \tag{5.4}$$

beschreiben lassen ($m_{0,0}$ gibt dabei die Anzahl der Pixel, $m_{1,0}$ bzw. $m_{0,1}$ die x- bzw. y-Koordinate des Flächenschwerpunktes an). Der Orientierungswinkel ergibt sich aus den Momenten zweiter Ordnung zu:

$$\phi = \frac{1}{2} \arctan(\frac{2m_{1,1}}{m_{2,0} - m_{0,2}}). \tag{5.5}$$

Die Formel lässt sich sowohl als Bestimmung des größten Eigenwertes des Eigenvektors der Kovarianzmatrix [Jäh97], wie auch als Minimierung eines Abstandstermes der Punkte von einer Geraden [JKS95, Pit00] herleiten. Die Segmente werden in 12 Merkmalskarten (jeweils 15° Orientierung) eingetragen, die eine Kategorisierung der Orientierungen vornehmen. Eine zusätzliche 13. Karte enthält die Segmente ohne dominante Vorzugsrichtung (s. Abb. 5.9).

Das Ausmaß an Salienz richtet sich nach der Exzentrizität, die sich nach Jähne [Jäh97] ebenfalls leicht anhand der zuvor beschriebenen Momente zweiter Ordnung des Segmentes berechnen lässt:

$$\varepsilon = \frac{(m_{2,0} - m_{0,2})^2 + 4m_{1,1}^2}{(m_{2,0} + m_{0,2})^2} \tag{5.6}$$

Die Exzentrizität ist 0 für ein rundes Objekt und 1 für ein linienförmiges und daher als Salienzmaß bereits geeignet normiert. Der errechnete Salienzwert wird allen Punkten des Segmentes zugewiesen. Abb. 5.10 zeigt einige einfache geometrische Formen und ihre Salienzwerte. Das Ergebnis gibt die angestrebte Bewertung der Exzentrizität wieder. Dass die Salienzbereiche nicht exakt den Formen entsprechen, liegt an der durchgeführten Dilation der Segmente. Abb. 5.11 gibt die Anwendung der Merkmalsberechnung auf das durchgehende Beispiel wieder.

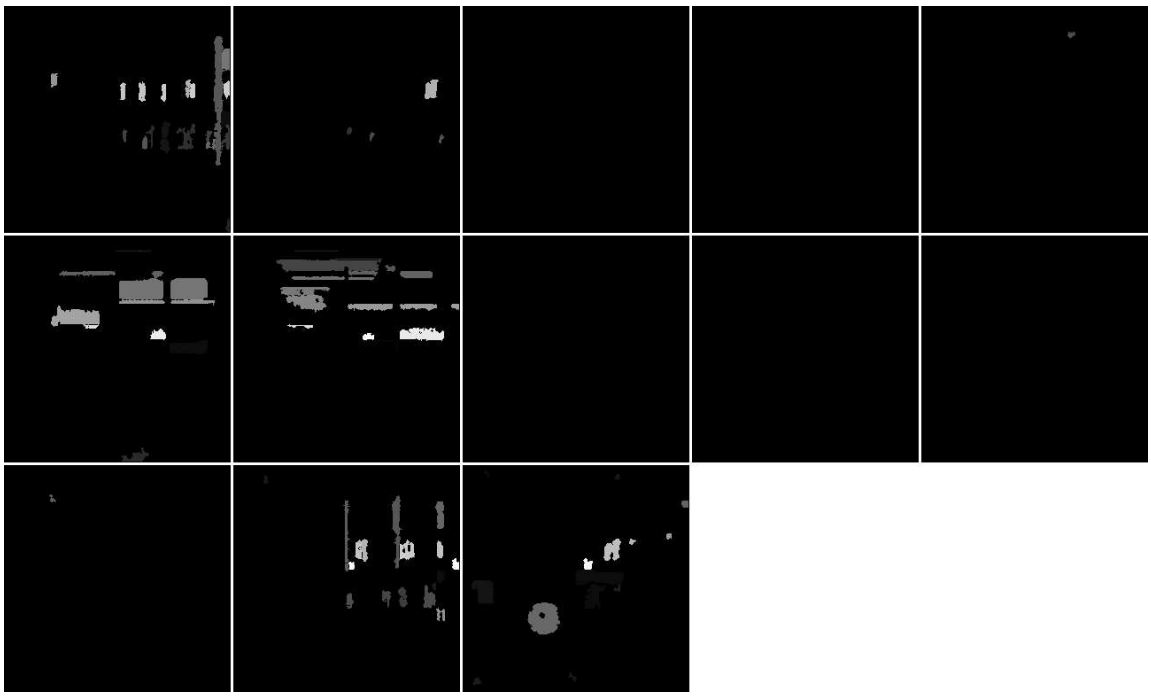


Abbildung 5.9: Merkmalskarten für die Kategorisierung der Orientierungen für das in Abb. 5.8 verwendete Eingabebild

Abbildung 5.10: Einfache geometrische Formen (links) und die ihnen zugeordnete Salienz (rechts). Hellere Bildpunkte bezeichnen höhere Salienzen.

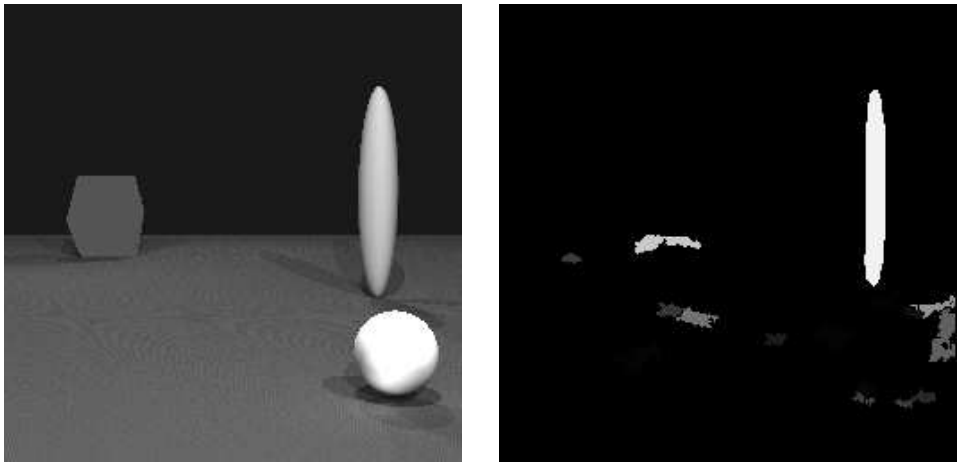


Abbildung 5.11: Ergebnis der Merkmalsberechnung Exzentrizität.

5.2.4 Experimente

An zwei zusätzlichen Beispielbildern sollen zuerst die mit den Berechnungen zur Symmetrie und Exzentrizität erzielten Ergebnisse in Abb. 5.12 illustriert werden. Genauere Eigenschaften der Algorithmen erhält man durch gezielte Variation der Eingabedaten. Hier sollten sich Effekte auf die Berechnungen qualitativ vorhersagen und mit den wirklichen Ergebnissen vergleichen lassen. Dazu gehören zum einen Eigenschaften, gegenüber denen eine Invarianz erhofft wird, wie Veränderungen der Größe, Rotation und Position von Objekten sowie das Einfügen von Rauschen. Dagegen sollte das Verfahren sensitiv auf die Modifikation gerade der Eigenschaft reagieren, auf die das Merkmal ansprechen soll. Letzteres demonstriert Abb. 5.13. Zu sehen ist eine Variation der Exzentrizität des rechten Objektes, das zu Anfang im Symmetriemerkmal stark repräsentiert ist, jedoch mit Zunahme der Exzentrizität dort immer geringere Salienzwerte verursacht und stattdessen eine immer stärkere Salienz anhand der Exzentrizität erreicht.

Zu den erwünschten Eigenschaften der Merkmale gehört eine gewisse Robustheit gegenüber Veränderungen der Eingabedaten, unter anderem gegen ein Rauschen, das auf die Bilddaten gelegt wird. Abb. 5.14 zeigt die Addition wachsender Anteile normalverteilten Rauschens auf das Eingabebild und die Konsequenzen für die Merkmalsberechnung. Wie angestrebt, zeigen sich beide Merkmale robust gegen diese Veränderungen und bestimmen in allen Fällen dieselben Bereiche maximaler Salienz, auch wenn durch das Rauschen andere Bereiche in ihrer Salienz verändert werden.

Eine Untersuchung des Symmetriemerkmales auf Abhängigkeit der gewählten Parameter ergibt sich, da das Verfahren im wesentlichen parameterfrei arbeitet. Bei der Exzentrizität sind vor allem die Schwellwerte von Bedeutung, die im folgenden untersucht werden.

Der Schwellwert für das initiale Bereichswachstum wurde empirisch auf 0.65 festgelegt, es werden also 65 % der Pixel als Flächen zugehörig und die restlichen 35 % als Randpixel angesehen. Umfangreiche Veränderungen des Parameters an einem Beispielbild (Abb. 5.15) zeigen, dass es einen großen Bereich gibt (etwa von 0.5 bis 0.75), in dem die Einteilung in Segmente plausibel erscheint (senkrechte Regalkanten, zwei Hälften des Bildes, durch Regal sichtbare Teile der Wand).

Später in der Verarbeitung erfolgt die Verschmelzung der Segmente, die von einem Schwellwert abhängt, der angibt, wie stark sich die mittleren Grauwerte unterscheiden dürfen. Der empirisch

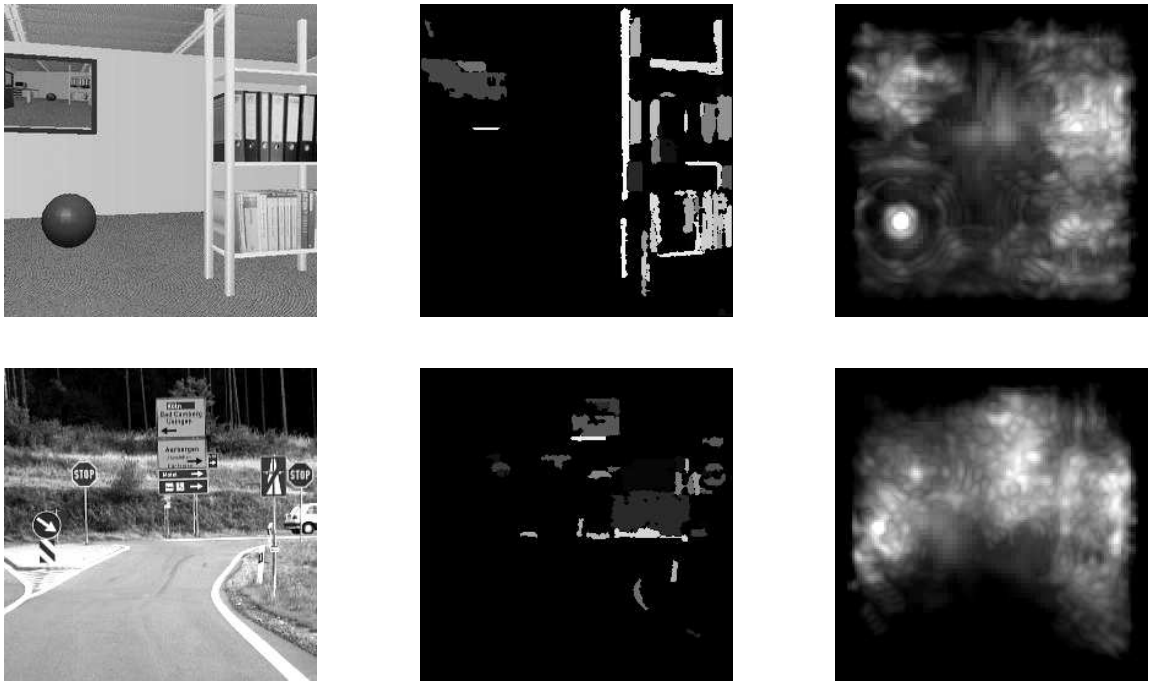


Abbildung 5.12: Beispiele für die grauwertbasierten Merkmalsberechnungen Exzentrizität (Mitte) und Symmetrie (rechts).

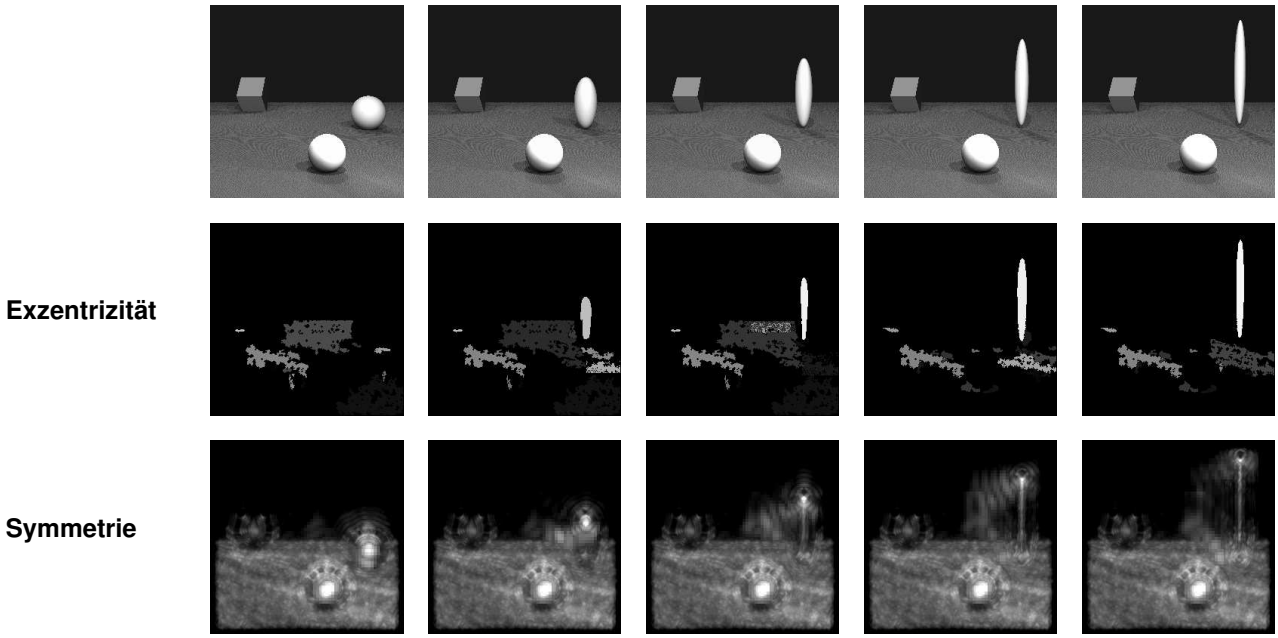


Abbildung 5.13: Variation von Exzentrizität und Effekt bezüglich der korrespondierenden Merkmale (Details im Text).

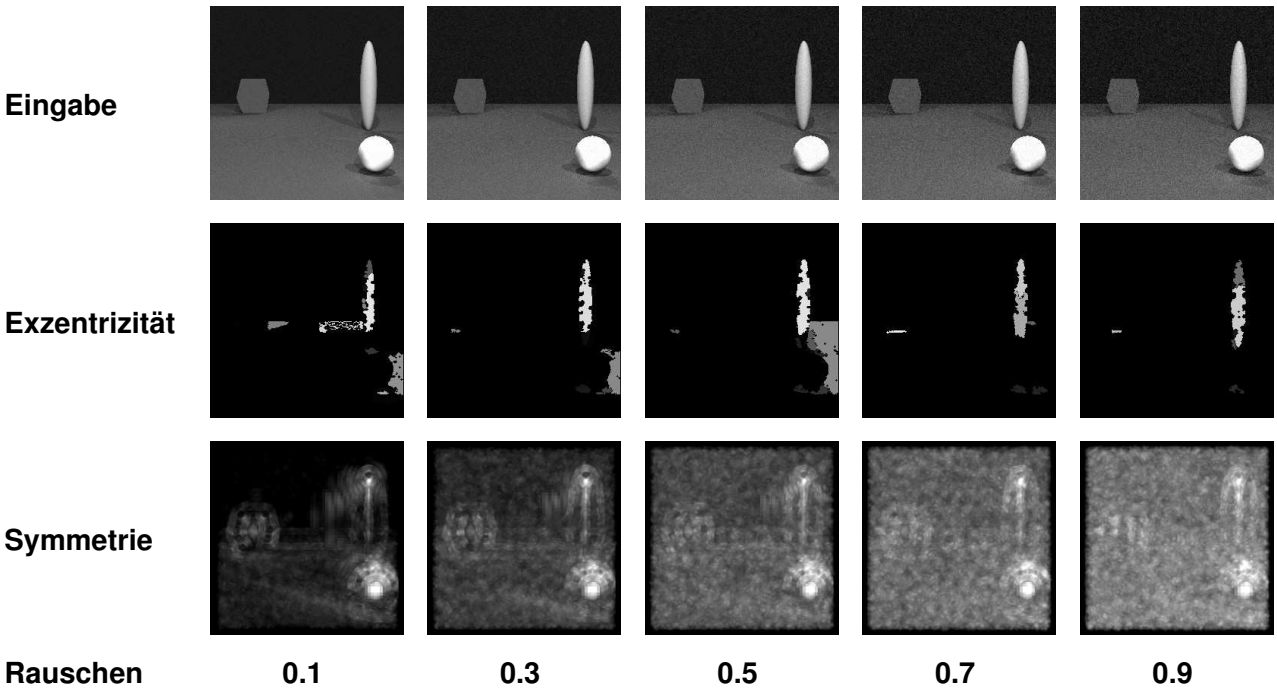


Abbildung 5.14: Empfindlichkeit der Merkmale gegen die Addition von Rauschen.

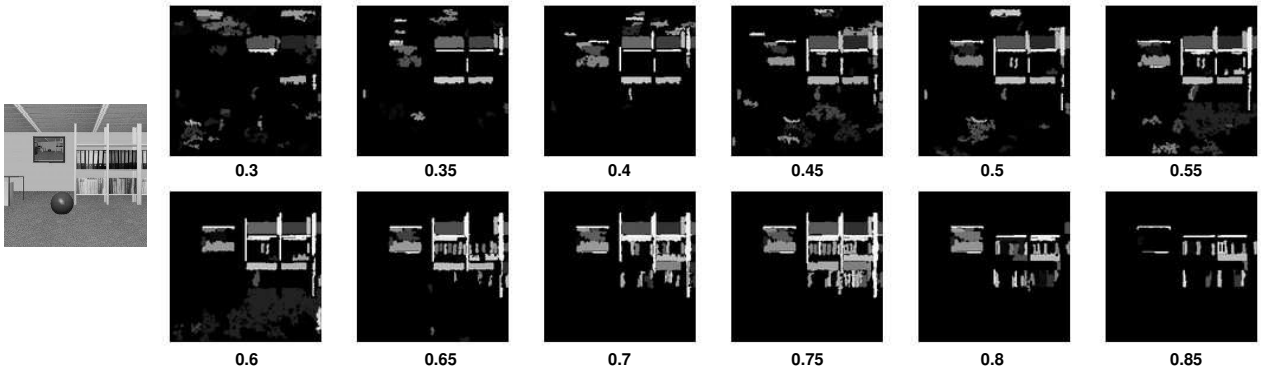


Abbildung 5.15: Einfluss des Schwellwertes auf die Segmentierungsergebnisse für das initiale Bereichswachstum.

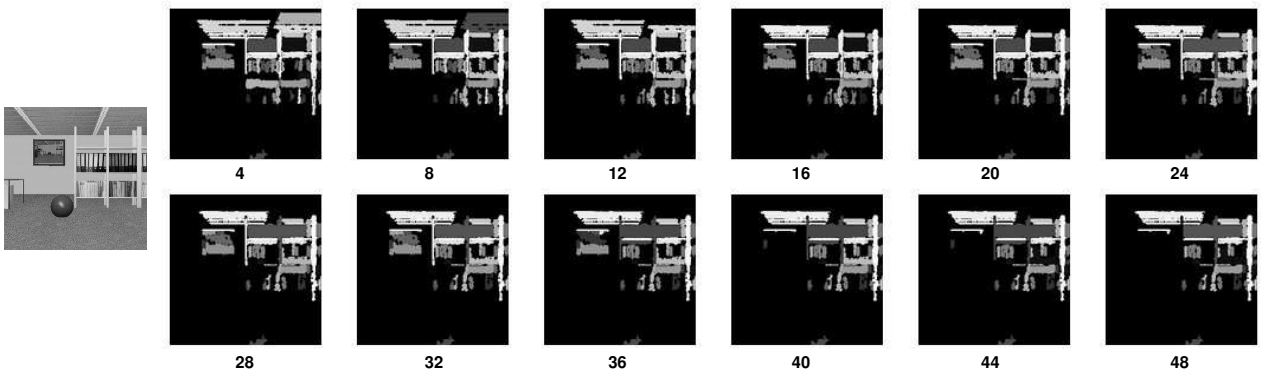


Abbildung 5.16: Einfluss des Schwellwertes für die Verschmelzung von Segmenten auf die Segmentierungsergebnisse

festgelegte Wert 20 liegt in einem Bereich von 12 bis 36, in dem man plausible Segmentierungen für das gegebene Bild findet (s. Abb. 5.16).

Die Merkmale Symmetrie und Exzentrizität zeigen sich also robust gegenüber Veränderungen der Szene, geben die gewünschte Eigenschaft wieder und sind unempfindlich gegen Veränderungen ihrer Parameter.

5.3 Farbbasierte Merkmale

5.3.1 Einführung

Die Verwendung von Farbe zur Kennzeichnung wichtiger Objekte ist sowohl aus der Natur (Warnfarbe giftiger Tiere) als auch aus der Technik (Verkehrsschilder, Gefahrenzeichen) bekannt. Auch umgekehrt wird die Angleichung von Farben benutzt, um eine Erkennung möglichst zu erschweren (Tarnung von Tieren oder militärischen Objekten). Demnach ist Farbe sowohl als wichtiger Hinweis auf Objekte und deren Grenzen anzusehen, als auch als auffällige und hinweisende Eigenschaft. Schließlich motiviert auch die eindeutige Einordnung von Farbe in die präattentiven Merkmale in Experimenten zur Visuellen Suche ihre Verwendung in einer Aufmerksamkeitssteuerung.

Da es hier einerseits um Farbe als Eigenschaft des Objektes und weniger der Szenenbeleuchtung geht, andererseits bekannt ist, dass Farben gerade im Kontrast zu ihrer Umgebung wahrgenommen werden, wird als Merkmal, das die Farbinformationen ausnutzt, der Farbkontrast gewählt. Die Berechnung des Merkmals beruht auf dem von Bollmann vorgestellten Farbmerkmal [BMD95, BM95, BJM98, Bol00].

5.3.2 Farbkontrast

Farbraumtransformation

Um eine Bewertung der Farbe zu erreichen, die der menschlichen Farbwahrnehmung nahekommt, ist die Transformation vom technischen RGB-Farbraum in einen empfindungsgemäßeren Farbraum notwendig (s.a. 2.2.2). Zur Auswahl eines geeigneten Farbraumes wurden von Bollmann [Bol00] Untersuchungen angestellt, die den Munsell-Farbraum als Referenz ansetzten. Dabei stellten sich der

Abbildung 5.17: Beispiel für Munsell-Farbraumtransformation. Rechts sind die drei Komponenten dargestellt.

CIELab [Hun87] und der MTM-Farbraum [MY88] als geeignete Approximationen dar, wovon der erste über eine zylindrische, der zweite zusätzlich auch über eine orthogonale Repräsentation verfügt. Während die zylindrische Form für die Überführung in eine sprachliche Beschreibung der Farben besser geeignet ist, stellt die Singularität der Unbuntachse ein Problem für technische Verfahren, in diesem Fall die Segmentierung, dar, weswegen die Wahl auf den MTM-Farbraum fällt.

Die der Segmentierung vorausgehende Transformation in den MTM-Farbraum wird durch eine erste Transformation der RGB-Daten in einen XYZ-Tristimulus erreicht

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0,608 & 0,174 & 0,200 \\ 0,299 & 0,587 & 0,144 \\ 0,000 & 0,066 & 1,112 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5.7}$$

, aus denen eine Transformation in den Adams-Farbraum (M_1, M_2, M_3) stattfindet:

$$\begin{aligned} M_1 &= V(1,02 \cdot X) - V(Y) \\ M_2 &= 0,4 \cdot (V(0,847 \cdot Z) - V(Y)) \\ M_3 &= 0,23 \cdot V(Y) \end{aligned} \tag{5.8}$$

Dabei berücksichtigt $V(x) = 11,6 \cdot x^{1/3} - 1,6$ die Nichtlinearität der menschlichen Wahrnehmung. Der eigentliche MTM-Farbraum (S_1, S_2, L) entsteht nun durch:

$$\begin{aligned} S_1 &= (8,88 + 0,966 \cdot \cos\varphi) \cdot M_1 \\ S_2 &= (8,025 + 2,558 \cdot \cos\varphi) \cdot M_2 \\ L &= M_3 \end{aligned} \tag{5.9}$$

, worin $\varphi = \arctan(\frac{M_1}{M_2})$. Die drei Komponenten sind für ein Beispiel in Abb. 5.17 dargestellt.

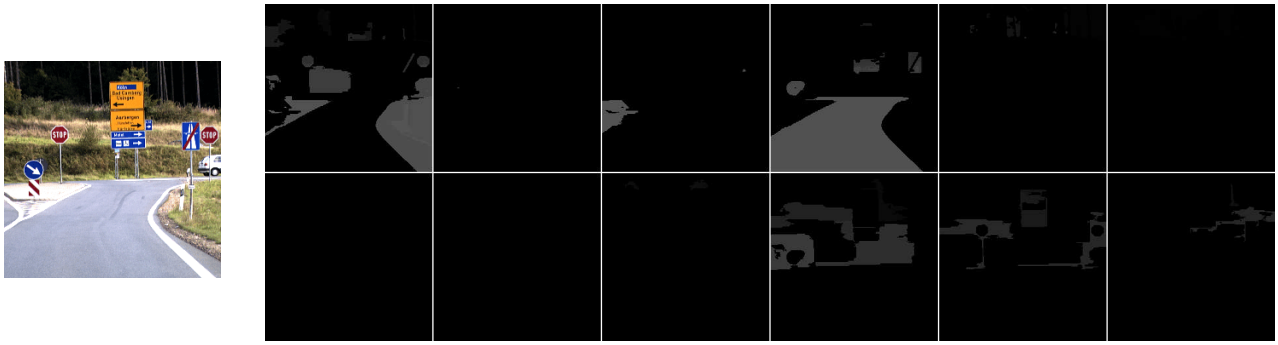


Abbildung 5.18: Segmentierungsergebnis für ein Beispielbild (links) mit Einteilung in 12 Farbklassen (rechts)

Segmentierung

Anschließend findet die eigentliche Segmentierung anhand von Farbe statt, die durch ein auf zentroider Verkettung beruhendes Bereichswachstumsverfahren umgesetzt wird. Hierbei wird immer der Wert des aktuellen Pixels mit den Mittelwerten in Frage kommender Segmente verglichen, wobei es dem ähnlichsten Segment zugeordnet wird, wenn die Distanz einen dynamischen Schwellwert ϑ_{Farbe} unterschreitet. Dieser Schwellwert bezieht die Farbvarianz in der Umgebung des Punktes additiv mit ein. Andernfalls wird mit diesem Pixel ein neues Segment begründet. Die Schwelle wird abhängig gemacht von der Varianz der Farbwerte in einer lokalen Umgebung um diesen Punkt, um eine Übersegmentierung in stark texturierten Bereichen zu vermeiden.

$$\vartheta_{Farbe} = c_{cadd} + c_{cmult} * \sigma^2 \tag{5.10}$$

Die Distanz, die hier verwendet wird, ist die Euklidische Distanz auf den MTM-Farbwerten. Um eine allzu starke Abhängigkeit von der Bearbeitungsreihenfolge zu vermeiden, wird die Richtung, in der die Zeilen bearbeitet werden, alterniert. Die Farbe der Segmente wird in 12 Kategorien eingeteilt. Abb. 5.18 zeigt ein Beispiel für das Resultat der Segmentierung anhand von Farbinformationen. Allzu kleine und große Segmente werden vor der weiteren Verarbeitung entfernt.

Salienzbestimmung

Die Salienz wird nun bestimmt als durchschnittlicher Farbkontrast entlang der Grenze zu den Nachbarsegmenten. Dabei wird der während der Segmentierung ermittelte Farbmittelwert der Segmente benutzt und als Kontrast der Euklidische Abstand dieser Mittelwerte berechnet, gewichtet mit der Länge der gemeinsamen Grenze:

$$\begin{aligned} F_i &= \frac{1}{U_i} \sum_{j \in B_i} b_{ij} \cdot d(< C_i >, < C_j >) \\ U_i &= \sum_{j \in B_i} b_{ij} \end{aligned} \tag{5.11}$$

Hierin bezeichnen U_i den Umfang des Segmentes i, B_i die Indizes aller Nachbarn zu diesem Segment, b_{ij} die Länge der gemeinsamen Grenze und d den Euklidischen Abstand der Farbmittelwerte

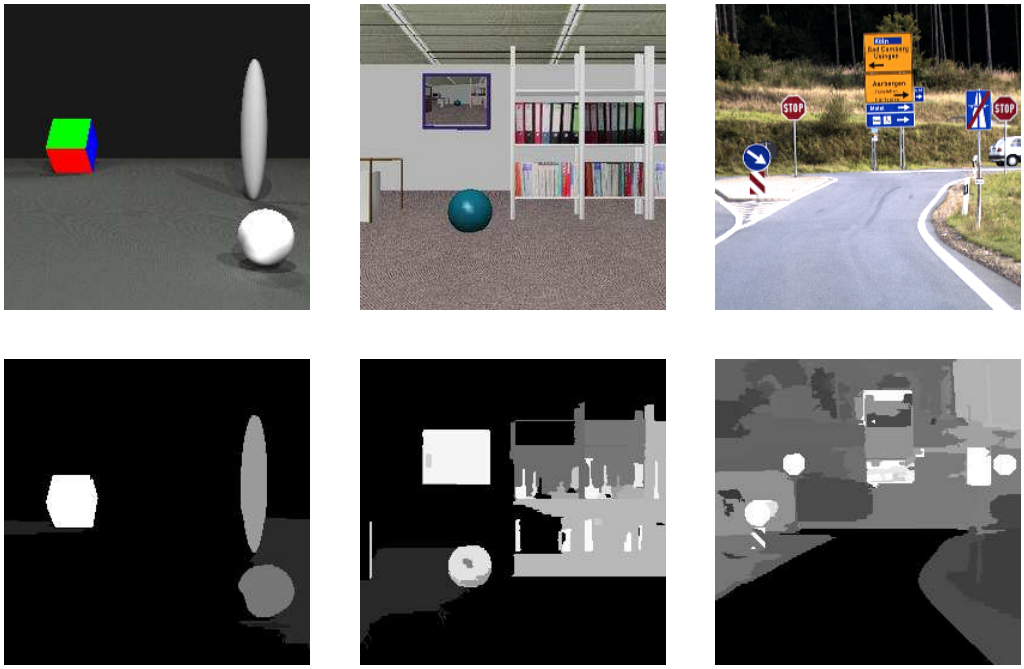


Abbildung 5.19: Bestimmung der Salienz anhand des Farbkontrastes (unten) für drei Beispielbilder (oben).

$< C_i >$, $< C_j >$ der Segmente i und j. Damit wächst die Salienz mit der Distanz der Farben im MTM-Farbraum und wird für Komplementärfarben maximal. Da es sich bei der Salienz für jedes Segment um eine gewichtete Mittelung von Euklidischen Abständen handelt, liegt der resultierende Wert im Intervall $[0, 1]$. Da hohe Werte insgesamt eher selten vorkommen, wird noch eine Nichtlinearität eingeführt, die kleine Werte unterdrückt und hohe Werte betont, wozu in diesem Fall eine sigmoide Funktion Verwendung findet:

$$feat_{color}(x, y) = \frac{1}{1 + \exp(-\beta * (2 * F_i - 1))} \tag{5.12}$$

mit $(x, y) \in S_i$ und $\beta = 3$.

5.3.3 Experimente

Das Charakteristik des Merkmales Farbe lässt sich in Abb. 5.19 erkennen, wo die Ergebnisse der Merkmalsberechnung für drei Eingabebilder unterschiedlicher Domänen und Qualitäten dargestellt werden. Abb. 5.20 zeigt das mit der Erhöhung des Farbkontrastes im Bild einhergehende Anwachsen der Salienz. Die Unempfindlichkeit gegen Rauschen demonstriert Abb. 5.21.

Die Farbsegmentierung wird beeinflusst durch den Schwellwert für die Verschmelzung eines Punktes mit einem benachbarten Segment. Dieser wird zwar dynamisch bestimmt, enthält aber einen konstanten faktoriellen Einfluss c_{cmult}, sowie einen additiven Einfluss c_{cadd} (siehe Formel 5.10). Die Parameter sind für alle gezeigten Experimente empirisch auf $c_{cmult} = 5$; $c_{cadd} = 8$ festgelegt. Dass es keiner spezifischen Optimierung dieser Parameter bedarf, ist Abb. 5.22 zu entnehmen. Dort wurden beide unabhängig voneinander deutlich verändert. Die Ergebnisse zeigen für die wesentlichen salienten Bildbereiche (Ball und Bild) in einem breiten Bereich der Parameter keine deutlichen Veränderun-

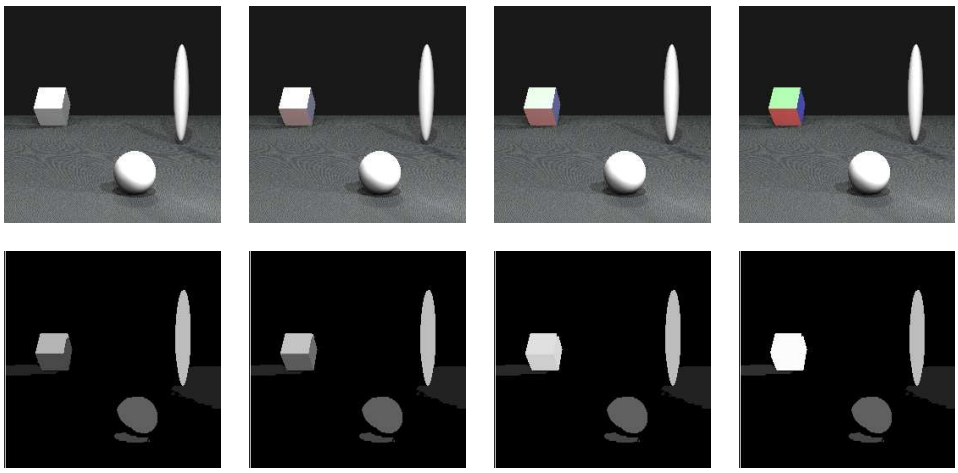


Abbildung 5.20: Erhöhung des Farbkontrastes und Effekt bezüglich des korrespondierenden Merkmals (Details im Text).

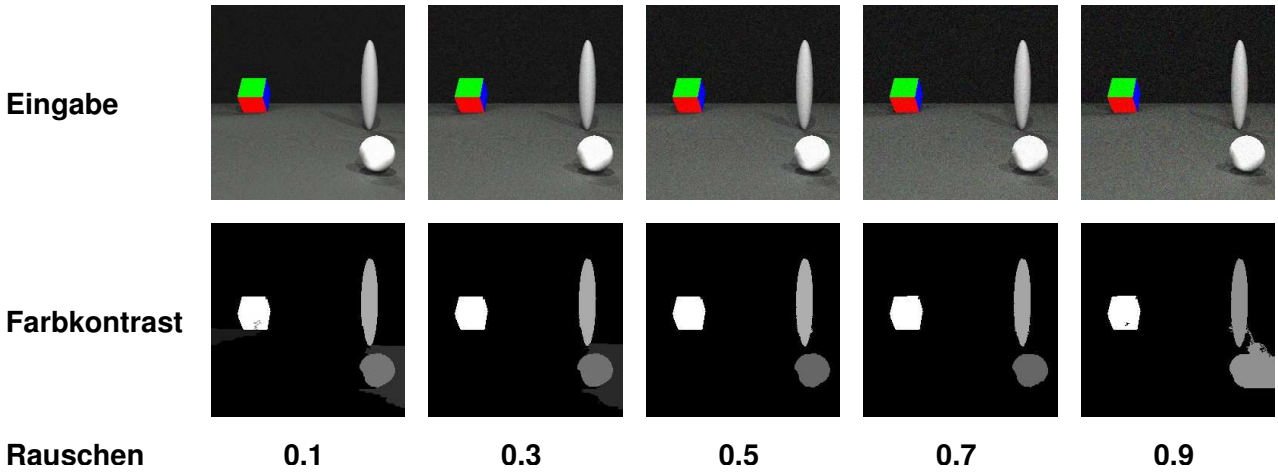


Abbildung 5.21: Robustheit des Farbmerkmals gegenüber der Addition normalverteilten Rauschens.

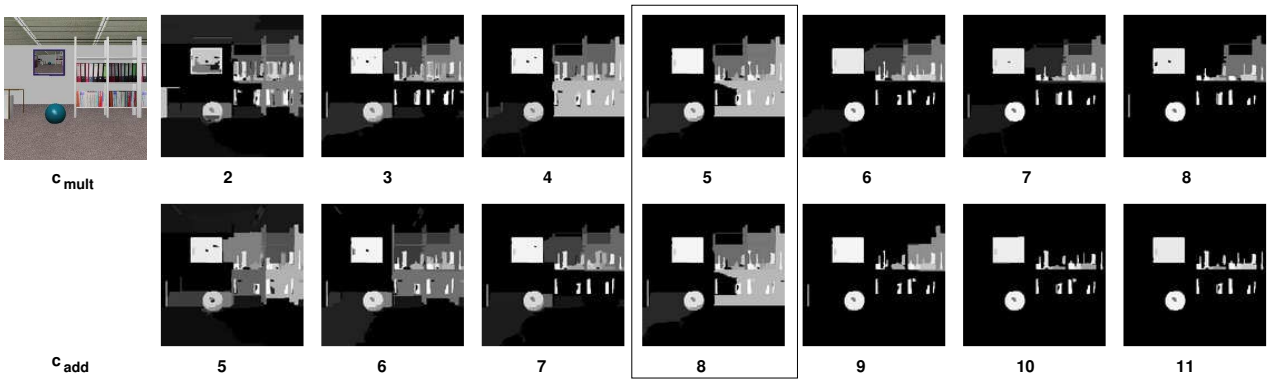


Abbildung 5.22: Abhängigkeit des Farbmerkmals von den Schwellwerten c_{mult} und c_{add}. Die gewählten Parameter sind hervorgehoben.

gen. Erst bei deutlich veränderter Wahl der Parameter ergeben sich signifikante Veränderungen der Salienz.

5.4 Stereobasierte Merkmale

5.4.1 Einführung

Räumliche Tiefe lässt sich einerseits als Eigenschaft eines Punktes im zweidimensionalen Bild, andererseits auch als zusätzliche Dimension eines dreidimensionalen Bildes auffassen. In dieser Alternative zeigt sich die besondere Rolle von Tiefe, die sich auch im Bereich der menschlichen Aufmerksamkeit niederschlägt. So haben Nakayama et al. [NS86] gezeigt, dass sich in der Visuellen Suche Konjunktionen von Tiefe und einem weiteren Feature im Gegensatz zu anderen Konjunktionen sehr effizient ausführen lassen.

Man kann also davon ausgehen, dass Tiefe eine gesonderte Rolle zwischen den üblichen Merkmalen wie Farbe oder Orientierung auf der einen Seite und der zweidimensionalen retinalen Position als Einheit der Selektion einnimmt. Um dem gerecht zu werden, wird Tiefe auf zwei Weisen im vorgestellten Modell verwendet. Zum einen als normales Merkmal zur Berechnung lokaler Salienz, dessen Umsetzung in diesem Abschnitt beschrieben wird. Zum weiteren wird Tiefe verwendet werden, um die vollständige Salienzrepräsentation mit Tiefeninformation zu versehen, was in Kapitel 5.5.6 beschrieben wird.

In beiden Fällen soll dabei keine akkurate dreidimensionale Rekonstruktion der Umgebung vorgenommen werden. Diese würde eine exakte Kalibrierung des Stereosystems voraussetzen und somit die Anforderungen an die Hardwareumgebung der vorgestellten Aufmerksamkeitssteuerung dramatisch erhöhen und die Einsetzbarkeit des Systems einschränken. Auch ist der damit verbundene Aufwand für ein präattentive Berechnung zu hoch. Vielmehr geht es um eine qualitative Bestimmung der Entfernung und vor allem der relativen Nähe. Das Vorgehen ist in Übereinstimmung mit Faugeras [Fau92], der selbst für stärker raumbezogene Aufgaben als die Aufmerksamkeitssteuerung, wie etwa die Roboternavigation, aufwendige Kamerakalibrierung als nicht notwendig identifiziert hat.

Als Merkmal ist die Tiefe eines Objektes entscheidend für Interaktionen mit Objekten. Aufgaben, zu deren Lösung Tiefeninformationen wesentlich beitragen kann, sind die Navigation, die Kollisions-

vermeidung, die Manipulation von Objekten, aber auch ihre Erkennung. Die Tiefe von Objekten ist weiterhin ein wichtiger Hinweis auf die Zusammengehörigkeit von Bildteilen. Ein homogenes Objekt besteht aus Oberflächen, deren Tiefe sich je nach Lage nur allmählich ändert. Sprünge in der Tiefe sind dagegen ein Hinweis auf unterschiedliche Objekte. Auf die Art kann Tiefe verwendet werden, Objekte vom Hintergrund zu trennen.

Die Entfernung eines Objektes kann auch als Maß dienen, das die Bedeutung oder Wichtigkeit bestimmt. Nahe Objekte interagieren im Normalfall eher mit dem Beobachter und müssen daher früher erkannt oder klassifiziert werden als entferntere Objekte. Diese Heuristik spiegelt sich in der Redensart „zuerst das Naheliegende" deutlich wieder. Ebensolches gilt für Bewegungen des Beobachters, für dessen Navigation und die Vermeidung von Kollisionen: die nahen Objekte haben eine größere Bedeutung. Auch für das natürliche Vorbild sind Nahrung und Verfolger um so wichtiger, je näher sie sich befinden.

Im Sinne der datengetriebenen Aufmerksamkeit (top-down) ist die gezielte Auswahl bestimmter Tiefenebenen von Bedeutung, deren Effizienz eine dreidimensionale Repräsentation der Salienzdaten voraussetzt. So wird es möglich, die Aufmerksamkeit gezielt auf einen Entfernungsbereich auszurichten und die Auswahl der auffälligsten Objekte auf solche, die sich in dieser Entfernung befinden, einzuschränken. Dieser Aspekt wird in den Kapiteln 5.5.6 und 6.4 weiter diskutiert, setzt jedoch in jedem Fall eine präattentive Bestimmung der Tiefeninformation für das gesamte Bild voraus.

5.4.2 Disparität

Merkmalsextraktion

In den Abschnitten 2.1.3 und 2.2.3 wurde die Bestimmung von Tiefe in natürlichen und technischen Systemen vorgestellt. Um Tiefe als Merkmal zu verwenden, ist zuerst die Dichte der Tiefendaten von entscheidender Bedeutung. Weiterhin ist wichtig, dass keine speziellen Anforderungen oder Einschränkungen bezüglich der beobachteten Umgebung möglich ist. Verfahren wie Tiefe aus Bewegung, Tiefe aus Textur und Tiefe aus Schattierung können diese allgemeine Tiefeninformation nicht zur Verfügung stellen. Somit stellt die Verwendung von Stereoinformationen den zu wählenden Weg dar.

Direkt korrelationsbasierte Verfahren haben Probleme mit der mangelnden Eindeutigkeit von Grauwerten zur Korrespondenzbildung wie mit der mangelnden Invarianz von Grauwerten durch den perspektivischen Unterschied und kommen deswegen nicht in Frage. Phasenbasierte Ansätze sind eingeschränkt hinsichtlich der detektierbaren Disparitäten. Daher fiel die Wahl auf einen merkmalsbasierten Ansatz. Das verwendete Merkmal sollte häufig genug vorhanden sein, um dichte Tiefeninformation zu erhalten. Weiterhin sollte es weitgehend invariant gegen die perspektivischen Bildunterschiede sein. Mit dem Hintergrund der Biologienähe fiel die Entscheidung zugunsten von Gaborfilterantworten aus. Dies macht es möglich, Zwischenergebnisse aus der Berechnung des Merkmales Symmetrie (s. 5.2.2), das ebenfalls auf der Berechnung von Gaborfilterantworten beruht, auszunutzen und so die Effizienz des Systems zu steigern.

Für die Berechnung von Tiefe sind ausschließlich Kanten mit vertikaler Komponente relevant, denn aus der horizontalen Komponente lässt sich aufgrund des Aperturproblems (visualisiert in Abb. 2.9) keine Disparität berechnen. Außer der vertikalen Antwort kommen also nur noch diagonale Orientierungen mit dominanter vertikaler Komponente in Frage. Jede zusätzliche Orientierung erhöht

die Dichte und Eindeutigkeit des Merkmales, gleichzeitig aber auch entsprechend den Rechenaufwand. Die Auswahl der Orientierungen wird durch die späteren Experimente determiniert.

Korrespondenzbildung

Zur Bestimmung der Disparität ist nun für jeden Orientierungskanal das Korrespondenzproblem zu lösen. Dazu wird ein Ähnlichkeitsmaß definiert, das in Anlehnung an Arbeiten von Trapp und Lieder [TDM95, Tra96, Lie99] festgelegt wird. Für die Berechnung ist davon auszugehen, dass die Epipolarlinien den horizontalen Bildzeilen entsprechen. Dies lässt sich durch entsprechende Rektifikation der Bilder bei bekannter Kamerageometrie einfach erreichen. Die Ähnlichkeitsfunktion entspricht einer leicht modifizierten Kreuzkorrelation:

$$\rho_{lr}(x,d) = \frac{w(x) * r_l(x) * r_r(x+d)}{\sqrt{w(x) * |r_l(x)|^2} * \sqrt{w(x+d) * |r_r(x+d)|^2}} \tag{5.13}$$

Dabei bezeichnet ρ_{lr} die Ähnlichkeit am Ort x für die Disparität d anhand der Gaborfilterantworten r_l und r_r für das linke bzw. rechte Bild und einer Fensterfunktion w, die der Einhüllenden des Gaborfilters entspricht. ρ_{lr}^{α} bezeichnet die Ähnlichkeit für die Orientierung α.

Nicht alle so berechneten Werte geben eine verlässliche Disparitätsinformation wieder. Am Bildrand finden sich Bereiche, für die die Gaborfilterantworten durch die Unvollständigkeit der zugrunde liegenden Eingabeinformation verfälscht ist. Diese Bereiche werden in der weiteren Verarbeitung unterdrückt. Weiterhin gibt es Bildbereiche, die nicht genügend Struktur und damit Information aufweisen, um die Ähnlichkeitsfunktion sinnvoll auszuwerten. Um sie auszuschließen, wird eine untere Schwelle festgelegt, über der die Varianz der Werte liegen muss, um in die Berechnung einzugehen. Die Festlegung des Schwellenwertes erfolgt empirisch. Abbildung 5.23 zeigt die Ergebnisse der Ähnlichkeitsfunktion für drei Orientierungen und 12 Disparitäten nach Ausschluss der ungültigen Werte. Das Beispielbild wird im Folgenden häufiger verwendet, da die Disparitätswerte weitgehend intuitiv überprüft werden können.

Weiterhin muss bestimmt werden, in welchem Disparitätsbereich die Ähnlichkeitsfunktion berechnet werden soll, um dort nach Korrespondenzen zu suchen. Der Bereich ergibt sich aus den erwarteten Objektentfernungen und der Abbildungsgeometrie auf der einen Seite und Erwägungen zum Berechnungsaufwand auf der anderen Seite. Aus der Abbildungsgeometrie lässt sich zuerst die Korrespondenz von Entfernung und Disparität herleiten, die vom horizontalen Abstand der Kameras, ihrem Öffnungswinkel und der Bildauflösung abhängt.

Salienzbestimmung

Die Korrespondenzberechnung liefert typischerweise jedoch keine eindeutigen Ergebnisse. Es bleiben vielmehr, wie Abb. 5.24 zeigt, für mehrere Orientierungen getrennt unter Umständen mehrere Kandidaten oder Hypothesen für die Disparitätswerte an jedem Ort. Diese Information soll für die spätere Lokalisation der Salienz in der Tiefe (s. 5.5.6), also die Erzeugung einer Repräsentation der aus allen Merkmalen gebildeten Auffälligkeit, genutzt werden.

Um zuvor jedoch die allein durch Tiefe induzierte Salienz zu bestimmen, muss für jeden Ort ein eindeutiger Wert aus den Tiefeninformationen ausgewählt werden. Die Spanne von Möglichkeiten

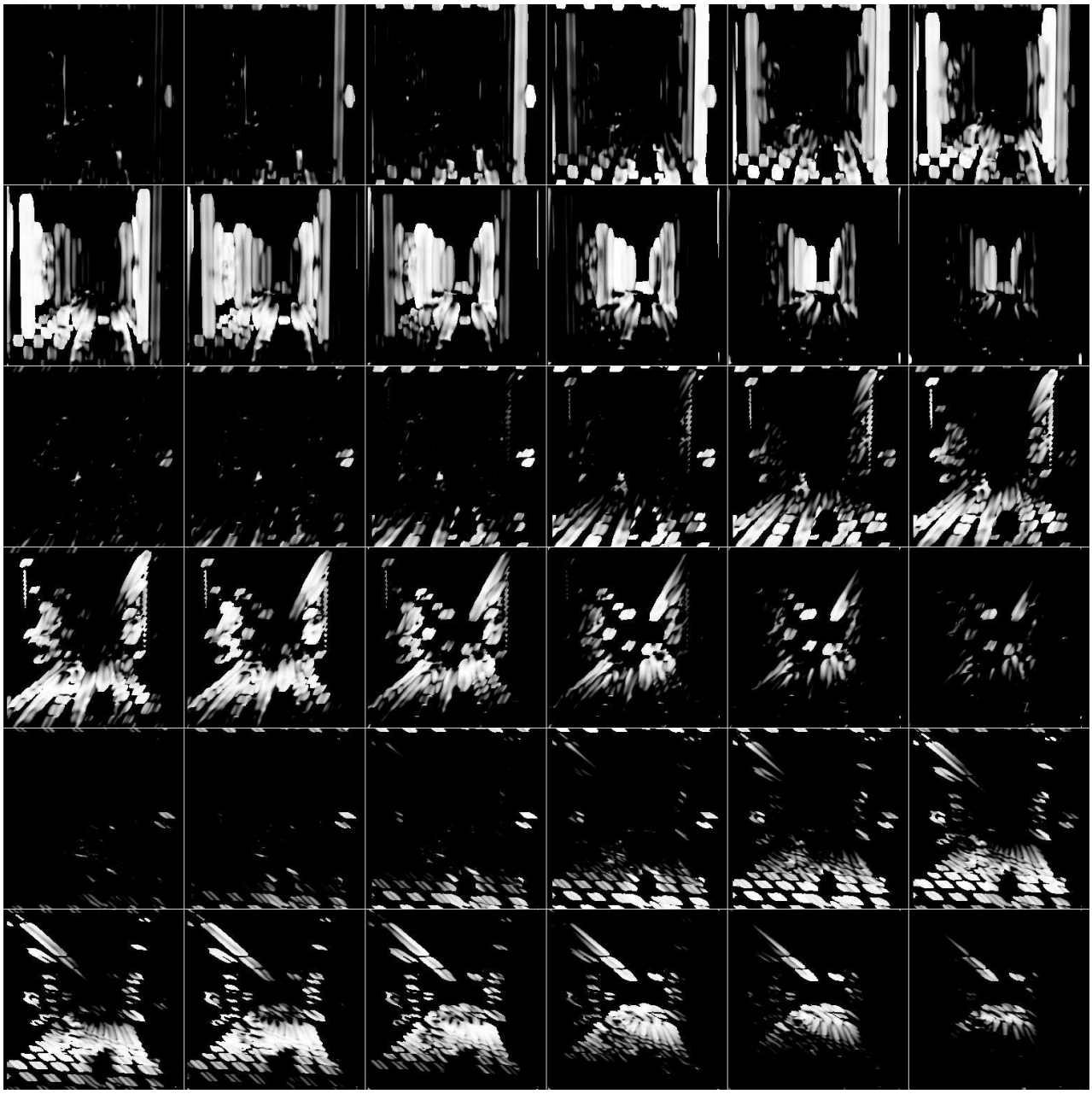


Abbildung 5.23: Für das unten gezeigte Stereopaar sind die Korrelationswerte für die drei Orientierungen (Reihe 1 und 2: senkrecht, Reihe 3 und 4: 30° nach rechts, Reihe 5 und 6 30° nach links) und in Leserichtung jeweils die Disparitäten von 11 bis 0 dargestellt. Es ist zu erkennen, dass zum Zentrum des Bildes hin die Korrelationen geringer werden, was der wachsenden Entfernung entspricht.

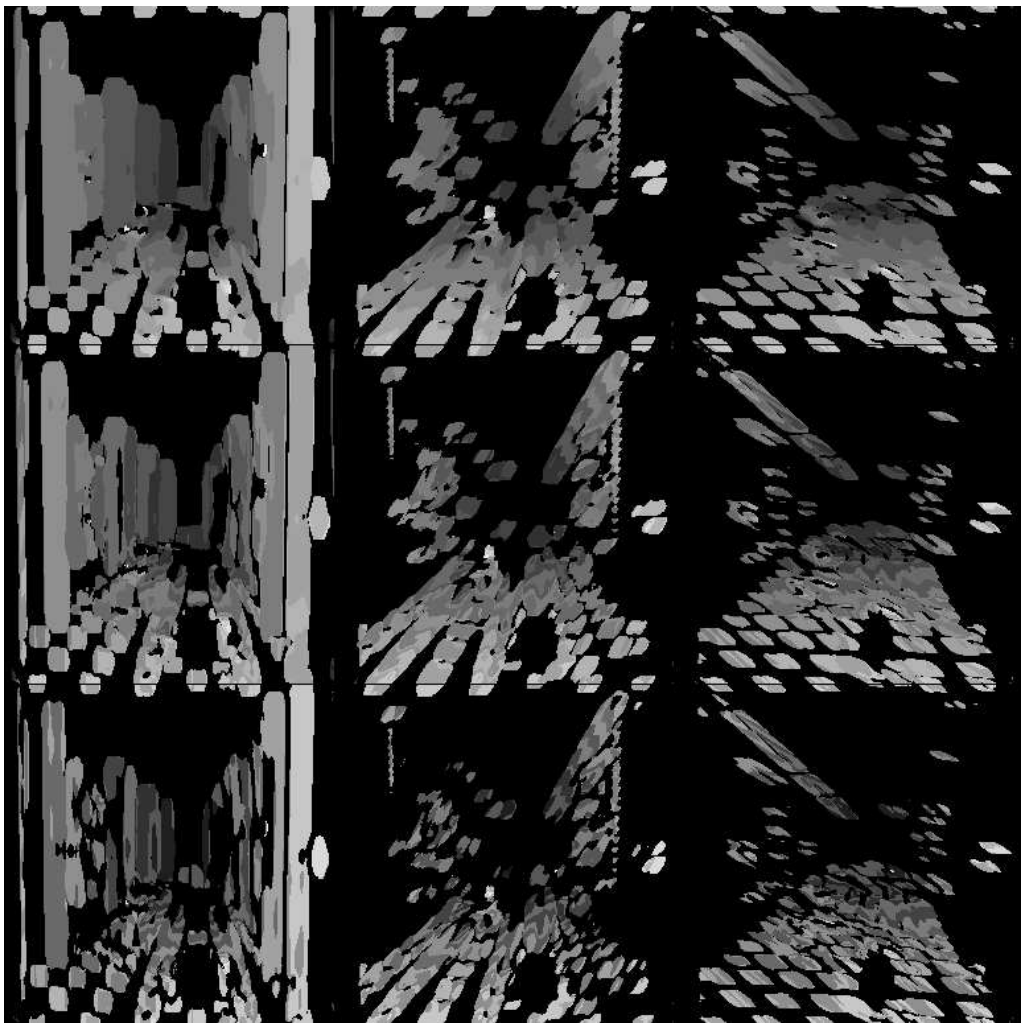


Abbildung 5.24: Für drei Orientierungen (links senkrecht; Mitte 30° nach rechts; rechts 30° nach links) sind für das Beispielbildpaar aus Abb. 5.23 von oben nach unten die drei besten Kandidaten für jeden Bildpunkt angegeben (schwarz: kein Disparitätswert, sonst Disparität steigend mit der Helligkeit).

geht von einer einfachen Selektion anhand des maximalen Korrelationswertes bis hin zu Selbstorganisationsverfahren, aus denen sich ein eindeutiges Maximum ergeben soll. Für Selbstorganisationsverfahren werden Hinweise auf die Plausibilität bestimmter Konfigurationen so umgesetzt, dass sich nach einigen Iterationen eindeutige Maxima für die Korrespondenzen ergeben. Als zu nutzende Hinweise kommen Eindeutigkeit, Vollständigkeit und Kontinuität in Frage. Experimente mit derartigen Verfahren [Lie99] zeigten, dass sie zwar geeignet sind, eine brauchbare Auswahl von Korrespondenzen durchzuführen, mit ihnen jedoch ein grundsätzlich sehr hoher Aufwand verbunden ist. Dieser Aufwand erscheint in diesem Kontext nicht angemessen, in dem es nicht um die Rekonstruktion von Tiefeninformationen zum Aufbau eines 3D-Modells der Szene geht, sondern alleine um die Bestimmung von Auffälligkeiten. Hier ist davon auszugehen, dass Fehleinschätzungen der Tiefe durch einfachere Verfahren in der weiteren Verarbeitung unterdrückt werden. Zu dieser Verarbeitung gehört eine räumliche und temporale Integration der Auffälligkeitswerte. Somit sollte die Leistungsfähigkeit der gesamten Aufmerksamkeitssteuerung nicht wesentlich beeinflusst werden.

Als Konsequenz wird ein nichtiteratives Verfahren zur Bestimmung der besten Korrespondenz durchgeführt, das ebenfalls die Kontinuität der Tiefe berücksichtigt. Die Nachbarschaft N eines Punktes geht gewichtet mit einer Gaußfunktion w_σ in das Maß der Zuverlässigkeit $conf(x,d)$ für eine Disparität d am Ort x ein, das auf den Korrelationsergebnissen ρ_{lr}^{α} für die Orientierung α beruht, die mit der Bedeutung b_α versehen sind:

$$conf(x,d) = \sum_{\alpha} \sum_{x' \in N(x)} w_\sigma(x - x') \cdot \rho_{lr}^{\alpha}(x', d) \cdot b_\alpha \tag{5.14}$$

Das Ergebnis für die beste Disparität an einem Ort ergibt sich als Maximum der Zuverlässigkeit:

$$\rho_{conf}(x) = d\,; \bigwedge_{d'} : \; conf(x,d') \leq conf(x,d) \tag{5.15}$$

Die Orientierungen werden für die Konfidenzbestimmung entsprechend ihres vertikalen Anteils gewichtet. Es wird also für jeden Ort die Disparität anhand der Korrelationsfunktion bestimmt und hieraus der normierte Salienzwert $feat_{depth}$ berechnet durch

$$feat_{depth}(x) = \frac{\rho_{conf}(x)}{maxdisp - mindisp} \tag{5.16}$$

, wobei $maxdisp$ und $mindisp$ den Suchbereich der vorkommenden Disparitätswerte eingrenzen und somit für eine Normierung der Salienzwerte sorgen. Abb. 5.25 zeigt die Auswirkung dieser Auswahl und der Salienzbewertung an einem Beispiel. Auch wenn einige wenige Fehlklassifikationen (etwa der kleine saliente Bereich im Bildzentrum) und einige Bereiche, in denen eine Tiefenbestimmung aufgrund mangelnder Strukturen nicht möglich war (etwa die Fläche der Kugel im Vordergrund oder Teile der Decke) auffallen, ist die Disparität in den allermeisten Bereichen korrekt bestimmt worden, wie sowohl die Abnahme der Salienz zum Zentrum hin, als auch die Ausnahmen im Bereich der sich vom Boden erhebenden Strukturen zeigen.

Es zeigt sich allerdings, dass die Breite der Gaborfilterantworten und damit der Fensterfunktion dafür sorgt, dass die berechneten Tiefenwerte eine gewisse Breite aufweisen. Selbst an einer lokal stark begrenzten vertikalen Kante würden sich Tiefenwerte rechts und links der Kante finden. Dies kann man als Schwäche des Verfahrens werten. Jedoch ist es zugleich eine Stärke, erlaubt es doch die

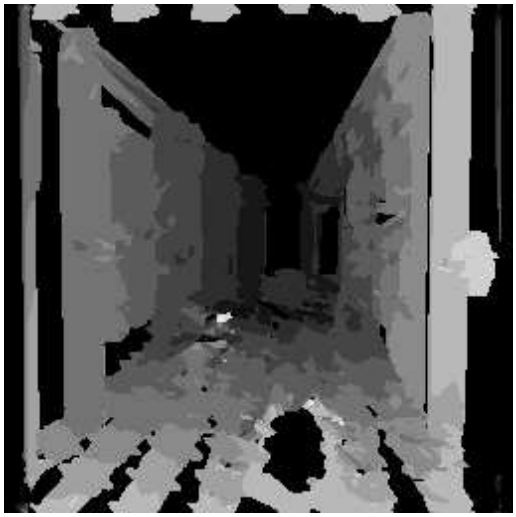


Abbildung 5.25: Salienz anhand des Merkmales Tiefe am Beispiel aus den Abb. 5.23 und 5.24

Ausdehnung der Bereiche, für die eine Tiefe bestimmt werden kann, sorgt somit für eine Verbesserung der Dichte der Tiefenkarte. Da in diesem Kontext der Dichte eine stärkere Bedeutung zukommt, ist der Effekt also durchaus gewünscht. Es wäre jedoch denkbar, das Verfahren um Berechnungen zu erweitern, die diese Kantenverbreiterung korrigieren.

Der Aufwand für die Berechnung des Stereomerkmals wird von der Korrespondenzbildung dominiert. Er wächst linear mit der Bildgröße, mit der Breite der Fensterfunktion und mit der Anzahl der möglichen Disparitäten, die überprüft werden müssen. Da diese wiederum der horizontalen Bildgröße proportional sind, erscheint eine Reduktion der Auflösung zur Beschleunigung besonders vielversprechend. Dies soll in der Multiskalenberechnung ausgenutzt werden, um gleichzeitig die Verlässlichkeit der Daten zu erhöhen und den Berechnungsaufwand zu vermindern.

Erweiterung auf Multiskalenberechnung

Kerngedanke der auf mehreren Skalen beruhenden Berechnung von Disparitäten ist die Einschränkung des Suchbereiches für Disparitäten bei höheren Auflösungen durch die Ergebnisse der Berechnungen für geringere Auflösungen. So lässt sich in der zeitkritischen maximalen Auflösung der Berechnungsaufwand deutlich senken.

Die Berechnung der Disparitäten findet dabei von der kleinsten Auflösung ausgehend so statt, wie zuvor beschrieben. Allerdings wird für jeden Bildpunkt in höheren Auflösungen der Konfidenzwert der niedrigeren Auflösung ausgewertet, um den Disparitätsbereich zu finden, in dem die Summe der Konfidenzen maximal wird. Die Breite des Disparitätsbereiches wird in der kleinsten Auflösung so gewählt, dass er dem relevanten Disparitätsbereich in der maximalen Auflösung entspricht. In der hier gewählten Parametrisierung mit zwei Skalen bedeutet dies, mit der Hälfte des gesamten relevanten Disparitätsbereiches in der geringsten Auflösung zu beginnen. Für n_{scale} Skalen bedeutet es, dass die Konfidenz für die niedrigst aufgelöste Skala wie bisher bestimmt wird $\rho_{lr}^{n_{scale}}(x, d)$. Von hier ausgehend werden für jeden Punkt der jeweils nächsten Skala die Grenzen für die Disparitätsberechnung so

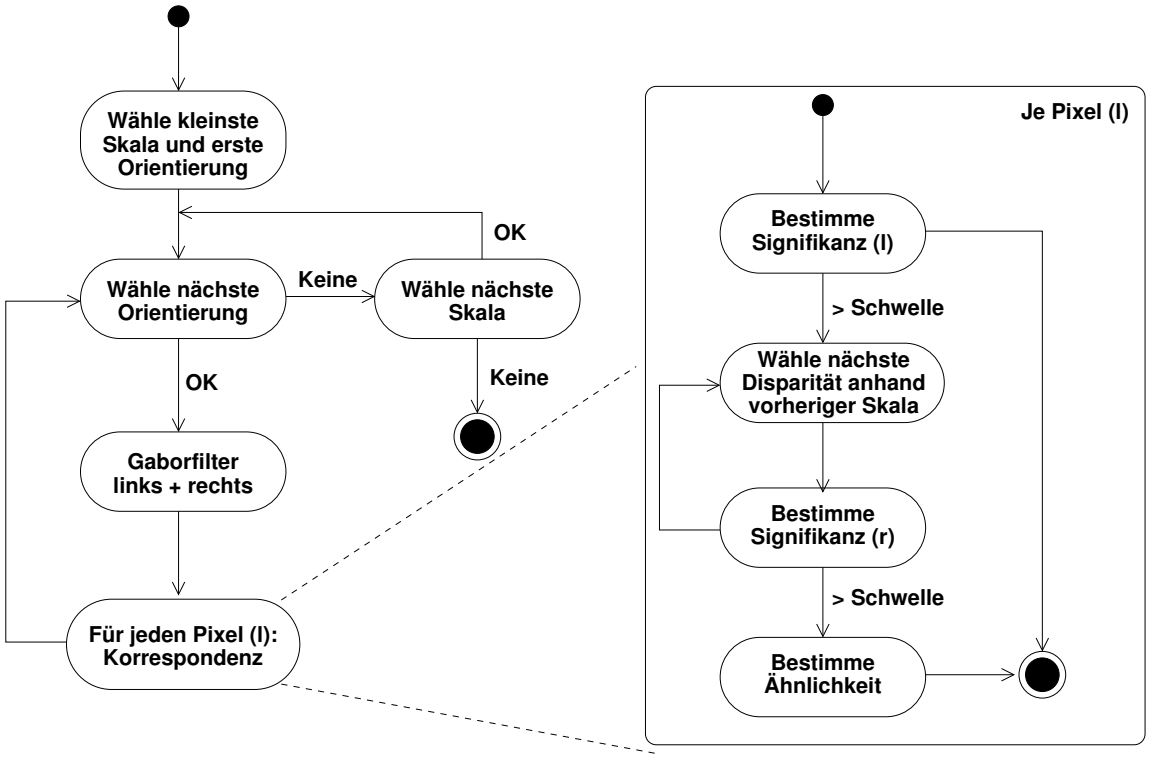


Abbildung 5.26: UML-Aktivitätsdiagramm zur Berechnung der Stereokorrespondenz anhand mehrerer Skalen.

bestimmt:

$$disp^i_{min}(x,d) = d : \bigwedge_{d'} f_{i-1}(d) >= f_{i-1}(d') \tag{5.17}$$

$$disp^i_{max}(x,d) = disp^i_{min}(x,d) + disp_{width} \tag{5.18}$$

$$f_i(d) = \sum_{d'=d}^{d+\frac{disp_{width}}{2}} \rho^i_{lr}(x,d') \tag{5.19}$$

Die Maximumsbildung zur endgültigen Bestimmung der Disparität summiert nun die Konfidenzen der verschiedenen Skalen. Der Ablauf der Korrespondenzbestimmung ist in Abb. 5.26 dargelegt, ein Beispiel zur Salienzberechnung gibt Abb. 5.27 wieder.

5.4.3 Experimente

Zur Überprüfung der Merkmalsberechnungen wurde in einem Bild die Tiefe eines Objektes variiert. Diese Veränderung sollte einen entsprechenden Einfluss auf die Merkmalskarte haben und als Konsequenz eine Variation in der Salienz bewirken (s. Abb. 5.28). Abb. 5.29 zeigt weitere Beispiele für die Bestimmung des Stereomerkmals. In der letzten Zeile wurde ein random-dot-Stereogramm eingesetzt (s. Kap. 2.1.3). Die Robustheit gegen Rauschen ist Abb. 5.30 zu entnehmen. Zwar nehmen die Variationen des Stereomerkmales mit dem Rauschen zu, jedoch ist der Effekt so, dass sich die Schätzung der Disparität nur in einem sehr kleinen Bereich verändert. Dies ist auf die Multiskalenberechnung zurückzuführen, die die Schätzung der Disparität für höhere Auflösungen einschränkt.

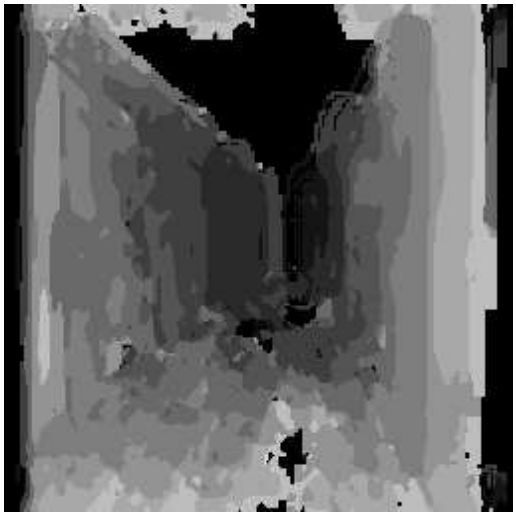


Abbildung 5.27: Salienz anhand des Merkmales Tiefe bei Multiskalenberechnung am Beispiel aus den vorhergehenden Abbildungen.

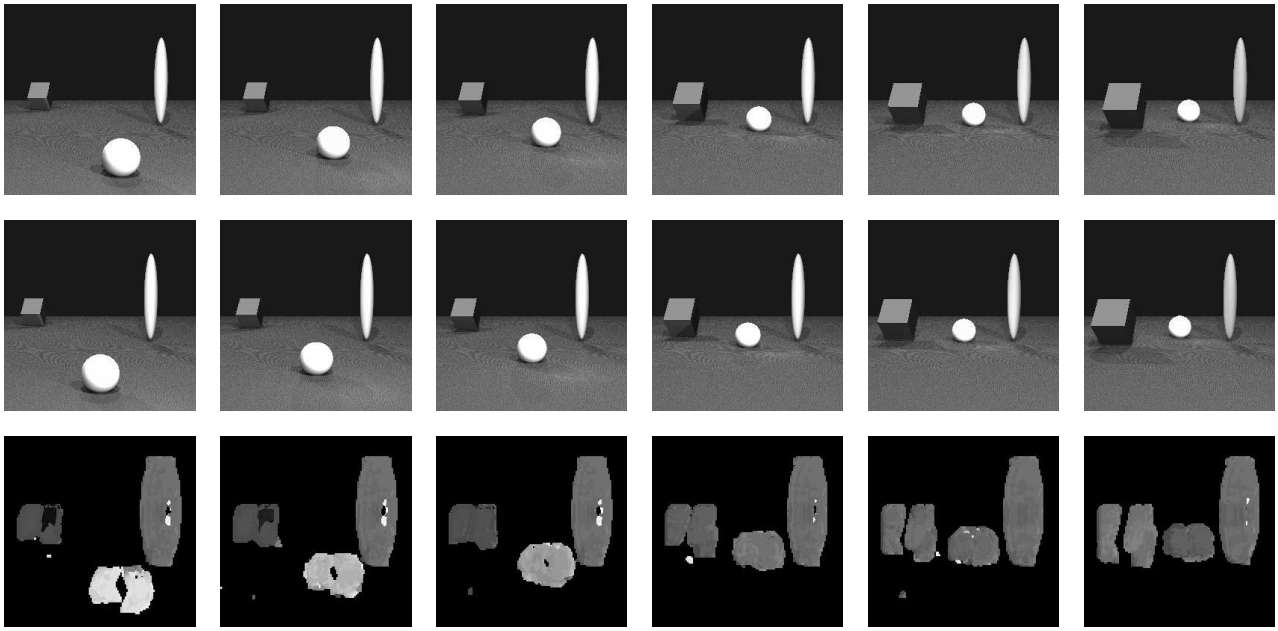


Abbildung 5.28: Variation der Entfernung und Effekt bezüglich des korrespondierenden Merkmals. Die obere Zeile zeigt jeweils das linke, die mittlere das rechte Kamerabild, das Ergebnis der Berechnung ist in der unteren Zeile dargestellt.

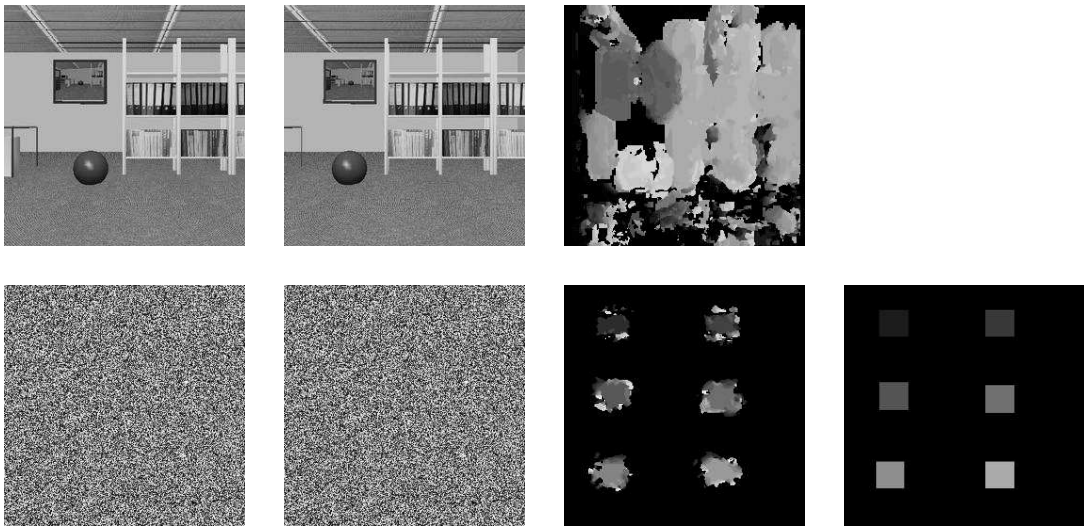


Abbildung 5.29: Beispiele unterschiedlicher Domänen für die Berechnung des Stereomerkmales. Links und in der Mitte sind die Eingabebilder (linke und rechte Kamera) dargestellt, rechts das Ergebnis der Berechnung. Für das random-dot-Stereogramm wird zusätzlich (ganz rechts) die verwendete Tiefenkarte angegeben.

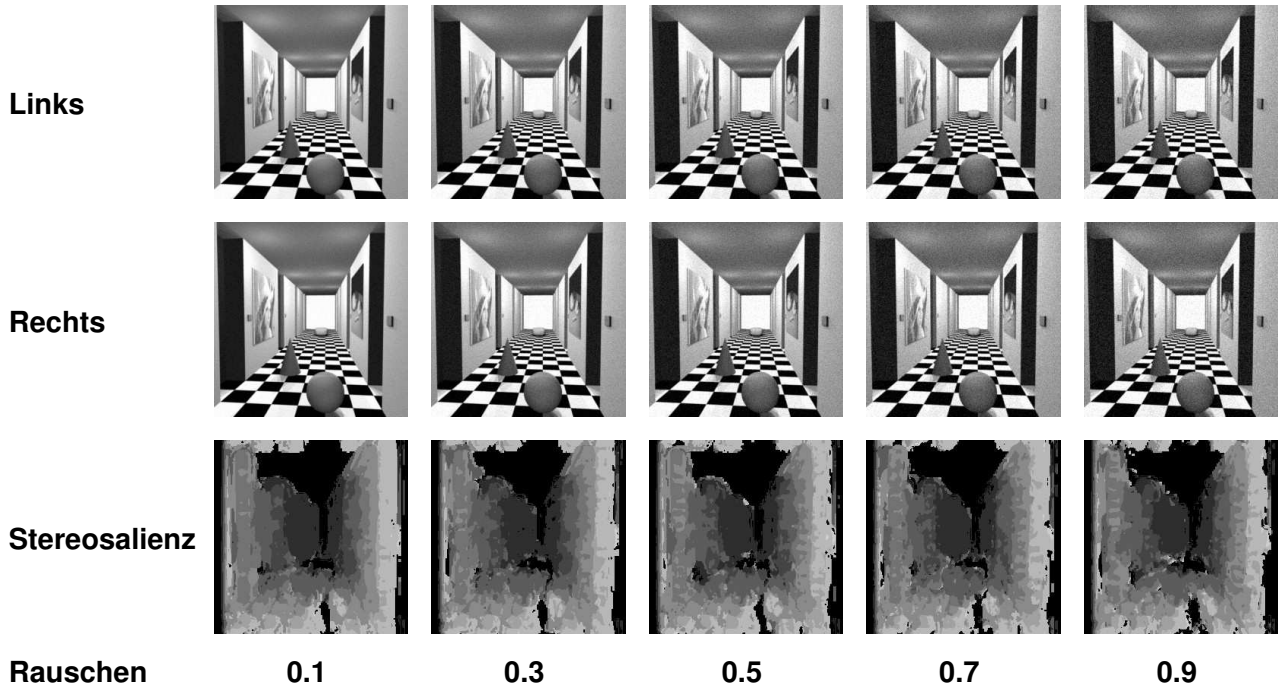


Abbildung 5.30: Der Einfluss von normalverteiltem Rauschen auf die Stereomerkmalsberechnung. Das Rauschen ist auf den beiden Stereobildern jeweils unabhängig.

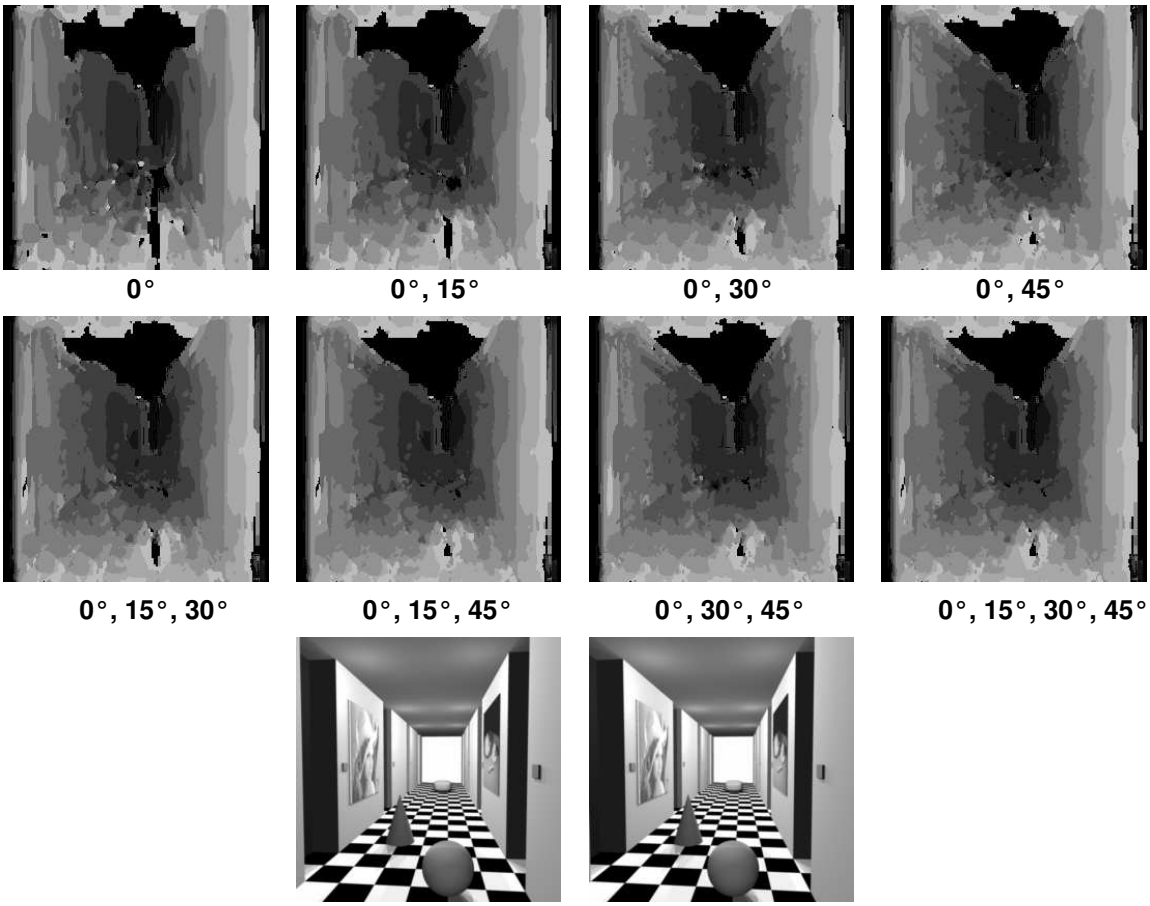


Abbildung 5.31: Auswirkung der Verwendung unterschiedlicher Orientierungen für die Gaborfilterung auf die Stereoberechnung.

Als wesentliche empirisch bestimmte Parameter der Merkmalsberechnung sind die Anzahl der Orientierungen und die untere Schwelle für die Varianz des Signals anzusehen. Als Orientierung für die Filterung kommen, wie bereits diskutiert, nur solche mit vertikaler Komponente in Frage. Die rein senkrechte Orientierung sollte dabei wegen ihrer Eignung und ihrer Häufigkeit (speziell in von Menschen gestalteten Umgebungen) verwendet werden, außerdem aus Symmetriegründen jeweils Paare von nach links und rechts abweichenden Orientierungen. Eine Auswahl von Orientierungen und die Ergebnisse der Merkmalsberechnung sind Abb. 5.31 zu entnehmen. Es ist zu sehen, dass das Verfahren zwar von der Verwendung mehrerer Orientierungen profitiert, die Ergebnisse sich jedoch außer für den Fall nur einer Orientierung nur leicht unterscheiden.

Auch der Schwellwert wurde experimentell verändert und der Effekt auf die Berechnung in Abb. 5.32 dargestellt. Es ist festzustellen, dass erst bei sehr hohen Schwellwerten einige korrekte Ergebnisse ausgelassen werden und bei sehr niedrigen Schwellwerten die Anzahl der fehlerhaft klassifizierten Pixel zunimmt (z.B. im Bildzentrum oder auf der Oberfläche des Balls im Vordergrund).

5.5 Integration der Merkmale

Die Existenz mehrerer Hinweise auf die Salienz verlangt nach einer Integration. Diese Integration der Merkmale sollte so stattfinden, dass eine informative Repräsentation der Salienz entsteht. Diese

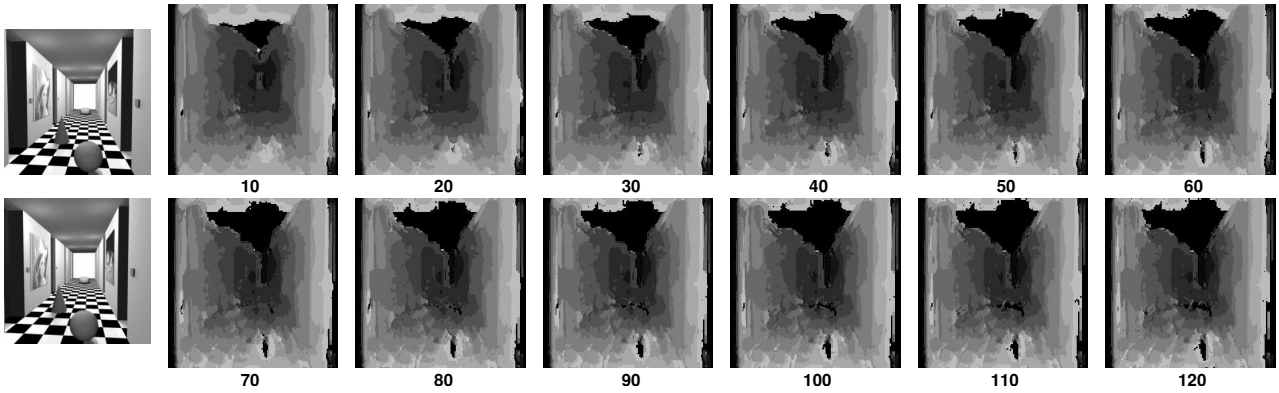


Abbildung 5.32: Veränderung der Salienzberechnung in Abhängigkeit des Varianzschwellwertes.

Repräsentation ist aber nicht ganz unabhängig von der Art der Selektion, die anhand dieser Repräsentation geschehen soll. Daher wird die Diskussion der unterschiedlichen Möglichkeiten hier nur im Hinblick auf die Informativität geführt und in Kapitel 6 fortgesetzt. Itti und Koch [IK01b] untersuchten folgende vier Alternativen zur Integration:

- Normalisierte Summation
- Linearkombination mit gelernter Gewichtung
- Globale nicht-lineare Normalisierung mit Summation
- Lokaler nicht-linearer Wettbewerb mit Summation

5.5.1 Getrennte Verwendung der Merkmale

Ebenso wie Modelle, die die Auffälligkeit anhand eines einzelnen Merkmals bestimmen, gibt es Modelle, die auf die eigentliche Integration mehrerer Merkmale verzichten und Selektionskandidaten nur anhand der Salienzinformation der Merkmale getrennt bestimmen. Dies war in der ursprünglich in NAVIS [Bol00, MBHS99] verwendeten Aufmerksamkeitssteuerung der Fall. Für jedes verwendete Merkmal wurden hier diskrete Aufmerksamkeitspunkte bestimmt und mit zusätzlichen Informationen annotiert. Diese Informationen enthielten die 2D-Position, das Maß an Auffälligkeit, das Merkmal und eine Größe für den Bereich, auf den sich diese Salienz bezog. Dabei wurde die räumliche Relation zwischen Aufmerksamkeitspunkten unterschiedlicher Merkmale ignoriert, weswegen man eigentlich nicht von einer Integration sprechen kann. Dieses Verfahren könnte man durch eine Maximumsuche innerhalb der Auffälligkeitskarten auch für die hier vorgestellten Merkmale durchführen, profitiert davon jedoch höchstens im Zusammenhang mit den darauf zugeschnittenen Methoden der Objekterkennung.

Ebenso getrennt werden die beiden Merkmale Tiefe und (horizontaler) Bildfluss bei Maki [MNE00, Mak96] verwendet. Das ausgewählte Segment ergibt sich anhand einer der beiden Eigenschaften, ein übergeordnetes System ist für die Auswahl zuständig.

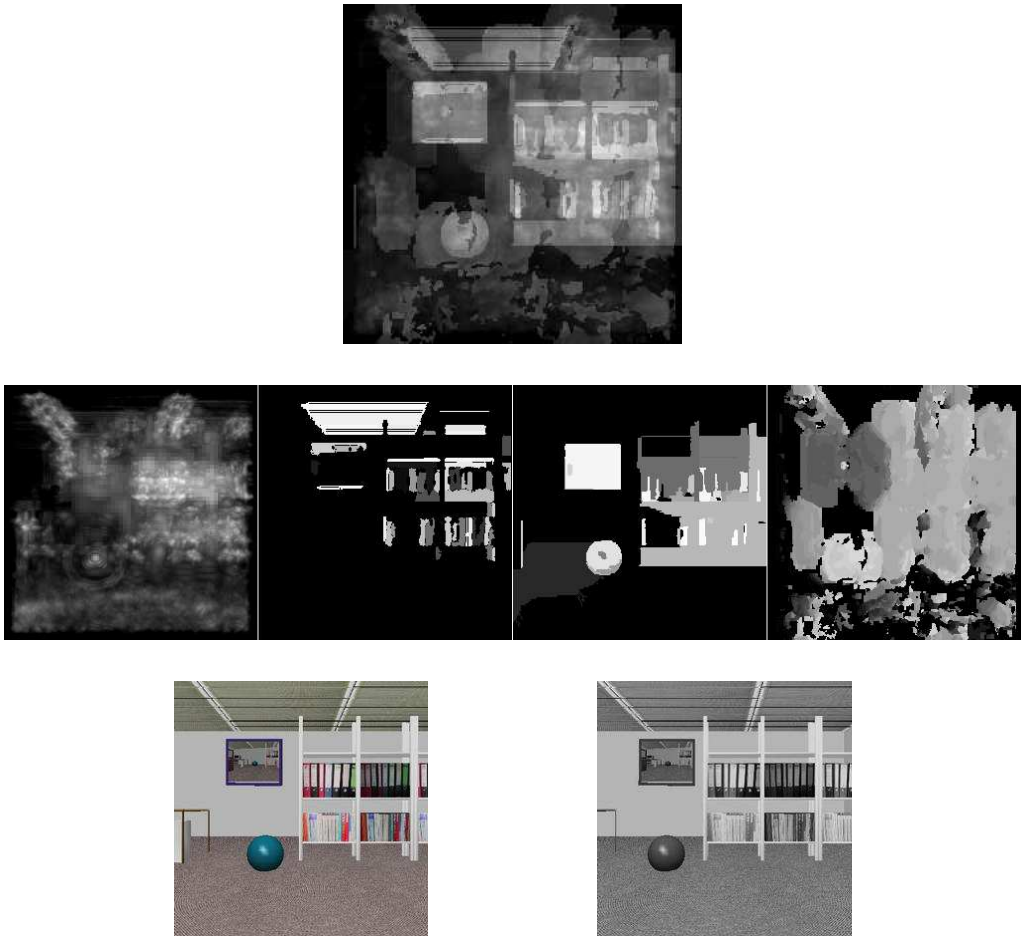


Abbildung 5.33: Superposition der Merkmale an einem Beispiel. Für die beiden Eingabebilder in der unteren Zeile wurden die vier zuvor beschriebenen Merkmale (von links nach rechts: Symmetrie, Exzentrizität, Farbkontrast, Tiefe) berechnet (mittlere Zeile), die additiv in der oben abgebildeten Mastermap zusammengefasst wurden. Abgesehen von Tiefe wurden die Merkmale nur für das linke Eingabebild berechnet.

5.5.2 Gewichtung der Merkmale

Die einfachste Möglichkeit zur Integration der Merkmale, so sie als Salienzkarten vorliegen, ist ihre Superposition. Vorausgesetzt, dass die Merkmalssalienzen einen vergleichbaren Wertebereich annehmen, reicht eine gewichtete Addition als Mechanismus aus. Resultat einer solchen Operation ist eine zweidimensionale *master map of attention*, wie sie viele Modelle (unter anderem [CF92, CRD96, FH96, IK00, KU85, Kop96, MS99, WCF89, Wol94, WG96]) vorsehen. Dieses Vorgehen basiert darauf, dass das Maß an Auffälligkeit für alle Merkmale in einer vergleichbaren Form vorliegt.

Eine Superposition der Merkmale anhand vorab definierter Gewichte fw_i

$$mm(x) = \sum_i fw_i * feat_i(x) \tag{5.20}$$

würde in dem verwendeten Beispiel mit identisch gesetzten Gewichten $fw_i \equiv 1$ das in Abb. 5.33 dargestellte Resultat liefern.

Das Lernen oder Adaptieren von Gewichten setzt voraus, dass eine Rückmeldung erfolgt, inwieweit ein Maximum in der Mastermap zu einem gewünschten oder einem unerwünschten Ziel gehört. Während dies bei der Abarbeitung einiger Aufgaben möglich ist, so etwa für die Visuelle Suche, fehlt solch eine Rückmeldung in einer normalen Exploration. Damit gehört dieser Aspekt in den aufgabenabhängigen Teil des Systems, der in Kapitel 8.5 diskutiert wird. Dies geschieht in Übereinstimmung mit dem Guided Search-Modell von Wolfe [WCF89, Wol94, WG96], das in der Gewichtung der Merkmale eine Aufgabe sieht, die primär top-down also zielgetrieben gelöst wird. So werden bei der Durchführung einer Visuellen Suche gerade die Merkmale hoch gewichtet, die eine Unterscheidung von Zielreiz und Ablenkern ermöglichen.

5.5.3 Bewertung der Exklusivität

Ein Problem der einfachen Integration der Merkmale besteht darin, dass die Exklusivität oder der Kontrast der Merkmale nicht in die Ergebnisse eingeht. So können leichte Schwankungen eines Merkmales, das im ganzen Bild stark vertreten ist, den einen Bereich verdecken, der anhand eines anderen Merkmales wesentlich auffälliger als der Rest ist. In der Visuellen Suche wird der Effekt, dass ein Objekt, das einzigartig ist, wesentlich auffälliger erscheint, als eine Gruppe gleichartiger Objekte als *odd-man-popout* bezeichnet. Er lässt sich leicht veranschaulichen, wenn man sich zum Beispiel ein einzelnes schwarzes Schaf in einer Herde weißer Tiere vorstellt

Itti und Koch [IK01b] konnten zeigen, dass eine Nichtlinearität in der Bevorzugung von Merkmalen mit lokal hoher Salienz gegenüber solchen mit breit verteilter gleichmäßiger Salienz einen Vorteil mit sich bringt. Die hier verwendete Methode soll expliziter die Exklusivität bewerten, so dass häufig auftretende Merkmale unterdrückt werden. Dazu wird für jedes Merkmal getrennt bestimmt, welche Ausprägungen der Merkmalseigenschaften wie häufig vorhanden sind.

Allerdings ist die Exklusivität dabei nicht für jedes Merkmal bestimmbar. Im vorgestellten Modell macht es etwa für das Merkmal Symmetrie keinen Sinn, eine Exklusivität zu bestimmen, da sich alleine ein Maß an Salienz bestimmen lässt, jedoch keine Unterteilung in unterschiedliche Typen von Symmetrie. Für die übrigen Merkmale ist dies jedoch möglich, wie im Folgenden gezeigt wird.

Die Orientierung der Segmente beim Merkmal Exzentrizität ist geeignet, eine Kategorisierung der Flächensegmente vorzunehmen. Eine Einteilung der Segmente in Bereiche zu jeweils 15° (s. Abb. 5.9) mit einer zusätzlichen Karte für Segmente ohne deutliche Orientierung dient als Basis. Für jede dieser Kategorien i wird die Anzahl der enthaltenen Segmente n_i bestimmt. Die Salienz des Segmentes wird zur Bewertung der Exklusivität durch $c_{exkl}^{n_i}$ dividiert. Dabei beschreibt c_{exkl} einen Parameter, der die Stärke des Exklusivitätsbewertung einstellt. Er muss mindestens 1 sein (dann erfolgt keine Bewertung der Exklusivität) und ist in allen Experimenten für alle Merkmale auf 1.1 festgelegt. Vergleichbar zur Exzentrizität bietet sich die Einteilung der Elemente nach Farbtönen (s. Abb. 5.18) für das Merkmal Farbkontrast an, aus der sich die Exklusivität analog durch Division ableiten lässt.

Für die Tiefe gibt es keine diskrete Anzahl von Segmenten, die eine den zuvor beschriebenen Merkmalen entsprechende Exklusivitätsbewertung erlaubt. Stattdessen wird direkt die Anzahl der Pixel np_i in jeder Disparitätsstufe i verwendet. Alle Pixel der Disparitätsstufe werden durch $c_{exkl}^{\frac{np_i}{np}}$ dividiert, so dass Disparitätsstufen mit besonders vielen Pixeln geringere Salienzwerte erhalten.

Abb. 5.34 zeigt für ein Beispiel jeweils die direkte Bestimmung der Salienz und zusätzlich den Effekt der Bewertung von Exklusivität anhand der drei Merkmale, die mit einer solchen Bewertung

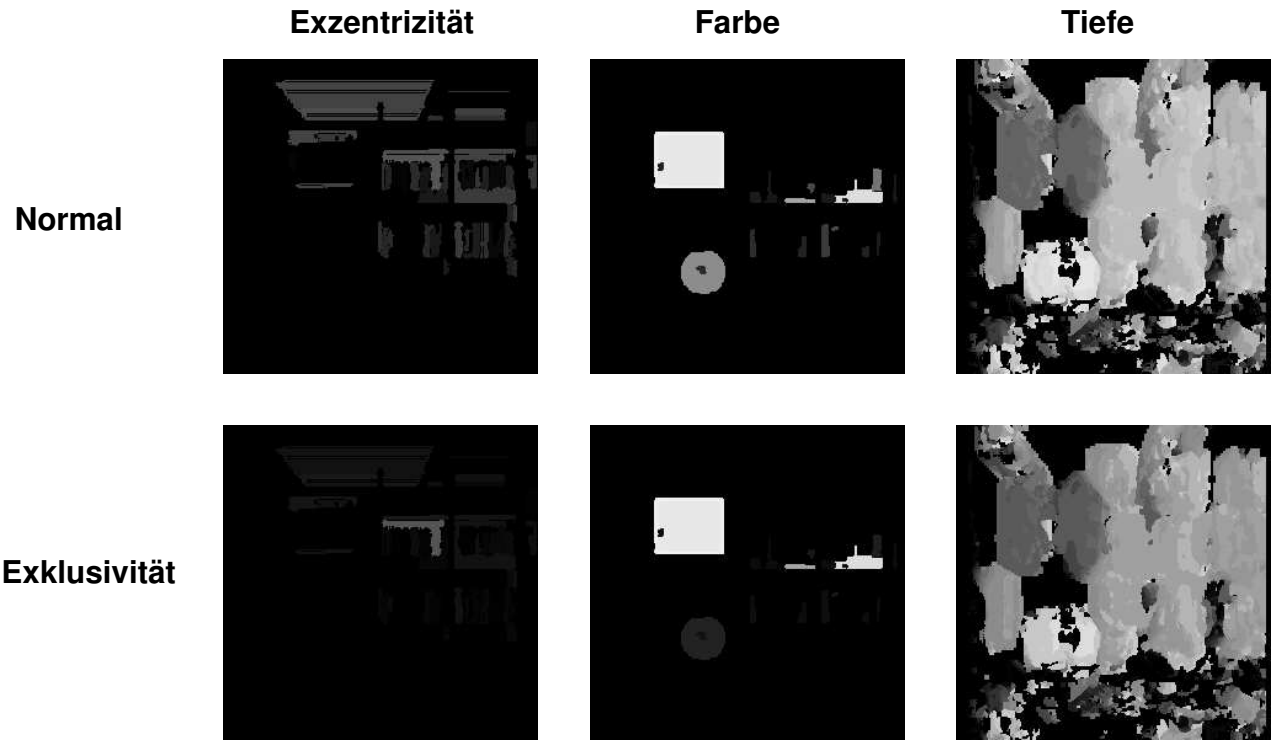


Abbildung 5.34: Effekt der Exklusivität auf die Berechnung der einzelnen Merkmale.

ausgestattet sind. Im Beispiel ist zu erkennen, wie etwa für die Tiefe die Bereiche, in denen sich das große Regal und die Tischkante befinden, durch die Bewertung der Exklusivität in ihrer Salienz reduziert werden. Dasselbe gilt bei der Exzentrizität für die senkrecht orientierten Bücher im Regal, sowie beim Farbkontrast für den Ball und die ähnlich gefärbten Bücher.

Dieser Mechanismus entspricht somit einer Umsetzung des odd man popout beim natürlichen Vorbild. Problematisch ist jedoch, dass gleichzeitig eine Form der Rauschverstärkung vorgenommen wird. Werden fälschlich Elemente detektiert, die in dieser Form im ganzen Bild nicht vorhanden sind, werden sie durch die Bewertung der Exklusivität noch verstärkt. Daher ist es wichtig, dass das Verfahren in dieser Form mit anderen Verfahren kombinierbar ist. Erreicht wird dies, da es sich um eine reine Bewertung innerhalb des Merkmales handelt und man als Resultat zuerst eine modifizierte Form der Merkmalssalienz erhält. In einer konkreten Anwendung und Umgebung ist empirisch zu bestimmen, welche Stärke der Exklusivitätsbewertung in Kombination mit welchen weiteren Verfahren zur Merkmalskombination die besten Resultate liefert.

5.5.4 Konditionale Verknüpfung

Weniger in allgemeinen Modellierungen visueller Aufmerksamkeit als in technischen Systemen zur Integration mehrerer Cues findet eine Verknüpfung der Cues in der Art statt, dass ein Typ von Cues nur ausgewertet wird, wenn ein anderer Cue eine Schwelle überschreitet. Dies ist vor allem dann interessant, wenn bestimmte Merkmale nicht für das ganze Bild ausgewertet werden sollen, sondern zur Einsparung von Rechenaufwand nur dort, wo ein anderer, einfacher zu berechnender Cue darauf hinweist, dass hier ein hoher Salienzwert zu erwarten ist. Diese Verknüpfung lässt sich von der Verwendung zweier Merkmale auf verschiedene Arten erweitern.

Braumann [Bra01] verknüpft die verwendeten Merkmale zur Gesichtsdetektion durch Fuzzy-Operationen, die kein arithmetisches Mittel, aber auch keine Maximumsselektion durchführen, sondern parametrisierbar dazwischen arbeiten. Im allgemeinen ist dieser Ansatz dann geeignet, wenn es nicht um unterschiedliche Quellen von Salienz geht, sondern um eine speziellere Eigenschaft, für deren Präsenz mehrere Hinweise ausgewertet werden können. Dies ist jedoch nicht in der Bestimmung allgemeiner Auffälligkeit der Fall, wie sie in diesem Modell vorgenommen wird. Sie wird daher hier nicht weiter verfolgt.

5.5.5 Multiple Gewichte

Um die Probleme mit der Gewichtung der einzelnen Merkmale, wie man sie in der Literatur findet, zu umgehen, wird im Rahmen des vorgestellten Modelles versucht, mehrere Gewichtungen gleichzeitig zu verwenden. Dies führt prinzipiell zu mehreren Auffälligkeitsrepräsentationen in der Form jeweils einer *master map*, in denen unter Umständen auch unterschiedliche Rangfolgen der auffälligen Bereiche auftreten. Offen ist für einen solchen Ansatz zweierlei:

- Wie bestimmt man die Gewichte?
- Wie werden die unterschiedlichen resultierenden Auffälligkeitskarten weiter verwendet?

Als Einflüsse auf die Gewichte kommen zielgetriebene Präferenzen, statistische Auswertungen der Merkmalssalienzen, die Historie der Selektion und die jeweils anderen Gewichte in Frage. Die Diskussion der Verwendung stellt einen gewissen Vorgriff auf die im folgenden Kapitel vorgestellte Selektionsstufe dar, die wie zuvor erwähnt die Repräsentation beeinflussen kann. In dieser Architektur spielt die Selektion mehrerer Einheiten eine wichtige Rolle. Für diese ist es naheliegend, die mehrfachen Repräsentationen zu nutzen, um anhand jeder Karte ein (anderes) Maximum auszuwählen. Ist das einmal geschehen, ergibt sich auch eine Antwort auf die erste Frage. Die Gewichte können jetzt nämlich anhand der Merkmale bestimmt werden, die im ausgewählten Bereich im Gegensatz zur mittleren Präsenz der Merkmale vorhanden sind. Die Auswertung einer Art von Signal-Rausch-Verhältnis führt so zu Gewichten, die die Auswahl stabilisieren.

In [Bac98] wurde ein derartiger Ansatz vorgestellt. Er beruht auf einer Selektion von einzelnen Bereichen hoher Salienz anhand getrennter Gewichtungen der Merkmalssalienzen. Die Art und Weise, auf die diese Auswahl durchgeführt wird, ist an dieser Stelle noch nicht von Bedeutung und wird im folgenden Kapitel beschrieben. Zu Beginn werden die Gewichte so initialisiert, dass für jeden Satz von Gewichten ein anderes Merkmal das höchste Gewicht erhält und alle Merkmale in der Summe etwa gleich gewichtet werden. Findet nun eine Selektion statt, wird bestimmt, wie stark die verschiedenen Merkmale im ausgewählten Bereich im Vergleich zum gesamten Bild präsent sind. Dabei stellt sich heraus, welche Merkmale wie sehr zur Selektion des Bereiches beitragen. Die Gewichte werden nun in genau diese Richtung beeinflusst, was zu einer Stabilisierung der Selektion führt. Weiterhin wird dafür Sorge getragen, dass die Gewichtung zu einer Unterscheidung der selektierten Objekte beiträgt, vor allem dann, wenn sich diese Objekte in räumlicher Nähe befinden. Dazu wird eine Abstoßung der Gewichtssätze eingeführt, deren Stärke genau mit der Nähe der selektierten Bereiche wächst.

Die Beschreibung einer Variante dieses Verfahrens wird in Kapitel 6.3.2 vorgenommen. Das ist notwendig, da sie stark von der weiteren Verarbeitung der Salienz durch Neuronale Felder abhängt, die erst im folgenden erläutert wird.

5.5.6 Dreidimensionale Repräsentation

Wie bereits in 3.2.4 diskutiert, stellt die Tiefe insofern einen Sonderfall dar, als sie als Merkmal dienen kann, aber auch eine räumliche Dimension beschreibt. Man kann sie dazu verwenden, anstelle einer zweidimensionalen *master map* eine dreidimensionale Karte zu erstellen, die dann eine Adressierung und räumliche Selektion auch anhand der dritten Dimension zulässt. Das Ziel ist dabei nicht, eine 3D-Rekonstruktion der Szene zu erstellen, vielmehr geht es um eine eher qualitative Einordnung in nahe und ferne Objekte und eine Einteilung in einige Tiefenebenen, die die Relation zwischen den Objekten erschließt.

Weiteres Argument für die Nutzung einer solchen Repräsentation ist, dass die dafür benötigten Tiefendaten ohnehin durch die Berechnungen im Rahmen des stereobasierten Merkmals (s. 5.4) bereit gestellt werden. Hier ist jedoch die Festlegung einer einzelnen Disparität nicht entscheidend, vielmehr können die nachfolgenden Mechanismen gerade eine Unterdrückung sporadisch auftretender Fehlklassifikationen ausgleichen, indem der räumliche und temporale Kontext mitbetrachtet wird. Daher wird hier wieder von der Ähnlichkeitsfunktion anhand unterschiedlicher Orientierungen ausgegangen, die für jede Disparität mehrere Hinweise auf ihre Gültigkeit an einem Ort angibt (s. Gleichung 5.13).

Ziel des Vorgehens ist eine dreidimensionale Entsprechung der zweidimensionalen *master map*. Die Tiefeninformationen werden also nicht in einer zweidimensionalen Karte annotiert, sondern es wird eine dreidimensionale Karte für die Lokalisation der Salienz in allen drei räumlichen Dimensionen gebildet. Dabei geht es nicht darum, eine hochwertige Rekonstruktion der Tiefe vorzunehmen, die zur Interpretation oder Erkennung von Objekten oder auch zur Navigation notwendig wäre. Vielmehr soll die Karte eine grobe Unterteilung der Objekte in nähere und weiter entfernte erlauben oder auch die Trennung von Bereichen anhand ihrer Tiefe. Dafür reicht es aus, eine vergleichsweise kleine Anzahl von Tiefenebenen zu verwenden, um die relative Tiefe der Objekte auszuwerten. Die genaue Spezifikation ist natürlich auch abhängig von der weiteren Verarbeitung, die in Kap. 6 erläutert wird. Die Karte entsteht nun aus einer Faltung der Salienzwerte mit den Ähnlichkeitswerten an jeder 2D-Bildposition, wobei je nach Anzahl untersuchter Disparitäten und verwendeter Tiefe des Neuronalen Feldes eine Streckung oder Stauchung der Ähnlichkeitswerte vorgenommen werden muss:

$$\begin{aligned} S^{3D}(x,d) &= S(x) * (disp(x,d) + \frac{1 - \sum_d disp(x,d)}{nfsize_z}) \\ disp(x,d) &= \sum_b \rho_{lr}^{\alpha}(x,d) * b_\alpha \end{aligned} \tag{5.21}$$

Die Gewichtung der Ähnlichkeitswerte anhand der Orientierungen mit der Bewertung der vertikalen Komponente entspricht der vorgenommenen Gewichtung bei der Berechnung des Stereomerkmales. Ein Beispiel zeigt Abb. 5.35.

5.6 Zusammenfassung und Diskussion

In diesem Kapitel wurden Vorgehensweisen zur Berechnung lokaler datengetriebener Auffälligkeit vorgestellt. Diese zeichnen sich gegenüber den in der Literatur vorgestellten Verfahren durch ihre Allgemeinheit, ihre Objektbezogenheit, ihre Invarianz, Informativität, die Nähe zum natürlichen

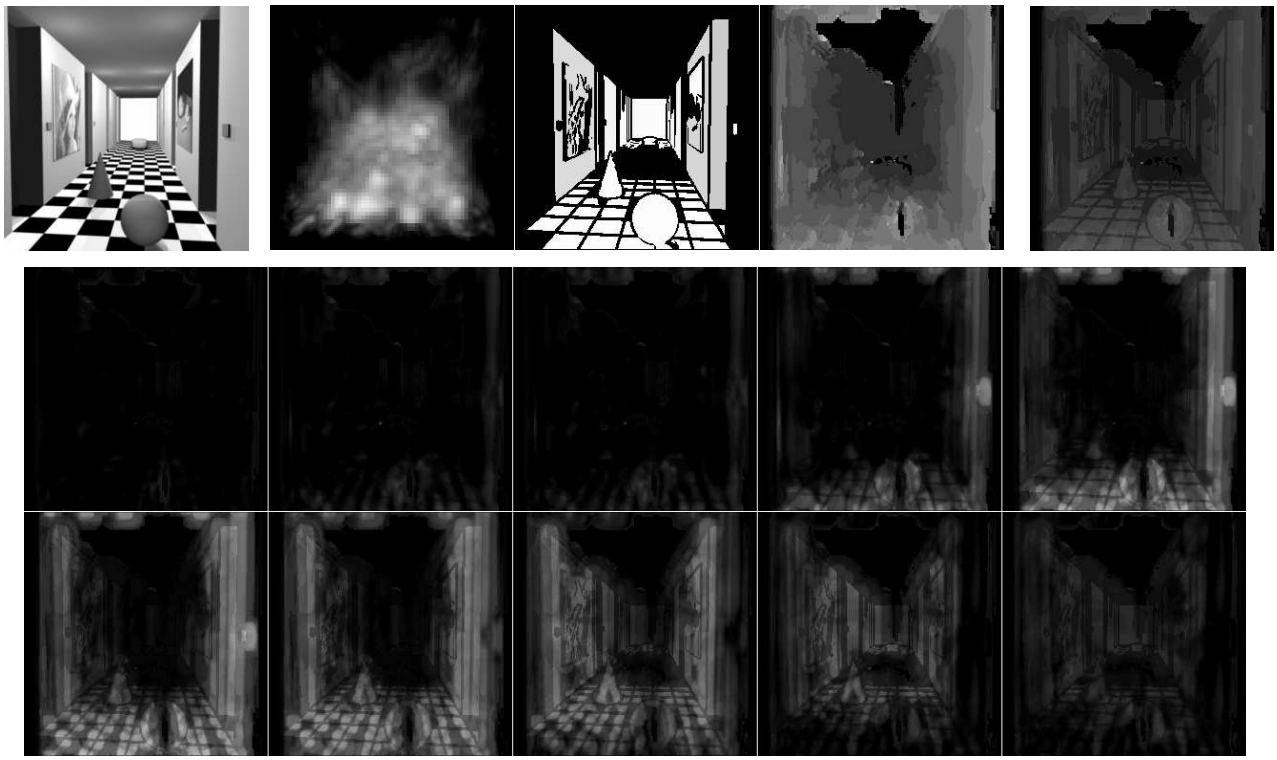


Abbildung 5.35: Die Integration der Merkmale (hier Mitte oben: Symmetrie, Exzentrizität und Tiefe) für ein Beispielbild (links oben) in eine übliche 2D-Mastermap (rechts oben) und in eine 3D-Repräsentation (in Leserichtung von links oben nach rechts unten nimmt die Disparität ab, die Entfernung also zu)

Vorbild und ihre Komplementarität untereinander aus. Trotzdem sind für konkrete Anwendungen soweit möglich zusätzliche spezialisierte Merkmale, die auf die Anwendungsdomäne hin optimiert wurden, einzubeziehen.

Neben der klassischen Repräsentation der Auffälligkeit in einer zweidimensionalen *master map of attention* wurden alternative Repräsentationen vorgestellt, die den nachfolgenden Selektionsverfahren umfangreichere Informationen zur Verfügung stellen, speziell eine dreidimensionale Auffälligkeitskarte. Die Eignung dieser Repräsentation für unterschiedliche Verfahren zur Selektion werden im folgenden Kapitel diskutiert.

Kapitel 6

Erste Selektionsstufe: Die Auswahl mehrerer visueller Objekte

Die im vorigen Kapitel beschriebenen Repräsentationen werden in anderen Modellen direkt verwendet, um anhand des Maximums einen einzelnen Fokus der Aufmerksamkeit auszuwählen. Das hier diskutierte Modell führt jedoch zuvor eine Selektion einiger Objekte in einer ersten Selektionsstufe aus, die in diesem Kapitel vorgestellt wird. Aufbauend auf eine Motivation dieser ersten Selektionsstufe werden die erwünschten Eigenschaften bestimmt, aus denen sich die Verwendung Dynamischer Neuronaler Felder für diese Selektionsstufe ableitet. Die Neuronalen Felder werden vorgestellt und hinsichtlich ihrer relevanten Eigenschaften untersucht. Mehrere Architekturen des Einsatzes dieser Felder werden entwickelt und analysiert, die jeweils in Zusammenhang mit den zuvor vorgestellten alternativen Salienzrepräsentationen stehen.

6.1 Ziel

In anderen Modellen zur visuellen Aufmerksamkeit (s. Kap. 3 und 4) wird eine *all-to-one* Selektion betrieben, also die Auswahl eines einzelnen Bereiches oder Objektes aus dem vollständigen Bild. Die Operationen gehören entweder in die präattentive parallele Stufe oder aber in die attentive serielle Stufe. Davon wird das vorgestellte Modell abweichen und zuerst eine *all-to-some* Selektion einführen, also in einem ersten Schritt eine kleine Anzahl von Elementen auswählen. Die darauf folgende Selektion eines einzelnen Fokus der Aufmerksamkeit aus diesen Elementen ist Thema des nächsten Kapitels. Im folgenden werden zuerst die Gründe für diese Entscheidung diskutiert, um anschließend die Anforderungen an eine solche Stufe abzuleiten.

Die Auswahl einiger Elemente leitet sich vor allem aus den Anforderungen ab, die die Verarbeitung dynamischer Szenen mit bewegten Objekten an die Aufmerksamkeitssteuerung stellt. An erster Stelle steht die Bindung von Informationen, die durch die serielle attentive Verarbeitung erhalten wurden, an die Elemente. Diese kann im dynamischen Fall nicht durch die Bindung an eine konstante Position erzielt werden. Stattdessen muss die Position des Elementes, auf das sich die Informationen beziehen, ständig aktualisiert werden. Dies entspricht einem Tracking des Elementes. Um also die attentiv

extrahierten Informationen aktuell zu halten, ist die Verfolgung der zuletzt selektierten Objekte notwendig. Damit wird es möglich, Informationen über die wichtigen Objekte für ein Objekt nach dem anderen zu extrahieren und so kontinuierlich mehr über die Umgebung zu lernen, obwohl sich diese Umgebung dynamisch verändert.

Der zweite entscheidende Unterschied in der Verarbeitung dynamischer gegenüber statischer Szenen liegt in der Hemmung bereits mit Aufmerksamkeit versehener Objekte oder Bereiche, um einen Wechsel der Aufmerksamkeitszuweisung zu ermöglichen. Die sonst übliche Verwendung einer statischen Inhibitionskarte reicht im dynamischen Fall nicht aus, sobald sich eines der inhibierten Objekte aus dem markierten Bereich herausbewegt. Für das Vorbild der natürlichen Aufmerksamkeit haben Tipper et al. [TDW91] eine solche Bindung der Hemmung an bewegte Objekte nachgewiesen, die sie in aktuellen Arbeiten [TW98b] mit einer zusätzlichen raumbasierten Hemmung verbinden.

Die Aufrechterhaltung der Bindung verlangt eine Verfolgung der zuvor fokal selektierten Objekte. Die primäre Aufgabe der ersten Selektionsstufe ist aber die Auswahl von auffälligen Elementen als Kandidaten für eine spätere fokale Selektion. Auch diese müssen verfolgt werden, um die Auswahl nicht nur von der aktuellen Salienzrepräsentation, sondern auch von der Historie dieser Daten abhängig machen zu können. Dazu sollte Evidenz für die Salienz über mehrere Frames hinweg integriert werden. Die zeitliche und räumliche Integration dieser Evidenz ist vor allem deswegen wichtig, weil die Berechnung der datengetriebenen Salienz auf Merkmale setzt, die ständig parallel für das ganze Bild bestimmt werden müssen und daher auf Effizienz optimiert sind. Man wird also vor allem in Umgebungen, deren Charakteristik von vornherein nicht zu bestimmen ist, mit Rauschen und Fehlern rechnen müssen. Um diese wenigstens zum Teil auszugleichen, kann eine zeitliche und räumliche Integration der Salienz einen wesentlichen Beitrag leisten. Die zu erfüllende Aufgabe dieser zweiten Verarbeitungsstufe geht somit über die reine Verfolgung von Elementen hinaus.

Die Verfolgung stellt jedoch eine wichtige Aufgabe dar. In diesem Fall muss das Tracking modellfrei ablaufen, da es vor der Erkennung der Objekte abläuft, die ja eine fokale Aufmerksamkeitszuweisung voraussetzt. Stattdessen spielt die Salienz hier eine entscheidende Rolle, denn um für das Tracking relevant zu werden, muss ein Objekt eine signifikant erhöhte Salienz im Vergleich zu seiner Umgebung aufweisen. Um wiederum selektiert zu werden, wird die erhöhte Salienz vorausgesetzt, die sich mit dem Objekt bewegen kann. Insofern sind Selektion und Verfolgung Prozesse, die miteinander interagieren und auf derselben Datenbasis operieren. Optimal wäre es somit, einen einzigen Mechanismus zu verwenden, der beide Aspekte in sich vereint: ein modellfreies Tracking der salientesten Objekte und eine robuste Selektion dieser salientesten Objekte. Vor der Entscheidung für einen solchen Mechanismus sollen noch die Anforderungen an die Selektion genauer geklärt werden.

Auf die Notwendigkeit zur räumlichen und zeitlichen Integration wurde bereits verwiesen. Die Selektion soll also keineswegs das Maximum der momentanen *master map of attention* auswählen, sondern die berechnete Auffälligkeit der letzten verarbeiteten Frames miteinbeziehen. Ebenso geht es nicht um einen einzelnen Punkt, dessen Salienz bewertet wird, das Maß der Auffälligkeit soll räumlich integriert werden. Schließlich ist Hysterese eine gewünschte Eigenschaft, um die Selektion auch bei leichten Schwankungen der Salienz stabil zu erhalten. Ein einmal selektierter Bereich soll also selbst dann selektiert bleiben, wenn seine Auffälligkeit für kurze Zeit leicht unter die Salienz eines nicht selektierten Objektes fällt.

Da hier mehrfach von der Selektion von Objekten die Rede war, soll geklärt werden, welcher Ob-

jektbegriff diesen Gedanken zugrunde liegt. Es kann sich dabei nicht alleine um Objekte handeln, die dem System bekannt sind. Genausowenig handelt es sich um gründliche, vom Hintergrund getrennte Segmente oder Gruppierungen, denn solche Prozesse würden in diesem System die Zuweisung von Aufmerksamkeit voraussetzen, die aber eben erst nach der Selektion erfolgt.

Diese Problematik der Reihenfolge von Gruppierung und Aufmerksamkeit ist auch für die natürliche Wahrnehmung bekannt. Es ließ sich dort sowohl zeigen, dass Gruppierung ohne Aufmerksamkeit stattfindet [ME97], als auch, dass Aufmerksamkeit auf bereits gruppierten Strukturen operiert. Dies führte Trick und Enns [TE97] zu der Aufteilung der Gruppierung in einen präattentiven Prozess des Clustering, der eine einfache Sammlung von Teilen darstellt und einer attentiven Formierung, in der unter anderem die Form der entstehenden Gruppe ausgewertet wird. Was zu der präattentiven Objekthaftigkeit von räumlichen Bereichen beiträgt, ist hingegen eine Konstanz hinsichtlich der berechneten Merkmale ebenso wie ein räumlich-zeitlicher Zusammenhang. Bereiche, die sich also homogen gegenüber den Merkmalen zeigen, einen räumlichen Zusammenhang aufweisen und Raum und Merkmale über die Zeit nur langsam ändern, werden als Objekte angesehen. Diese Eigenschaften treffen natürlich ebenso auf Objekte wie auf Teile von Objekten oder aber Gruppen von Objekten mit Ähnlichkeit und „gemeinsamem Schicksal" zu, deren Gruppierung oder Aufteilung in einzelne Objekte aber attentiven Prozessen vorbehalten sein soll.

Aus der Psychophysik entspricht Pylyshyn's in Kap. 3 vorgestelltes FINST-Modell der visuellen Indizes [PBF+94, Pyl98] am ehesten den genannten Anforderungen an diese Selektionsstufe. Leider gibt es kein Computermodell der FINST-Theorie. Dass jedoch auch beim Menschen ein vergleichbarer Zusammenhang zwischen Aufmerksamkeit und Tracking besteht und sich dieses Tracking von der sonstigen Bewegungswahrnehmung unterscheidet, konnten Culham et al. [CVAC00] anhand eines Nacheffekt zeigen, der nur für attentives Tracking auftritt.

Zu beachten ist, dass an dieser Stelle ein definierter Übergang von subsymbolischer zu symbolischer Verarbeitung stattfindet, da aus der signalnahen Salienzrepräsentation einige diskrete Elemente ausgewählt werden.

6.2 Dynamische Neuronale Felder

6.2.1 Dynamische Neuronale Felder nach Amari

Dynamische Neuronale Felder (kurz Neuronale Felder, DNF oder NF) wurden von Amari [Ama77] bereits 1977 als Modell von großen Verbünden kortikaler Neuronen vorgestellt und von Takeuchi und Amari [TA79] weiter analysiert. Die zugrunde liegenden Untersuchungen führen eine Mittelung und zeitliche Integration der Feuerraten solcher Neuronen durch. Das Entscheidende am Modell von Amari ist seine strukturelle Einfachheit und die daraus resultierende Eignung für mathematische Analysen. Man geht von homogenen Verbindungen in einem einschichtigen Netzwerk aus, also Verbindungen, die alleine von der räumlichen Distanz der beteiligten Neuronen abhängen. Zu den relevanten Eigenschaften Neuronaler Felder, die im Folgenden erläutert werden, gehören Hysterese und Bifurkation sowie eine räumliche und zeitliche Integration, die es für die im Rahmen dieser Arbeit relevante Selektionsaufgabe interessant erscheinen lassen. Sie werden typischerweise in Kontexten verwendet, in denen es um die Selektion in stark verrauschten Daten geht.

Formal wird die Dynamik Neuronaler Felder anhand ihrer lokalen Aktivation u am Ort $\mathbf{x}$ zum Zeitpunkt t definiert durch:

$$\tau \frac{d}{dt} u(\mathbf{x}, t) = -u(\mathbf{x}, t) + h + \int w(\mathbf{x} - \mathbf{x}') S[u(\mathbf{x}', t)] d\mathbf{x}' + i(\mathbf{x}, t) \tag{6.1}$$

Die Veränderung der Aktivation ist abhängig von der aktuellen Aktivation, einem (negativen) Ruhewert h, der durch eine sigmoide Funktion S und die Gewichte $w(\mathbf{x} - \mathbf{x}')$ zwischen den Neuronen vermittelte Aktivation der anderen Neuronen, einer externen Eingabe i und einer Zeitkonstanten τ. In der Interpretation als realem Neuronennetzwerk spricht man von einem Membranpotential u, das über die nicht-lineare Schwellwertfunktion S als neuronale Aktivität in Form einer Puls-Emissionsrate weitergegeben wird. Diese gelangt über Verbindungen mit den Gewichten w als Eingabe neben der externen Eingabe i an die benachbarten Neurone. Speziell die Definition der Gewichte w beeinflusst das Verhalten eines Neuronalen Feldes.

In jedem Fall geht man von lokal exzitatorischen Verbindungen aus, die mit zunehmender Distanz inhibitorisch werden. Man unterscheidet zuerst solche Gewichtsfunktionen, die außerhalb einer lokalen Umgebung 0 werden, von jenen, die für zunehmende Distanzen negativ bleiben. Im ersten Fall spricht man von einer lokalen Feldinhibition, während andere Gewichtsfunktionen zu einer sogenannten globalen Feldinhibition führen. Typische Definitionen für die Gewichte sind im Falle einer lokalen Feldinhibition DoG-Funktionen (*Difference of Gaussians*) mit lokal positiven Werten, während für globale Feldinhibition meist eine Normalverteilung abzüglich eines konstanten Wertes Verwendung findet (s. Abb. 6.1). Die Schwellwertfunktion S wird entweder als harter Schwellwert oder als sigmoide Funktion $S(x) = \frac{1}{1+\exp(-\beta * x)}$ umgesetzt. In jedem Fall ist S nicht-linear, monoton steigend und es gilt $\lim_{x \to -\infty} S(x) = 0$ und $\lim_{x \to \infty} S(x) = 1$.

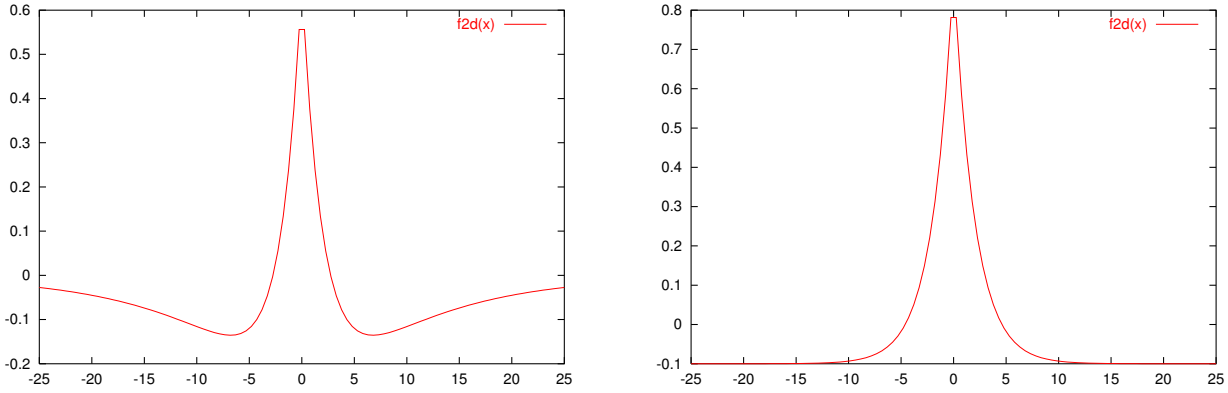


Abbildung 6.1: Gewichtsfunktionen für Neuronale Felder lokaler Feldinhibition (links) und globaler Feldinhibition (rechts).

Neuronale Felder können abhängig von der Wahl der Parameter stabile Zustände erreichen, die von der Art der Eingabe abhängen oder ein instabiles Muster von Aktivationen aufweisen. Man unterscheidet nach [KA79] folgende Fälle für prinzipielle Lösungen der Dynamikgleichung (Gleichung 6.1) bei Neuronalen Feldern mit lokaler Feldinhibition:

- die homogene leere Lösung (ϕ-Lösung) mit:
  $u(\mathbf{x}) \leq 0$ für alle $\mathbf{x}$ (kein aktiver Bereich)

- die homogene vollständige Lösung (∞-Lösung) mit:
  $u(\mathbf{x}) > 0$ für alle $\mathbf{x}$ (vollständige Aktivation)
- die instabile lokalisierte Lösung (a_1-Lösung) mit:
  $\bigwedge_t \bigvee_{x,t_1,t_2} : u(x,t_1) < 0 \wedge u(x,t_2) < 0,\ t_1, t_2 > t$
- die stabile lokalisierte Lösung (a_2-Lösung) mit:
  $u(\mathbf{x}) > 0$ für $\mathbf{x}_1 < \mathbf{x} < \mathbf{x}_2$

Die Beweise von Amari beziehen sich auf eine Stufenfunktion, Veit [Vei97] zeigt, dass sich für eine kontinuierliche Schwellwertfunktion dasselbe Verhalten ergibt. Von besonderem Interesse sind natürlich die a_2-Lösungen. Zur besseren Veranschaulichung der Eigenschaften sollen für ein eindimensionales Neuronales Feld Untersuchungen der Dynamik mit verschiedenen Vereinfachungen der Aktualisierungsregel durchgeführt werden. Ohne externe Eingabe endet ein DNF unabhängig von der initialen Aktivation bei einer konstanten Aktivation, die dem Ruhewert entspricht. Bei zeitlich konstanter Eingabe und einem mit dem Ruhewert initialisierten Feld werden die folgenden stabilen Zustände erreicht:

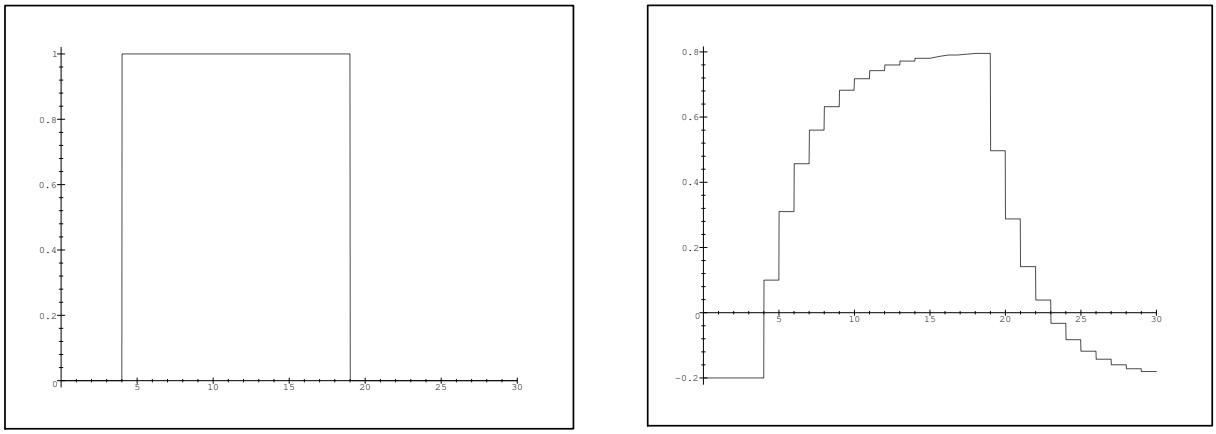


Abbildung 6.2: Aktivation eines Neurons (rechts) ohne Verbindung bei Rechteckimpuls als Eingabe (links).

Ohne Verbindungen: Mit $w(\mathbf{x}) \equiv 0$ verhalten sich alle Neuronen voneinander unabhängig und werden von ihrer Eingabe nur in einer Weise beeinflusst, die sich als *leaky integrator* charakterisieren lässt, sofern die Eingabe den Ruhewert überschreitet - andernfalls bleibt es bei der homogenen leeren Lösung. Für eine Eingabe in der Form einer (zeitlichen) Stufenfunktion ist in Abb. 6.2 die Aktivation des Neurons dargestellt.

Lineare Aktivationsfunktion: Verzichtet man auf die Nichtlinearität in der Schwellwertfunktion S fällt unter anderem die Eigenschaft der Bifurkation für das DNF weg.

Ohne Ruhewert: Der Ruhewert sorgt für einen Teil der Rauschunterdrückung, indem Eingaben, die unterhalb dieses Wertes liegen, ohne zusätzliche Anregung aus der Nachbarschaft ignoriert

werden, so dass also das Überschreiten des Ruhewertes eine notwendige Voraussetzung für die Ausbildung von Aktivitätsclustern ist, auch wenn sie nicht im gesamten Bereich des Aktivitätsclusters zutreffen muss.

Räumlich konstante Eingabe: Für eine Eingabe der Form $i(\mathbf{x}) \equiv s$ ergibt sich abhängig von s entweder die homogene leere Lösung oder aber die homogene aktivierte Lösung. Letzteres tritt genau dann ein, wenn $s + h > 0$ gilt.

Im Rahmen von Selektionsaufgaben ist man besonders an den Bereichen positiver Aktivation interessiert, im folgenden auch Aktivationscluster oder Bereiche sigmoider Aktivation genannt, definiert als:

$$R(u) = x|u(x) > 0 \tag{6.2}$$

Zur Charakterisierung der Gewichtsfunktion bedient sich Amari der Stammfunktion

$$W(x) = \int_0^x w(y)dy \tag{6.3}$$

und leitet daraus folgende charakteristische Maße her

$$W_m = \max_{x>0} W(x) \tag{6.4}$$

$$W_\infty = \lim_{x\to\infty} W(x) \tag{6.5}$$

, auf die im folgenden Bezug genommen wird.

Berechnung

Während es auf den ersten Blick so aussieht, als ob die Berechnung der Dynamik Neuronaler Felder für Felder globaler Inhibition die Faltung der Gewichtsfunktion, die für die gesamte Größe des Neuronalen Feldes definiert ist, mit der sigmoiden Aktivation des Feldes voraussetzt, wird stattdessen nur die Faltung eines kleinen Zentrums der Gewichtsfunktion vorgenommen. Außerhalb dieses Zentrums wird die Funktion als konstant angenommen. Die verbleibende Verknüpfung wird über die Verwendung eines globalen Inhibitionsneurons simuliert. Dieses Neuron summiert die Aktivation von allen Neuronen des Feldes und weist wiederum eine schwache inhibitive Verbindung zu allen Neuronen auf. Dies führt zu einer drastischen Beschleunigung der Berechnung, in deren Aufwand die Größe des Gewichtskernels eingeht. Die Beschleunigung führt sogar dazu, dass bei vergleichbaren Gewichtsfunktionen (d.h. wenn die anregende Normalverteilung eine vergleichbare Breite aufweist) die Berechnung für global inhibitive Felder schneller stattfinden kann als diejenige der lokal inhibitive Variante, da dort die Breite der zweiten Normalverteilung für einen größeren Gewichtskernel sorgt.

6.2.2 Allgemeine Anwendungen Neuronaler Felder

Zur Modellierung von Hirnaktivitäten auf den unterschiedlichen Ebenen Einzelneuron, Zellverbund bis hin zu großen Hirnstrukturen verwenden Jirsa et al. [JJFK01] eine modifizierte Version Neuronaler Felder, die aus Verbünden mehrerer Felder bestehen, welche untereinander nicht-symmetrische

Verbindungen aufweisen. Die Abbildung dieser Verbünde auf den Kortex erlaubt die Simulation von MEG- und EEG-Daten, die mit tatsächlichen Messungen in Korrespondenz gesetzt werden können.

Giese [GSH96, Gie99] verwendet die Dynamik der Neuronalen Felder zur Modellierung der Bewegungswahrnehmung, bei der durchaus mehrere Perzepte gleichzeitig vorhanden sein können. Die Stärke des Vorgehens liegt in der Kopplung perzeptueller Organisation mit den dynamischen Aspekten der Wahrnehmung.

Im Kontext mobiler Roboter stellen Schöner et al. [BS96, SDE96] Verwendung von sogenannten *behavioral variables* als Konsequenz aus dem Disput zwischen klassischer, hierarchischer Planung und verhaltensbasierter Robotik vor. Die *behavioral* variables beruhen in ihrer Implementation auf der Verwendung Neuronaler Felder, die subsymbolisch Sensorinformationen verarbeiten und symbolische Ergebnisse für die Handlungssteuerung bereitstellen.

Bruckhoff und Dahm [BD98, DBJ98] verwenden Neuronale Felder zur lokalen Pfadplanung und -steuerung eines mobilen Roboters. Dabei kodiert ein eindimensionales, zyklisches NF die möglichen Richtungen, in die sich der Roboter bewegen könnte. Als Eingabe wird eine Repräsentation des zu erreichenden Zieles mit aktuellen Sensordaten zur Detektion von Hindernissen und dem Kurzzeitgedächtnis entnommenen Informationen zu Hindernissen, die außerhalb der momentanen Sensoren liegen, kombiniert. Die Position des Maximums innerhalb des Neuronalen Feldes determiniert die Rotation, die Stärke dieses Maximums die Geschwindigkeit der Vorwärtsbewegung. Es konnte gezeigt werden, dass das Neuronale Feld auch bei multimodalen Verteilungen und deutlichem Sensorrauschen eine zuverlässige Selektion zur Planung und Steuerung aufwies.

Engels und Schöner [ES95] gehen ein ähnliches Problem an wie Bruckhoff und Dahm, verwenden jedoch eine reduzierte Variante des Neuronalen Feldes, die zwar schneller zu berechnen ist, der jedoch wichtige Eigenschaften wie das Clustering fehlen.

Als Modellierung einer Form von Arbeitsgedächtnis verwenden Laing et al. [LTGE02] Neuronale Felder, wobei sie ausführliche Untersuchungen zu den Bedingungen durchführen, unter denen sich mehrere Aktivitätscluster bilden können. Sie verwenden dazu jedoch eine andere Variante Neuronaler Felder, bei denen die Verbindungen nicht monoton fallen, sondern oszillieren.

Das Problem der geeigneten Parametrisierung der Neuronalen Felder wurde bei Igel et al. [IEJ01] durch die Anwendung von Genetischen Algorithmen gelöst.

Eine sehr verwandte Gruppe Neuronaler Modelle wird als *Dynamic Link Matching* bezeichnet. Sie wurden von Konen et al. [KMM94] zur gleichzeitigen Lokalisierung und Erkennung von Objekten verwendet.

6.2.3 Verwendung Neuronaler Felder zur Steuerung von Aufmerksamkeit

Das erste Modell zur Steuerung von Aufmerksamkeit, das auf der Dynamik Neuronaler Felder basiert, haben Kopecz et al. [Kop96] 1996 vorgestellt. Es verwendet ein DNF globaler Feldinhibition, so dass ein eindeutiges Ergebnis des WTA sichergestellt ist. Das System führt nun anhand der Position des Aktivitätsclusters ein einfaches Tracking des auffälligsten Objektes durch, bis ein anderes Objekt weit auffälliger wird oder das verfolgte Objekt den sichtbaren Bereich verlässt. Es handelt sich um die Modellierung offener Aufmerksamkeit mit einem einzigen einfachen Verhalten.

Eine Erweiterung des Systems [PKE98, PKE99] zeichnet sich durch die Ausführung langsamer Verfolgungsbewegungen zusätzlich zu sakkadenartigen Sprüngen aus. Die Unterscheidung ist aller-

dings eine quantitative, da Sakkaden und langsame Folgebewegungen im Gegensatz zum natürlichen Vorbild als Bewegungen desselben Typs, aber unterschiedlicher Weite angesehen werden. Es fehlt für das Modell eine ausführlichere Diskussion des Loslösens der Selektion, um eine nächste Selektion zu ermöglichen.

Hamker und Gross [HG97] stellen ein zweistufiges Selektionsmodell vor, das auf der Verwendung Neuronaler Felder beruht. Im Unterschied zum hier vorgestellten Modell geht es jedoch wie üblich um die Selektion eines einzigen Objektes, das anhand von Merkmalskarten in einem WTA-Prozess durch ein Neuronales Feld bestimmt wird. Die zweite Stufe dient der Segmentierung eines Objektes, beginnend mit dem Aktivationsbereich des Neuronalen Feldes. Damit wird eine betont objektbasierte Selektion umgesetzt. Es wird von einer alternativen Beschreibung der Neuronalen Felder nach Kaski und Kohonen [KK94] ausgegangen, deren Berechnung jedoch einen höheren Aufwand erfordert, wie Wilhelm [Wil98] gezeigt hat.

Eine interessante Modifikation der Struktur Neuronaler Felder stellt Ahrns [AN99, Ahr00] vor, indem die Struktur der Neuronalen Felder ortsvariant modelliert wird, was die Berechnung der Dynamik erheblich beschleunigt, ohne dabei die wesentlichen Eigenschaften der Felder zu verlieren. Die Modellierung von Auffälligkeit beruht jedoch nur auf einem einzelnen Merkmal, entspricht weitgehend Standardmodellen und kennt keine weiteren Verhalten.

In der Gruppe von Böhme und Gross wurde ein System zur visuellen Lokalisation von Personen vorgestellt, dessen Aufmerksamkeitskomponente eine dreidimensionale Struktur Neuronaler Felder enthält [CBB+98, CBBG98, Bra01], wobei jedoch für die dritte Dimension eine Auflösungspyramide für die Merkmale herangezogen wurde, bei denen die Größe als konstant und damit umgekehrt proportional zur Entfernung angenommen wurde. Es handelt sich um ein Neuronales Feld globaler Feldinhibition, die als Eingabe eine Auflösungspyramide von aggregierten Hinweisen auf die Präsenz von Köpfen (Hautfarbe, Form der Kopf/Schulterpartie) erhält. Diese Pyramidenrepräsentation wird jedoch nicht für das Neuronale Feld verwendet, es erfolgt stattdessen ein Verkleinern der übrigen Pyramidenschichten auf die geringste Auflösungsstufe. Das Modell wird jedoch rein statisch zur Selektion anhand eines einzelnen Bildes verwendet, dynamische Aspekte und Wechsel der Aufmerksamkeit wurden nicht berücksichtigt.

6.2.4 Selektion durch Neuronale Felder

Als wichtige Selektionseigenschaften der Neuronalen Felder gelten die Rauschunterdrückung, die räumliche und temporale Integration und die Hysterese. Zur Untersuchung dieser Eigenschaften in Neuronalen Feldern werden W_m, W_∞ und das Ruhepotential h als wesentliche charakterisierende Eigenschaft der Gewichtsfunktion definiert (s. Abb. 6.3). Die folgenden Experimente wurden immer mit zweidimensionalen Neuronalen Feldern durchgeführt, sofern jedoch nur eine Dimension zur Betrachtung relevant war, wurde für die Darstellung eine Projektion auf diese Dimension vorgenommen.

Unter Hysterese versteht man die Abhängigkeit eines Zustandswechsels vom aktuellen Zustand in der Art, dass eine Zustandsänderung nur mit einem höheren Aufwand zu erreichen ist als ein Beibehalten des aktuellen Zustandes. Zur Veranschaulichung der Hystereseeigenschaft wird eine Eingabe mit zwei lokalisierten Peaks definiert, deren Amplitude durch einen Parameter α bestimmt werde. Während der eine Peak an Position 16 jeweils die Amplitude αerhält, wird die Amplitude des zweiten Peaks an Position 48 durch $1 - \alpha$ gegeben, so dass die Summe jeweils konstant bleibt. Verändert

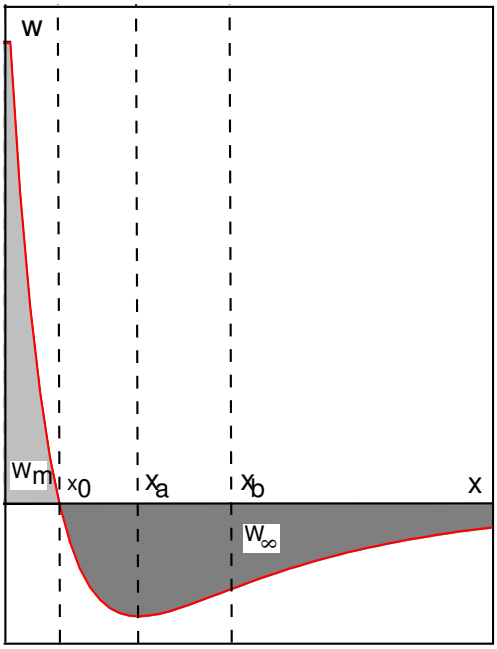


Abbildung 6.3: Zentrale Charakteristika der Gewichtsfunktion w in Abhängigkeit der Distanz x für lokal inhibitive Neuronale Felder (dargestellt für den eindimensionalen Fall).

man den Anteil der einzelnen Peaks kontinuierlich durch ein Erhöhen bzw. Vermindern von α, so ergibt sich die in Abb. 6.4 dargestellte Hystereseschleife für die Position des Aktivitätsmaximums im Neuronalen Feld. Es zeigt sich, dass die Position des Aktivitätsclusters keineswegs wechselt, sobald das Maximum sich verlagert, sondern erst nachdem die vorher schwächere Eingabe deutlich höhere Werte erreicht.

Die allmähliche Aufteilung eines einzelnen Maximums in zwei räumlich getrennte Maxima resultiert in der in Abb. 6.5 dargestellten Bifurkation. Die Entfernung, bei der die Trennung in zwei Aktivationsbereiche stattfindet, wird durch den Verbindungskernel determiniert. Er führt, wie von Amari [Ama77] gezeigt wurde, in Abhängigkeit der Distanz d zwischen zwei Maxima in der Eingabe zu folgenden Verhaltensweisen:

- lokale Maximumssuche für $0 < d < x_a$,
- Abstoßung der Maxima für $x_a < d < x_b$ und
- Koexistenz für $x_b < d$.

Die Unterdrückung von Rauschen im Selektionsprozess zeigt sich, wenn man einen Rechteckimpuls mit gleichverteiltem Rauschen unterschiedlicher Intensitäten überlagert und anschließend überprüft, ob der Rechteckimpuls ein Aktivationscluster verursacht und ob es weitere Aktivationscluster gibt. Die Ergebnisse solcher Experimente für unterschiedliche Amplituden zeigt Abb. 6.6. Bei einer Signalstärke von 1 findet man bis zu einer Stärke des Rauschens von 1.2 ausschließlich Aktivationen, die vom Signal erzeugt werden, erst darüber hinaus zeigt das Rauschen seinen Einfluss.

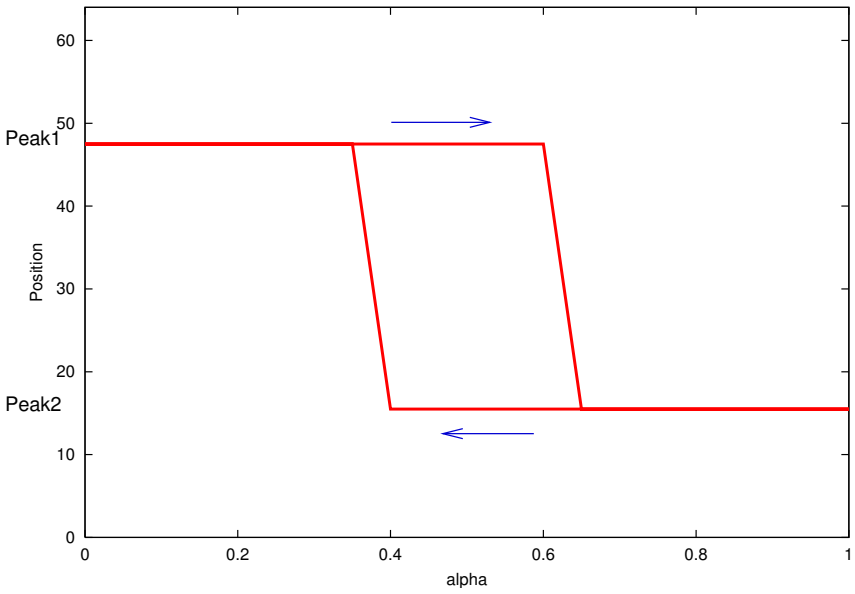


Abbildung 6.4: Hystereseschleife: Position des Aktivitätsclusters in Abhängigkeit von der Stärke der Eingabecluster (das untere Cluster mit der Amplitude α, das obere Cluster mit der Amplitude $1-\alpha$). Der Wert von α wurde zuerst von 0 bis 1 erhöht (oberer Verlauf) und dann von 1 wieder bis auf 0 vermindert (unterer Verlauf).

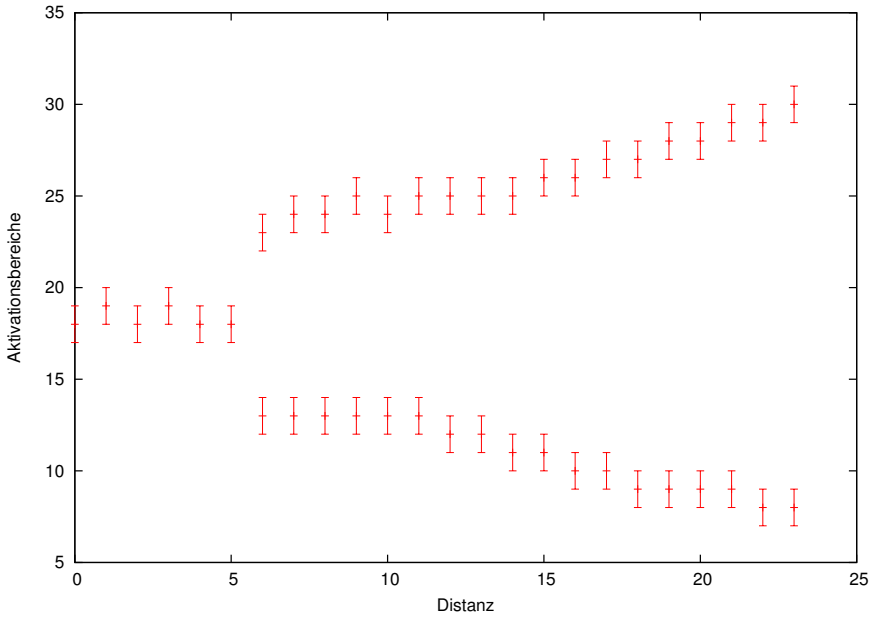


Abbildung 6.5: Bifurkation: durch Trennung zweier Maxima erhält man ab einer gewissen Distanz zwei Aktivationsbereiche.

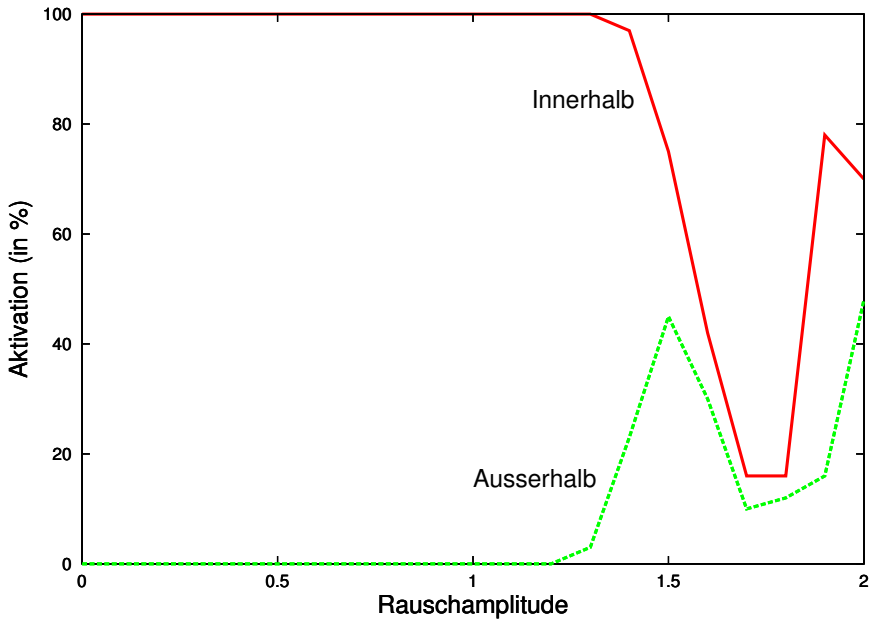


Abbildung 6.6: Überlagerung eines Rechteckimpulses mit normalverteiltem Rauschen: Häufigkeit des Auftretens aktivierter Neuronen innerhalb und außerhalb des Impulsbereiches.

6.3 Zweidimensionale Dynamische Neuronale Felder

Ausgehend von den in Kap. 5.5 vorgestellten zweidimensionalen Salienzrepräsentationen wird hier analysiert, in welcher Form Neuronale Felder die Selektionsstufe für derartige Repräsentationen darstellen können. Für jede Repräsentation wird dazu nach geeigneten Modellen Neuronaler Felder gesucht, die aus den jeweiligen Eigenschaften der Salienzrepräsentation Vorteil ziehen können. Verwandte Diskussionen sind in [BM02b] zu finden.

6.3.1 Verwendung eines einzelnen zweidimensionalen Neuronalen Feldes

Der naheliegende Typ von Neuronalen Feldern zur Selektion mehrerer Einheiten ist ein Feld lokaler Inhibition. Bei diesem Feld ist die Hemmung, die ein Aktivationscluster ausübt, räumlich begrenzt und wird durch eine DoG-Funktion definiert. Damit ist es möglich, dass mehrere Aktivitätscluster gleichzeitig auftreten. Als Eingabe für solch ein Feld ist eine einzelne Salienzkarte, wie sie durch die Gewichtung der Merkmalskarten (Kap. 5.5.2) entsteht, geeignet.

Abb. 6.7 zeigt die Struktur der Verwendung anhand einer einfachen Szene und den dazugehörigen Ergebnissen. Die Bereiche positiver Aktivation sind farblich hervorgehoben. Die Dynamik des Feldes entspricht Formel 6.1, wobei $i(x) = mm(x)$ der Aktivation der Mastermap entspricht (s. 5.20).

Die Veränderung der Aktivation, ausgehend von einem mit dem Ruhewert initialisierten Neuronalen Feld bis zu einem stabilen Zustand (als Kriterium für einen stabilen Zustand wird festgelegt, dass die gemittelte absolute Veränderung der Aktivation kleiner als 0.02 ist), ist in Abb. 6.8 illustriert. Beachtenswert ist die relativ kleine Anzahl von Zyklen für ein iteratives dynamisches System.

Die Verfolgungseigenschaften, die ein solches Feld aufweist, wurden mit einer künstlichen Eingabe untersucht. Dazu wurde ein Zielreiz in einer Umgebung aus normalverteiltem Rauschen bewegt. Untersucht wurde die Anzahl von Aktualisierungszyklen, die für ein korrektes Verfolgen des Reizes bei

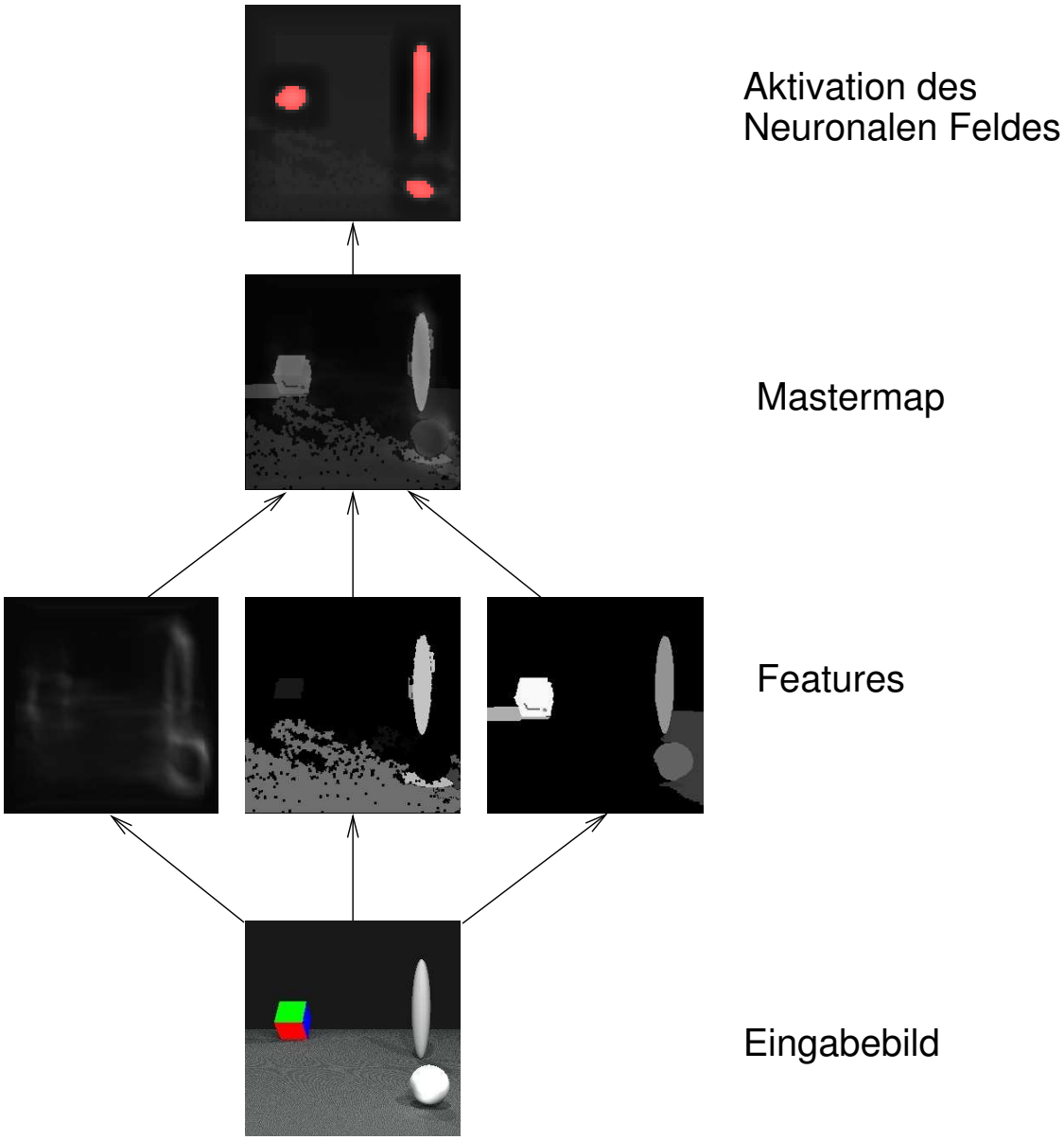


Abbildung 6.7: Verwendung des Neuronalen Feldes mit lokaler Feldinhibition am Beispiel. Die Bereiche positiver Aktivation (Aktivationscluster) sind farblich hervorgehoben.

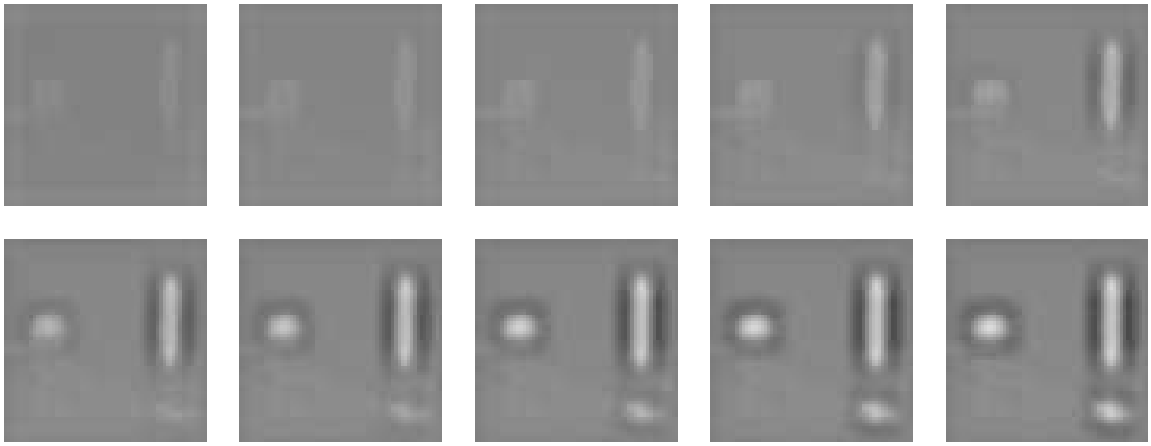


Abbildung 6.8: Entwicklung der Aktivation in einem Neuronalen Feld für 10 Zyklen. Ausgehend von dem Ruhewert des Feldes wird ein stabiler Zustand (durchschnittliche Aktivationsänderung je Neuron im letzten Zyklus bei unter 0.02) erreicht. Als Eingabe wurde die in Abb. 6.7 als Mastermap dargestellte Karte verwendet.

unterschiedlichen Geschwindigkeiten und Signal-Rausch-Verhältnissen notwendig war. Als Kriterium für die korrekte Verfolgung wurde eine minimale Überlappung von der Hälfte des Zielreizes durch das Aktivitätscluster des Neuronalen Feldes angenommen.

Abb. 6.9 zeigt das Ergebnis, wobei die Experimente nach jeweils 55 Aktualisierungszyklen abgebrochen wurden. Es zeigen sich also die Grenzen des Trackings bei einer Objektbewegung von mehr als 12 Pixeln und andererseits bei einer Objektamplitude von 0.5, sofern die Geschwindigkeit hoch genug ist. Bewegen sich die Werte jedoch innerhalb dieser Grenzen, reichen typischerweise schon 10 Aktualisierungszyklen des Neuronalen Feldes, um den Kontakt zum verfolgten Objekt nicht zu verlieren.

Diese Verfolgung eines Objektes demonstriert die prinzipielle Eignung des Feldes, der Einsatzzweck ist aber die gleichzeitige Selektion und Verfolgung mehrerer Elemente. Hier ist nach Amari [Ama77] zu unterscheiden, wie groß die Distanz der Objekte zueinander ist. Bei einer Entfernung größer als x_b beeinflussen sich die Objekte nicht gegenseitig. Sobald diese Grenze jedoch unterschritten wird, findet eine Interaktion statt.

Zur Untersuchung des Verhaltens wurde das Experiment zur Bifurkation repliziert mit dem Unterschied, dass sich die beiden Objekte einander diesmal annähern (Abb. 6.10). Es ist im Bereich zwischen x_a und x_b eine Abstoßung festzustellen, so dass die Positionen der Aktivitätscluster nicht mehr vollständig mit der tatsächlichen Position der Maxima übereinstimmen - die Aktivitätscluster werden weggedrängt. Sinkt die Distanz schließlich unter x_a, so findet eine Vereinigung der Aktivitätscluster statt. Diese würde in umgekehrter Richtung erst bei Überschreiten der Distanz x_a wieder aufgelöst. Auch hier zeigt sich also ein stabiles Verhalten, es gibt keine Entfernung, um die herum eine oszillierende Vereinigung und Trennung zweier Aktivitätscluster auftritt.

Um die Verfolgungsleistungen weiter bewerten zu können, muss man sich vergegenwärtigen, welche Art von Information vom System genutzt wird, um ein Objekt zu verfolgen. Bei dieser Verwendung der Neuronalen Felder beruhen Selektion und Verfolgung auf Bereichen hoher integrierter Salienz. Es stehen keine weiteren Informationen zur Verfügung, um ein Objekt von einem anderen zu unterschei-

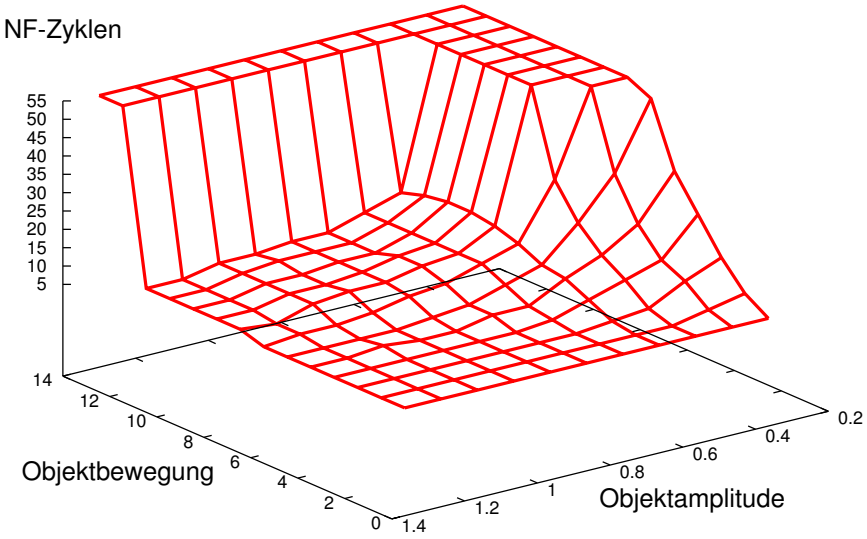


Abbildung 6.9: Anzahl notwendiger Aktualisierungszyklen des Neuronalen Feldes, um das Tracking eines Objektes angegebener Geschwindigkeit und Amplitude zu gewährleisten. Die Berechnung wurde nach 55 Zyklen abgebrochen.

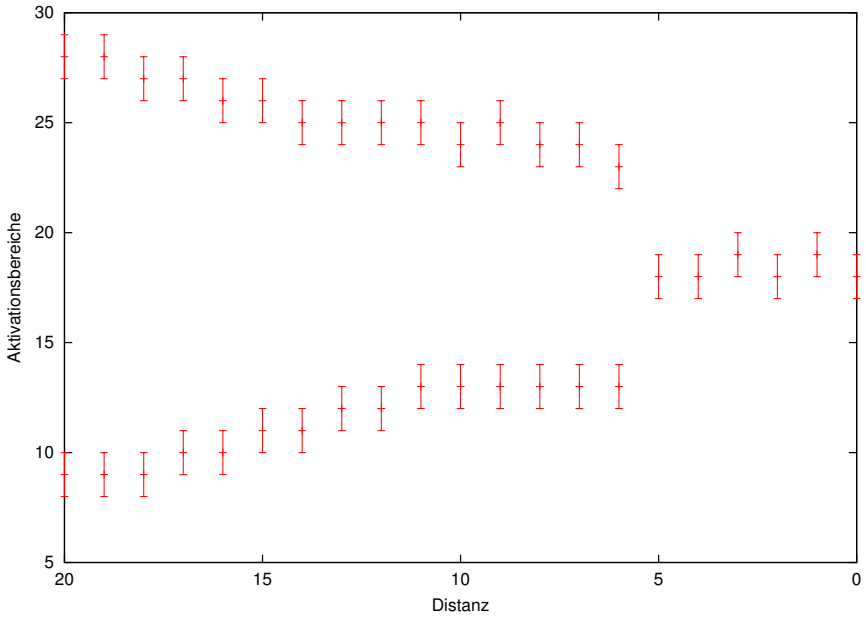


Abbildung 6.10: Verfolgung zweier Maxima: innerhalb des Interaktionsbereiches (ab Distanz 10) findet zuerst eine Abstoßung der Aktivationscluster und danach (ab Distanz 5) eine Vereinigung zu einem Cluster statt.

den. Dies kann vor allem dann problematisch werden, wenn saliente Objekte benachbart auftreten oder sich sogar temporär überlappen. Ein nachgeschalteter Prozess zur Bestimmung der Korrespondenz zwischen den verfolgten Objekten und den Aktivitätsclustern wäre notwendig. Während dies natürlich möglich wäre, widerspricht es jedoch der Integration von Selektion und Verfolgung. Deswegen wird die im folgenden Abschnitt dargelegte Verwendung mehrerer Neuronaler Felder bevorzugt, die vor allem von einer reicheren Objektrepräsentation profitiert.

6.3.2 Konnektivität zwischen mehreren Neuronalen Feldern

Um der Selektion und Verfolgung reichhaltigere Informationen zur Verfügung zu stellen als dies in der ersten vorgestellten Variante der Fall war, soll die individuell gewichtete Salienzrepräsentation (s. Kap. 5.5.5) als Eingabe in die Neuronalen Felder dienen. Entsprechend werden mehrere Neuronale Felder verwendet, um jedem der Felder ein individuelles Profil der Auffälligkeit darzubieten. Dies bedeutet, dass es für jede Kombination aus Feld und Merkmal ein Gewicht gibt und die Eingabe für jedes Feld eine (andere) Mastermap darstellt. Abb. 6.11 stellt schematisch die Verwendung dar. Die Eingabe in die Felder $i_n(x)$ wird nun beschrieben anhand einer Gewichtung $fw_t(m,n)$ für das m-te Merkmal und das n-te Neuronale Feld. Der Index t deutet schon an, dass diese Gewichte sich im Laufe der Zeit ändern werden. Ihre genaue Festlegung wird später diskutiert werden.

$$i_n(x,t) = \sum_m fw_t(m,n) * feat_m(x) \tag{6.6}$$

In diesem Fall fällt die Entscheidung eindeutig für Neuronale Felder globaler Feldinhibition. Der Grund liegt darin, dass die globale Feldinhibition das Vorhandensein eines einzigen Aktivitätsclusters im Feld garantiert. Davon profitiert einerseits die vereinfachte Bestimmung der Korrespondenz zwischen Objekt und Aktivitätscluster. Wichtiger noch ist jedoch, dass sich dadurch die Gewichte für ein Neuronales Feld auf ein einziges Objekt beziehen. Das erlaubt die im weiteren beschriebene Anpassung der Gewichte an die Eigenschaften des verfolgten Objektes zur Stabilisierung von Selektion und Tracking. Schließlich spricht auch die schnellere Berechnung der global inhibitiven Neuronalen Felder bei der Verwendung mehrerer Felder für diese Variante.

Um die gleichzeitige Verfolgung und Selektion desselben Objektes durch unterschiedliche NF zu vermeiden, ist eine inhibitive Verbindung zwischen den Feldern notwendig. Andernfalls würden ähnliche Gewichtsfunktionen zur mehrfachen Selektion desselben Objektes durch mehrere NF führen und so zu redundanten Berechnungen und der Verfolgung und Selektion nur weniger Objekte führen. Die notwendige Inhibition wird rein lokal vorgenommen und inhibiert von jedem Neuron aus die am selben Ort jedoch in den anderen Feldern befindlichen Neuronen. Damit ergibt sich die Dynamik mit dem Parameter c_{ib} für die lokale Inhibition vorläufig zu:

$$\tau \frac{d}{dt} u_j(x,t) = -u_j(x,t) + h + \int w(x-x') S[u_j(x',t)] + \sum_k c_{ib} S[u_k(x',t)] + i_j(x,t) \tag{6.7}$$

Bestimmung der Merkmalsgewichte

Festzulegen bleibt, auf welche Art und Weise die Gewichte bestimmt werden, um von der neuen Repräsentation tatsächlich profitieren zu können. Initial macht es Sinn, die Gewichte ein wenig voneinander

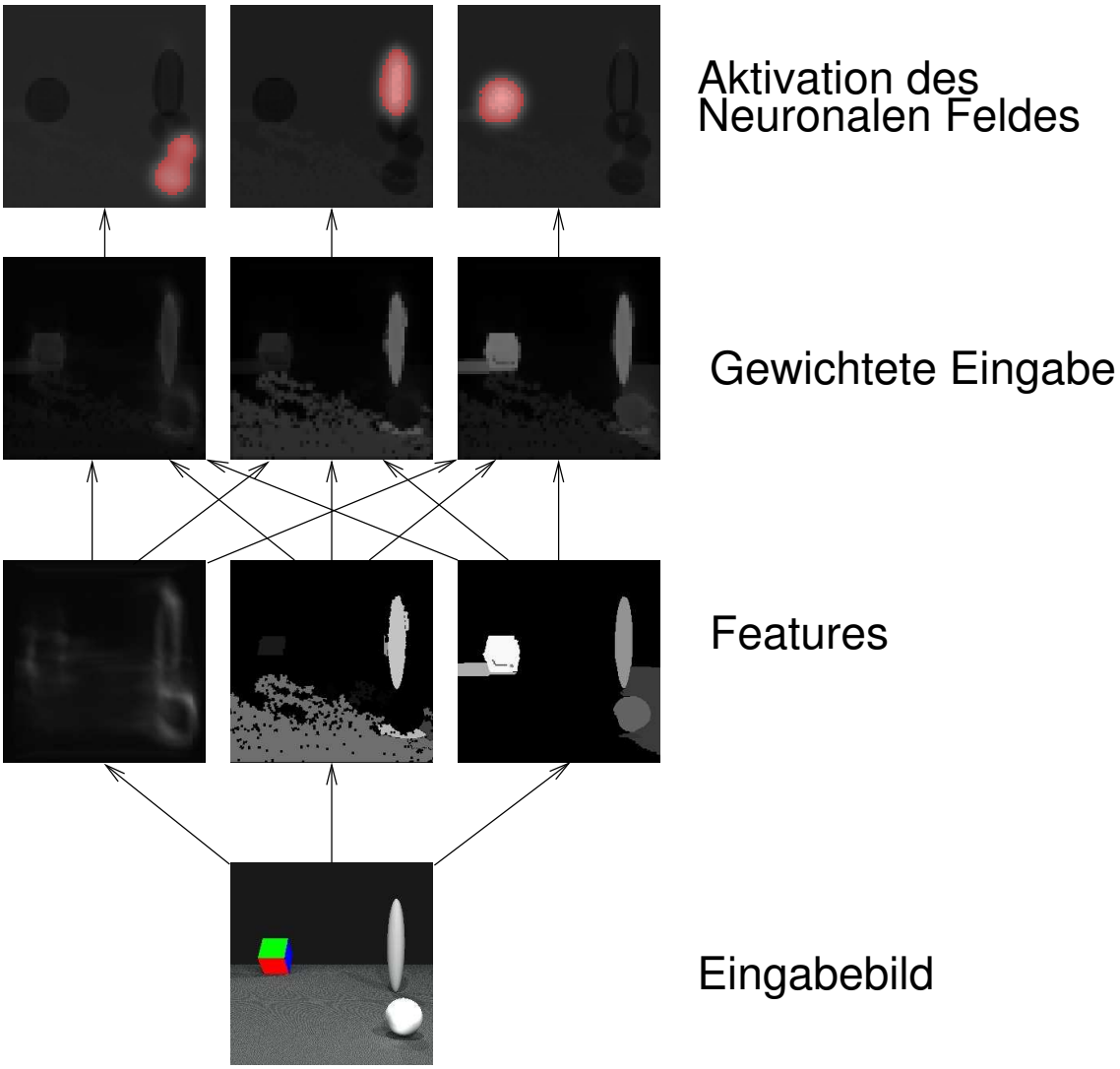


Abbildung 6.11: Verwendung eines Systems Neuronaler Felder mit individuell gewichteten Merkmalen. Die Bereiche sigmoider Aktivation sind farbig hervorgehoben.

zu unterscheiden, um Oszillationen bei identischen Eingaben in die Felder zu vermeiden. Dabei sollten die Merkmale jedoch, sofern kein übergeordnetes Wissen über die Wichtigkeit der Merkmale vorhanden ist, in der Summe gleich gewichtet werden. Sobald sich in einem der NF ein Aktivationscluster herausgebildet hat, ergibt sich die Möglichkeit, diese Selektion zu stabilisieren und damit auch das Tracking zu verbessern. Dazu wird die Präsenz der Merkmale im Bereich des Aktivitätsclusters im Verhältnis zu ihrer sonstigen Präsenz ausgewertet. Dies ergibt eine Art von Signal-Rausch-Verhältnis dafür, wie relevant jedes einzelne der Merkmale für das ausgewählte Element ist. An dieses Verhältnis werden die Gewichte in leaky-integrator-Form adaptiert:

$$fw'_t(m,n) = \lambda * fw_{t-1}(m,n) + (1-\lambda) * \frac{sn_t(m,n)}{n_{nf}} \tag{6.8}$$

Die Informativität der einzelnen Merkmale $sn_t(n)$ für ein Merkmal n und ein aktives Neuronales Feld m wird dabei anhand der Merkmalsinformation $feat_n(x,t)$ und der Fläche der Aktivation in diesem Feld $S_m(t)$ gegenüber der Gesamtfläche des Feldes NF_m bestimmt zu

$$sn_t(m,n) = \frac{\sum_{S_m(t)} feat_n(x,t)}{\sum_{NF_m} feat_n(x,t)} * \frac{NF_m}{S_m(t)} \tag{6.9}$$

Ein Haupteffekt der Gewichtung ist die Unterscheidung der Objekte. Diese Unterscheidung ist vor allem dann wichtig, wenn keine anderen Kriterien eingesetzt werden können. Als anderes Kriterium kommt primär die räumliche Position in Frage. Unterscheidet sich diese bereits stark für zwei Objekte, ist ihre unterschiedliche Gewichtung eher unerheblich. Sind sich die Objekte jedoch nahe, ist die Gewichtung zur Trennung beider Objekte entscheidend. Demnach wird eine Verknüpfung der Gewichte zwischen den Feldern in Abhängigkeit von der räumlichen Nähe vorgenommen. Diese Verknüpfung findet inhibitiv statt, so dass die Unterschiede zwischen den Profilen betont werden. Die Verknüpfung ist dabei jedoch durch die räumliche Nähe beeinflusst, so dass sie mit wachsender Entfernung abnimmt. Das Zentrum des Aktivitätsclusters im Feld m sei bezeichnet mit g_m:

$$fw''_t(m,n) = fw'_t(m,n) - \sum_{m'} c_{fwi}(g_m - g_{m'}) fw'_t(m',n) \tag{6.10}$$

Schließlich werden die Gewichte noch normalisiert:

$$fw_t(m,n) = \frac{fw''_t(m,n)}{\sum_{n'} fw''_t(m,n)} \tag{6.11}$$

Umgekehrt betrachtet ergibt die vorige Argumentation, dass die Bedeutung der lokalen Inhibition zur Vermeidung der mehrfachen Selektion desselben Elementes durch verschiedene Neuronale Felder von der Ähnlichkeit der Gewichtungen abhängt. Haben zwei Neuronale Felder sehr unterschiedliche Gewichte, laufen sie kaum Gefahr, tatsächlich dieselben Objekte zu selektieren. In diesem Fall lässt sich die lokale Inhibition durchaus vermindern. Davon profitiert das System im Falle naheliegender ähnlicher Objekte. Räumlich und zeitlich begrenzte Okklusionen selektierter Objekte müssen jetzt nämlich nicht zu vollständiger Unterdrückung der Aktivation in einem der Felder führen, sofern die Objekte anhand ihrer Profile ausreichend unterscheidbar sind.

Das Verfolgungsverhalten in dieser Verwendung Neuronaler Felder unterscheidet sich von dem innerhalb eines einfachen Feldes vor allem dadurch, dass durch die Anpassung der Gewichte eine

bessere Unterscheidung von verfolgtem Objekt und Hintergrund gewährleistet wird. Dies führt zu einer Erhöhung des Signal-Rausch-Verhältnisses, dessen Einfluss auf die Verfolgungsleistung Abb. 6.9 zu entnehmen ist.

Zusätzlich zu dieser quantitativen Steigerung, gibt es jedoch einen qualitativen Fortschritt bei der Behandlung von Okklusionen. Die dieser Verwendung zugrunde liegende Repräsentation von Salienz beruht auf individuell gewichteten Clustern hoher Salienz, d.h. auf Profilen von Merkmalen, die sich in Abhängigkeit von den anderen Gewichten und bereits getroffener Selektionen anpassen. Diese Repräsentation erlaubt eine Unterscheidung zwischen Objekten, die auch im Falle zeitlich begrenzter Verdeckung die weitergehende Verfolgung der betroffenen Objekte erlaubt. Ein Vergleich hinsichtlich der Verfolgung sich temporär verdeckender Objekte findet zusammen mit dem noch vorzustellenden 3D-Modell Neuronaler Felder in Kap. 6.4 statt.

Der Rechenaufwand für ein einzelnes NF wird dominiert durch die Bestimmung der Eingabe in jedes Neuron und wächst jeweils linear mit der Anzahl der Neuronen und der Fläche der Nachbarschaft für die Gewichtsfunktion. Dies gilt auch für den Fall der vorgeschlagenen Verknüpfung mehrerer Neuronaler Felder, da die Inhibition zwischen den Neuronalen Feldern rein lokal stattfindet. Somit hängt der Gesamtaufwand linear von der Anzahl der verwendeten Felder ab. Diese Zahl wird bestimmt anhand der Systemressourcen, der in typischen Szenen vorkommenden Anzahl von relevanten Objekten und der Aufgabe des Systems. Für die meisten Experimente in dieser Arbeit wurden vier Felder verwendet, da dies in etwa der entsprechenden Leistung des menschlichen visuellen Systems zur gleichzeitigen Selektion und Verfolgung bewegter Objekte [PS88, PBF+94, Pyl98] entspricht.

Experimente

Betrachten wir eine Situation, in der bereits mehrere Objekte unterschiedlicher Salienz durch die Neuronalen Felder selektiert wurden und untersuchen, was passiert, wenn nun ein zusätzliches Objekt auftaucht und alle bisher selektierten Objekte an Auffälligkeit übertrifft. Abb. 6.12 zeigt die Aktivation der Neuronalen Felder zusammen mit der Eingabe. Es ergibt sich, dass das neue Objekt sich gegen das schwächste der bisher selektierten Elemente durchsetzt, die anderen Aktivationscluster aber bestehen bleiben. Auch in dieser Hinsicht ist das Verhalten des Systems als robust und effizient zu bezeichnen, es wird nur das am wenigsten wichtige Objekt zugunsten eines neuen wichtigen Objektes betroffen, die sonstigen Zuordnungen von Aktivitätsclustern und Bereichen bleibt bestehen.

6.4 Dreidimensionales Dynamisches Neuronales Feld

Besonders zur Trennung von Objekten, die sich in der zweidimensionalen Abbildung in räumlicher Nähe befinden, ist die dreidimensionale Salienzrepräsentation, die in Kap. 5.5.6 vorgestellt wurde, geeignet. Um im Selektions- und Verfolgungsprozess davon profitieren zu können, müssen die verwendeten Neuronalen Felder ebenso in die Tiefe ausgedehnt werden. Dass auch der Mensch gerade bei diesen beiden Aufgaben von der Verwendung von Tiefenhinweisen profitiert, weisen übrigens Viswanathan und Mingolla [VM98, VM99] nach. Es ist daher erstaunlich, wie unbekannt eine dreidimensionale Salienzrepräsentation in der Literatur zur Steuerung von Aufmerksamkeit ist.

Von speziellem Interesse ist das Verhalten des Systems bei temporären Okklusionen, die für die Verfolgung durch Neuronale Felder bisher das Hauptproblem darstellten, aber auch hinsichtlich der

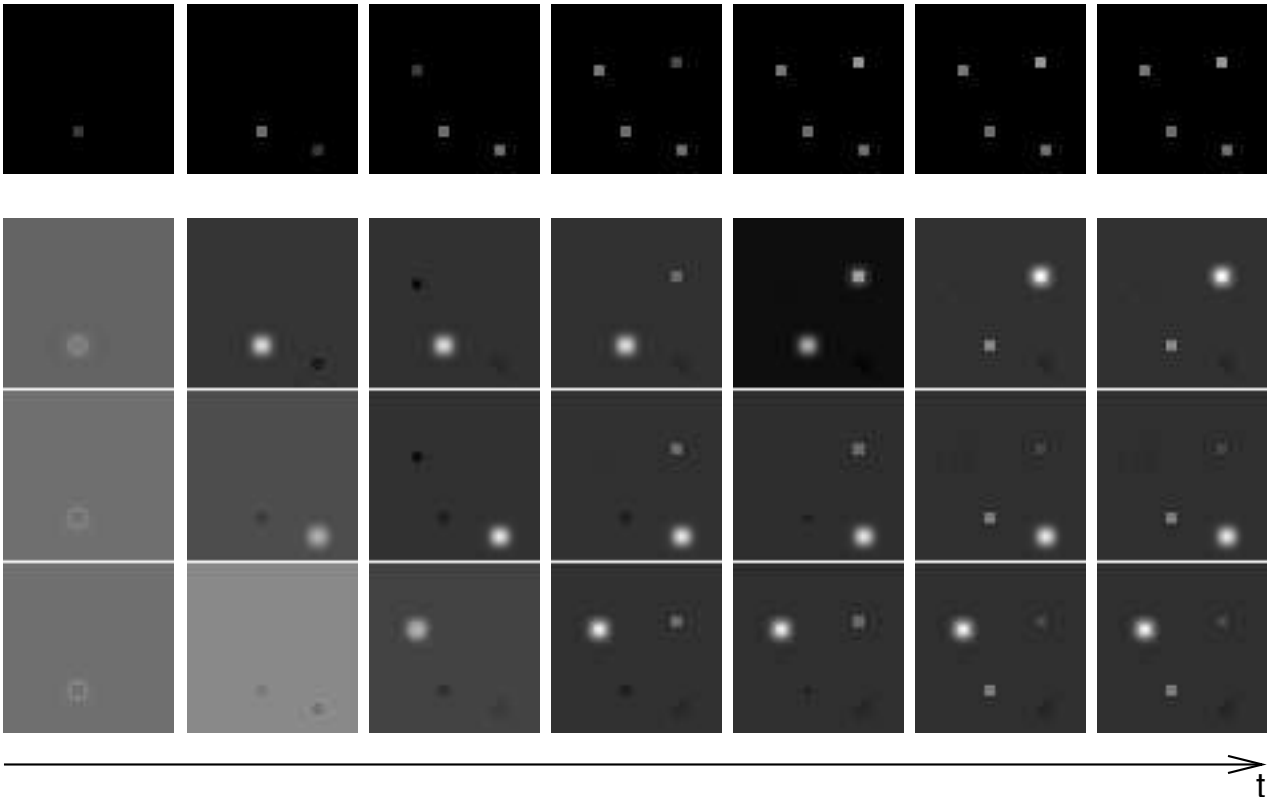


Abbildung 6.12: Verhalten eines Systems Neuronaler Felder bei Darbietung mehrerer auffälliger Objekte. Die obere Reihe zeigt die Eingabe für die Felder. Vier Objekte tauchen nacheinander auf. Unten ist die Aktivation der Felder zu sehen, wobei im dritten und vierten Frame ein Zustand erreicht ist, bei dem alle Felder ein Aktivationscluster aufweisen. Das neue vierte Objekt ist mit höherer Auffälligkeit ausgestattet und verdrängt das schwächste der vorhandenen Aktivationscluster.

Segmentation von Objekten für die genauere Selektion einzelner Objekte. Schließlich ist hier zu klären, inwieweit eine Ausdehnung auf eine zusätzliche Dimension den Rechenaufwand für die Aktualisierung der Felder unverhältnismäßig erhöht und wie dieser Aufwand in Grenzen gehalten werden kann.

Dreidimensionale Neuronale Felder wurden unabhängig von der in [BM00] vorgestellten ersten Variante von Braumann [Bra01] verwendet, wobei jedoch die dynamische Veränderung der Reize und eine damit zusammenhängende Verfolgung genausowenig eine Rolle spielen wie die Berechnung von Tiefendaten. Die Rechenzeitprobleme wurden durch Verwendung eines sehr kleinen Neuronalen Feldes (27*22*5 Neuronen) umgangen, wodurch die Konnektivität direkt von der zweidimensionalen Version ausgedehnt werden konnte.

6.4.1 Modellierung der Konnektivität in der Tiefe

Auch für diesen Fall ist die Diskussion zu führen, ob ein NF lokaler Feldinhibition oder aber mehrere Felder globaler Feldinhibition zum Einsatz kommen sollen. Da der Aufwand für die Berechnung der Dynamik Neuronaler Felder direkt von der Anzahl der Neuronen abhängt, ist davon auszugehen, dass im dreidimensionalen Fall die weitaus größere Anzahl von Neuronen, die durch die Berücksichtigung einer zusätzlichen Dimension zustande kommt, zu einer entsprechenden Erhöhung des Aufwandes führen wird. Diese noch zusätzlich durch die Verwendung mehrerer Felder zu vervielfachen, erscheint unangebracht. Außerdem soll die dreidimensionale Repräsentation zu einer besseren Trennung der Objekte führen, die sich anhand einer zweidimensionalen Repräsentation nicht erreichen lässt, was eine Verwendung mehrerer solcher Felder unnötig macht. Somit soll also ein dreidimensionales lokal inhibitives Neuronales Feld verwendet werden.

Neben der Anzahl der Neuronen geht auch die Größe des Verbindungskernels unmittelbar in den Berechnungsaufwand ein. Dieser würde sich bei Verwendung eines DoG-Kernels in allen drei Dimensionen um die Größe des Kernels in der dritten Dimension vervielfachen. Im Falle des Systems zweidimensionaler Neuronaler Felder wurde der Aufwand für die Verbindungen gering gehalten, indem Verbindungen zwischen den Feldern auf die identische (2D-)Position reduziert wurden. Ein ähnliches Vorgehen wird auch hier vorgeschlagen, um zusätzlich zu einem zweidimensionalen Verbindungskernel reine Verbindungen in einer einzelnen Dimension zu verwenden. Naheliegend wäre, auch an dieser Stelle einen DoG-Kernel zur Spezifikation der Verbindungsgewichte zu verwenden.

Allerdings muss man beachten, welcher Art die Eingabedaten sind. Es handelt sich hier um ursprünglich zweidimensionale Salienzdaten, die anhand mehrerer Tiefenhypothesen in eine dreidimensionale Struktur eingetragen werden. Das gleichzeitige Vorhandensein mehrerer Aktivitätscluster an derselben zweidimensionalen Position bei jedoch unterschiedlichen Tiefen ist insofern nicht erwünscht. Vielmehr soll die Dynamik des Feldes dazu führen, dass aus den Hypothesen mittels räumlicher und temporaler Integration ein einzelnes Cluster gebildet wird. Das würde dafür sprechen, die Gewichte in dieser Dimension auf Art und Weise der globalen Feldinhibition zu definieren, nämlich als Gaußverteilung abzüglich eines konstanten Wertes.

Eine weitere Vereinfachung zur reinen Inhibition, wie es für den erwähnten Fall des Systems zweidimensionaler Felder vorgenommen wurde, erscheint jedoch nicht sinnvoll. Dadurch würde die lokale Anregung wegfallen und Eingaben mit Disparitätswerten, die zwar nicht identisch sind, aber nahe beeinander liegen, würden sich gegenseitig hemmen, obwohl sie in der Tiefe dicht beieinanderliegen.

In diesem Zusammenhang muss auch die Größe des Feldes in der dritten Dimension diskutiert werden. Zu beachten ist der Aufwand, der durch die Größe entsteht, weswegen die Größe so gering wie möglich zu halten ist. Entscheidend ist hier eben nicht die genaue Repräsentation der Tiefe, um eine quantitative Rekonstruktion zu ermöglichen, sondern eine qualitative Trennung der Objekte in solche, die sich näher und andere, die sich weiter entfernt befinden, wobei Objekte mit ähnlicher räumlicher Tiefe als zusammenhängend angesehen werden sollen. Damit wird klar, dass sich die Auflösung in der Tiefe in einer anderen Größenordnung bewegt als die Auflösung in den anderen Dimensionen. Um eine Trennung mehrerer Objekte zu ermöglichen, wurde in Anlehnung an die unterschiedenen Disparitätsstufen der Stereoberechnung 11 Neuronenschichten in der Tiefe festgelegt. Bei dieser Größenordnung vereinfacht sich die Festlegung der Gewichte zu:

$$w(x-x') = \begin{cases} k * \exp(\frac{x-x'}{\sigma^2}) - k_2 * \exp(\frac{x-x'}{\sigma_2^2}) & , x_z = x'_z \\ H_1 & , x_y = x'_y,\ x_x = x'_x,\ |x_z - x'_z| = 1 \\ -H_2 & , x_y = x'_y,\ x_x = x'_x,\ |x_z - x'_z| > 1 \end{cases} \quad (6.12)$$

Dabei bezeichnen H_1und H_2 zwei positive Konstanten für die Anregung der Aktivität in den direkt benachbarten Tiefenschichten und die Inhibition in allen anderen Tiefenschichten. Dies stellt eine Vereinfachung der Verbindungen in der Tiefe dar, die erheblich zur Begrenzung des Aufwandes beiträgt.

Abb. 6.13 stellt die Verwendung dieser Struktur dar.

6.4.2 Experimente

Von primären Interesse im Vergleich der vorgestellten Architekturen Neuronaler Felder ist das Verhalten bei der Verfolgung von Objekten, die sich temporär verdecken. Das zur Untersuchung dieser Problematik konzipierte Experiment besteht aus einer großen Anzahl von Durchgängen, in denen eine Anzahl von Objekten (von 2 bis 7) für 15 Zyklen so bewegt wurden, dass es mindestens eine Verdeckungssituation gab. Gemessen wurde für jede Architektur die durchschnittliche Dauer der korrekten Verfolgung der Objekte. Wurden also alle Objekte in allen Durchgängen für den jeweils ganzen Durchgang korrekt verfolgt, ergibt sich ein Wert von 15.

Abb. 6.14 zeigt, dass die einfache Version eines einzelnen zweidimensionalen Neuronalen Felder lokaler Inhibition unter diesen Umständen die schlechtesten Ergebnisse aufweist. Für die beiden elaborierteren Varianten gilt, dass die Verfolgungleistung des Systems zweidimensionaler Felder globaler Inhibition eine etwas bessere Verfolgungsleistung aufweist, solange die Anzahl der Objekte nicht zu groß wird. Zu beachten ist dabei jedoch, dass die Information im dreidimensionalen Neuronalen Feld insofern reichhaltiger sind, als sie eine zusätzliche räumliche Lokalisierung erlauben.

6.5 Zusammenfassung und Diskussion

Es wurde gezeigt, dass Neuronale Felder eine geeignete Wahl sind, die neu definierte Aufgabe der ersten Selektionsstufe als integrierte Selektion und Verfolgung mehrerer salienter Elemente zu lösen. Neben der Verfolgung zeigen diese Felder Eigenschaften wie Hysterese, räumlich-zeitliche Integration und Bifurkation, die, worauf auch in der Literatur immer wieder hingewiesen wird, eine Selektion

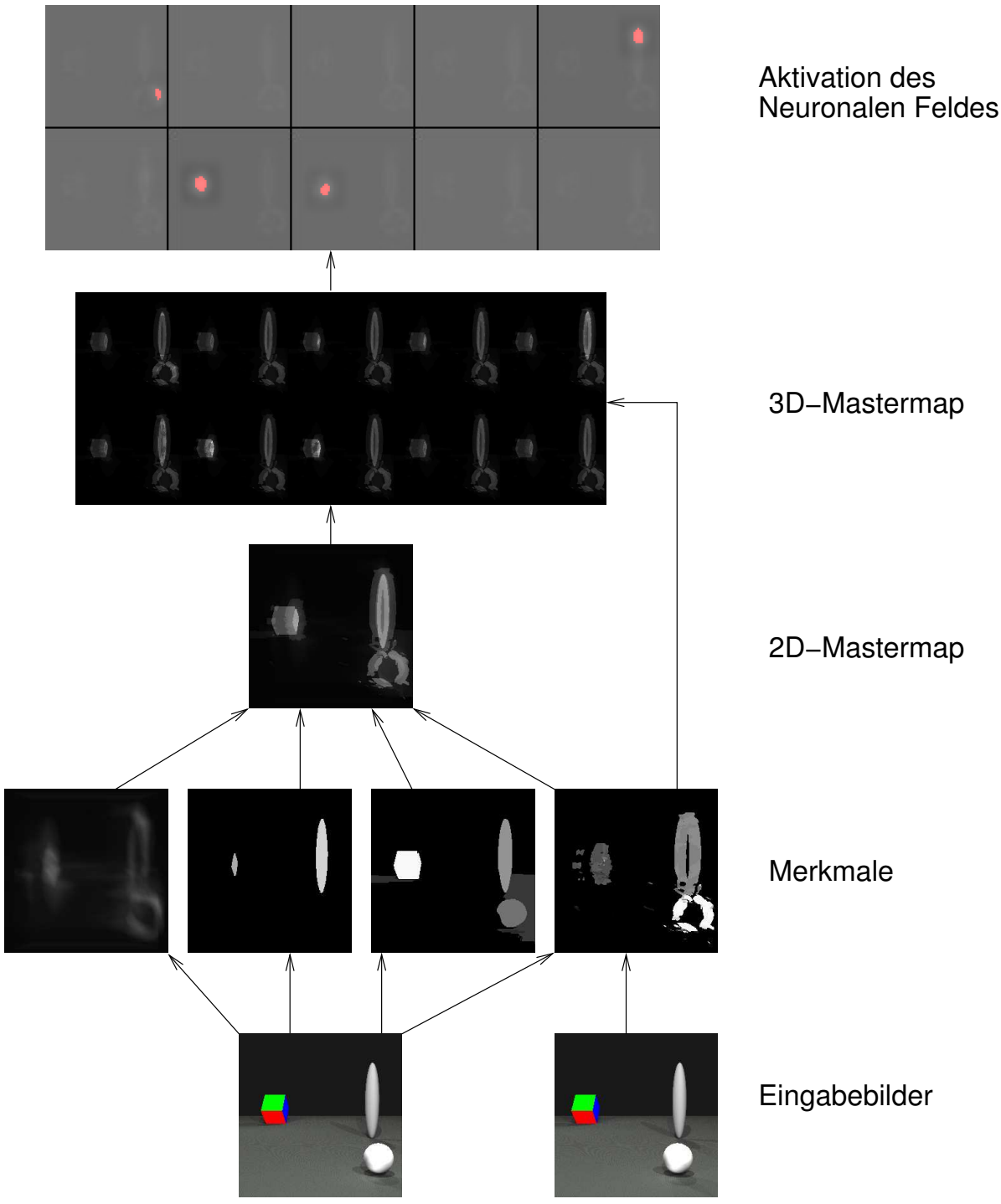


Abbildung 6.13: Verwendung eines dreidimensionalen Neuronalen Feldes mit lokaler Feldinhibition. Im Neuronalen Feld sind die Bereiche sigmoider Aktivation farblich hervorgehoben.

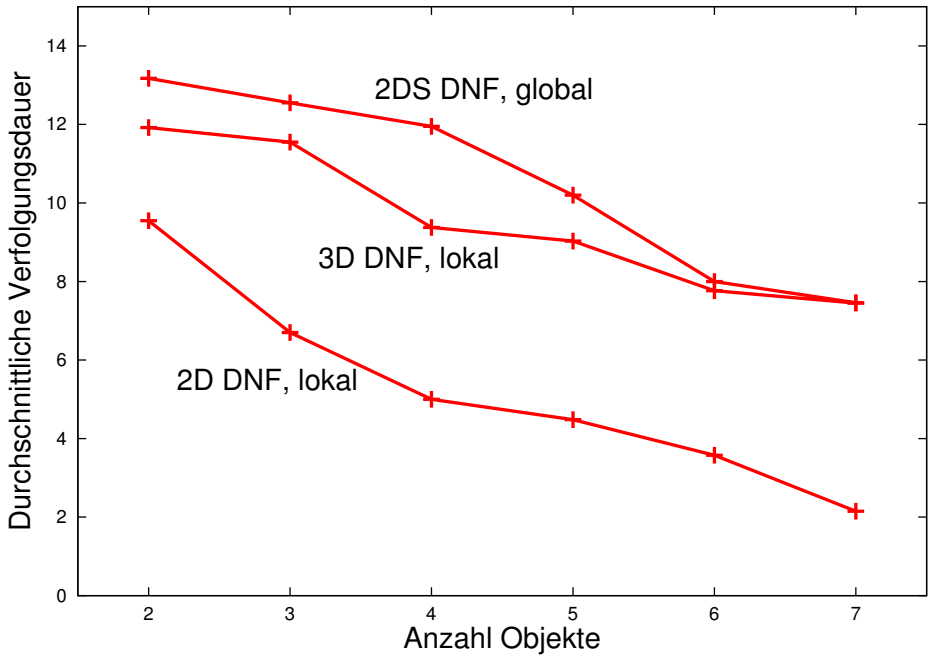


Abbildung 6.14: Verhalten bei temporärer Okklusion in der Verfolgung mehrerer Objekte durch verschiedene Varianten der Neuronalen Felder.

in stark verrauschten Eingabedaten erlauben. Abhängig von unterschiedlichen Möglichkeiten der Salienzrepräsentation sind unterschiedliche Architekturen der Neuronalen Felder geeignet, die im Rahmen dieser Arbeit entwickelt wurden.

Was die Auswahl der geeigneten Architektur angeht, sind folgende Hinweise zu beachten. Das einfache Modell eines zweidimensionalen Neuronalen Feldes lokaler Feldinhibition zeichnet sich zwar durch die schnellste Berechnung aus, ist aber hinsichtlich der Selektion und Verfolgung mehrerer Objekte, die in räumlicher Nähe auftreten können, den beiden anderen Modellen deutlich unterlegen. Die Entscheidung zwischen diesen Modellen ist primär von der Verfügbarkeit und Qualität von Disparitätsinformationen abhängig. Besteht das verwendete Aktive Sehsystem aus mindestens zwei Kameras, die zur Berechnung der Disparität geeignete Aufnahmen bereitstellen können, ist die Variante eines dreidimensionalen Feldes aufgrund der besten Verfolgungsleistung und der umfangreichen Salienzinformationen zu bevorzugen. Andernfalls, oder in Fällen, in denen die Rechenleistung für ein dreidimensionales Feld nicht ausreicht, sollte auf die Variante mehrerer zweidimensionaler Felder globaler Feldinhibition zurückgegriffen werden.

Zusammenfassend ist festzustellen, dass die erste Selektionsstufe unter Verwendung eines Systems Neuronaler Felder eine Möglichkeit darstellt, aus der subsymbolischen Kartenrepräsentation von Auffälligkeiten eine kleine Anzahl auffälliger Bereiche robust auszuwählen und auch unter schwierigen Umständen zu verfolgen, ohne dabei auf eine Erkennung der zugrunde liegenden Objekte angewiesen zu sein.

Die Verfolgung der selektierten Bereiche hoher Salienz gehört streng betrachtet in die mittlere, semiattentive Verarbeitungsstufe, weil sie nur mittelbar durch die zeitliche Integration von Information über bewegte Objekte mit der Selektion dieser Bereiche zu tun hat. Die Integration mit der ersten Selektionsstufe erlaubt jedoch ein einfacheres Modell und wird daher vorgezogen. Weiterer

Bestandteil der mittleren Verarbeitungsstufe sind - je nach Anwendung - Prozesse, die Information über die selektierten Elemente bestimmen, die sich so einfach berechnen lässt, dass sie keine fokale Selektion und damit zusammenhängende Serialisierung voraussetzt.

Natürlich könnte man prinzipiell die erste Selektionsstufe aus zwei Mechanismen zusammensetzen, von denen der eine für die Selektion, der andere für die Verfolgung zuständig ist. Gerade im Bereich der Verfolgung finden sich in der Literatur andere leistungsfähige Verfahren, in letzter Zeit werden vor allem die auf Partikelfilterung beruhenden Verfahren wie der Condensation-Algorithmus [IB98a, IB98b] oft und erfolgreich eingesetzt. Nachteil dieses Vorgehens ist jedoch, dass beide Bestandteile - das Tracking und die Selektion - gegenseitig voneinander abhängig sind. Diese Abhängigkeit in zwei Verfahren zu integrieren, würde das Modell weniger effizient und unnötig komplizierter werden lassen.

Kapitel 7

Zweite Selektionsstufe: Der Fokus der Aufmerksamkeit

Die Aufgabe der zweiten Selektionsstufe ist es, aus den Resultaten der ersten Selektionsstufe einen einzelnen klassischen Fokus der Aufmerksamkeit auszuwählen. Hier wird also der Übergang zur rein attentiven Verarbeitung hergestellt. Dies passiert aufgrund der zur Verfügung gestellten Datenbasis ausschließlich symbolisch. Die zweite Selektionsstufe ist - stärker als die erste Stufe - modellgetriebenen top-down-Einflüssen unterworfen und stellt so eine entscheidende Schnittstelle für die Verwendung der Aufmerksamkeitssteuerung durch weitere Systeme dar. Wichtigste Datenstruktur dieser Stufe sind die Objectfiles, die eine einfache symbolische Beschreibung der Selektionskandidaten darstellen.

7.1 Ziel

In Kapitel 3.5.2 wurde darauf verwiesen, dass die übliche Einteilung der Verarbeitung in einen parallelen, präattentiven und einen seriellen, attentiven Teil nicht ausreicht. Das Vorhandensein einer rein seriellen, attentiven Stufe ist jedoch kaum anzuzweifeln. Denn einerseits können Berechnungen so komplex sein, dass eine komplette Serialisierung notwendig ist, andererseits kann die Spezifikation von Aktionen die Auswahl eines einzigen verhaltensrelevanten Objektes verlangen. Somit muss ein Aufmerksamkeitsmodell auch über eine derartige Stufe verfügen, die zu jedem Zeitpunkt nur ein einzelnes Element enthält.

Die Selektion dieses einzelnen Elementes soll im vorgestellten Modell alleine auf den Resultaten der ersten Selektionsstufe beruhen, da die erste Stufe bereits alle in Frage kommenden Objekte ausreichender Auffälligkeit ausgewählt und mit zusätzlichen Informationen angereichert hat. Die zweite Selektionsstufe nimmt diese Objekte, erzeugt aus ihnen jeweils eine geeignete symbolische Struktur und trifft eine Auswahl unter diesen Strukturen. Die Auswahl geschieht in Abhängigkeit der Inhalte dieser Strukturen, der Historie der Auswahl und einem Verhaltensmodell. Die zur Verfügung stehenden Verhaltensmodelle werden in Kapitel 8.5 vorgestellt, in diesem Kapitel soll die Transformation in eine geeignete Datenstruktur und das prinzipielle Vorgehen bei der Auswahl im Vordergrund stehen.

Es ist zu beachten, dass in dieser Stufe der für Systeme des Bildverstehens wichtige Übergang von subsymbolischer, signalnaher und konnektionistischer Verarbeitung zur symbolischen Verarbei-

tung vorgenommen wird. Die symbolische Repräsentation ist vor allem deswegen angezeigt, da es sich um eine diskrete kleine Anzahl von Objekten handelt. Es kann hier von den genauen Signaleigenschaften abstrahiert werden, da diese - soweit bedeutsam - in die symbolische Repräsentation des Objektes eingegangen sind. In Anlehnung an die psychophysische Modellierung wird diese symbolische Repräsentation als Objectfile (OF) bezeichnet. Neben der Bedeutung für die Selektion stellen die Objectfiles aber auch eine entscheidende Datenstruktur für das Gedächtnis des Systems dar. Indem von den umfangreichen symbolischen Daten abstrahiert wurde und für die relevanten Teile des Bildes eine kompakte symbolische Struktur erstellt wurde, die eine kontinuierliche Beschreibung dieser Teile enthält, wird die Speicherung und Verarbeitung bedeutsamer Daten vereinfacht.

7.2 Objectfiles als symbolischer Speicher

Das Vorhandensein weniger diskreter Objekte legt die Verwendung symbolischer Verfahren zur Selektion nahe. Um jedoch genügend Informationen über diese Objekte zur Verfügung zu haben, ist eine symbolische Beschreibung der Information notwendig. Diese Erkenntnis führte in der Psychophysik zur Beschreibung von *object files* [KTG92, WB97] (siehe auch Kap. 3.3.1), die präattentiv gebildet werden. Entscheidende Eigenschaft der Objectfiles ist ihre Bindung an ein Objekt, die unabhängig von der Objektidentität stattfindet. Sie wird von Kahneman und Treisman [KTG92] anschaulich beschrieben durch die Zitierung von Menschen, die in einem Film dasselbe Objekt mit verschiedenen Identitäten belegen, aber eindeutig immer dasselbe Objekt meinen („It's a bird. It's a plane. It's superman!"). Diese Objekthaftigkeit unabhängig von der Identität, die Trennung von Objekten, bevor sie erkannt worden sind und bevor ihnen eventuell auch Aufmerksamkeit zugewiesen wurde, wird durch Objectfiles modelliert. Sie enthalten einerseits Informationen, die über das Objekt gesammelt wurden, aber vor allem einen - wie auch immer gearteten - Zeiger auf das Objekt, der dem Objekt folgt und den Bezug im Laufe der Verarbeitung des Objektes erhält.

7.2.1 Anlegen von Objectfiles

Im vorgestellten Modell wird ein Objectfile für jede Selektion der ersten Stufe angelegt, d.h. für jeden zusammenhängenden aktiven Bereich im Neuronalen Feld. Voraussetzung dafür ist ein einfaches Labeling der Aktivation, das Schwerpunkt, Anzahl der Pixel und den Bereich der jeweils zusammenhängenden Aktivationen liefert. Dieses wird nach jeder Aktualisierung der Neuronalen Felder durchgeführt. Prinzipiell existiert zu jedem Zeitpunkt für jedes dieser Aktivitätscluster genau ein Objectfile. Die weiteren Aktionen, die im Folgenden erläutert werden, sind:

- die Sammlung zusätzlicher Informationen für das Objectfile, das sich einerseits auf Bildeigenschaften, andererseits auf den Selektionsprozess selbst und die Resultate höherer Prozesse bezieht, sowie

- die Korrespondenzbildung zwischen Objectfiles, die festlegen soll, ob und wenn ja, welches der zuvor erstellen Objectfiles dem aktuell erstellten entspricht, sich also auf dasselbe Objekt bezieht.

7.2.2 Informationen in einem Objectfile

Die in einem Objectfile enthaltenen Informationen beziehen sich auf Merkmalsinformationen, räumliche und zeitliche Lokalisation, Selektion durch fokale Aufmerksamkeit, Ergebnisse höherer Verarbeitungsstufen und die Historie des Objectfiles. Jedes Objectfile wird bei der Erzeugung mit einem eindeutigen Label versehen. Die wichtigste Information eines Objectfiles ist der Verweis auf die aktuelle Position, sofern eine solche existiert (siehe inaktive Objectfiles später). Im Falle dreidimensionaler Neuronaler Felder bezieht die Position die Tiefe mit ein, andernfalls handelt es sich um eine zweidimensionale Position. Über diese findet die Adressierung der Objectfiles statt. Die Position wird in Bildkoordinaten angegeben und mit der Nummer des Frames assoziiert. Für jeden Frame ist die Ausrichtung der Kamera gespeichert, so dass für Objectfiles unterschiedlicher Zeitpunkte festgestellt werden kann, ob die Kameraposition identisch ist und somit die Bildkoordinaten vergleichbar sind. Andernfalls kann eine Anpassung der Koordinatensysteme anhand der Kamerabewegung unter der Voraussetzung erfolgen, dass sich die Plattform nicht bewegt hat.

Zusätzlich werden in einer Historie die Zeitpunkte des Auftretens mit den assoziierten Orten gespeichert. Ebenso verhält es sich mit der Information, ob das Objectfile momentan selektiert ist, beziehungsweise wann es zuvor selektiert war. Diese Informationen sind entscheidend, um dem Verhaltensmodell die Selektion der Objectfiles zu ermöglichen. Anhand der Positionen im Laufe der Zeit können zusätzlich Informationen über die Trajektorie und damit Bewegungsrichtung und -geschwindigkeit abgeleitet werden. Diese werden zwar im Rahmen des vorgestellten Modells nicht weiter verwendet, stehen jedoch übergeordneten Auswertungsprozessen zur Verfügung.

Die Position des Objectfiles ergibt sich aus dem zugehörigen Aktivitätscluster. Für den aktuellen Frame werden jeweils alle zugehörigen Pixel gespeichert. Daraus werden der Schwerpunkt, die Anzahl der zugehörigen Pixel und eine Bounding Box, also die Grenzen eines minimalen Rechtecks, das das Aktivitätscluster enthält, berechnet, die für alle Frames gespeichert werden. In Abschnitt 7.3.2 wird eine Erweiterung diskutiert, mit deren Hilfe eine verbesserte Segmentierung der Objekte vorgenommen wird.

Sowohl zur Klassifikation und Erkennung als auch zur Bestimmung von Korrespondenzen werden Merkmalsinformationen in Objectfiles abgelegt. Diese werden anhand des bekannten Bereiches des Aktivitätsclusters aus den Merkmalskarten extrahiert. Zusätzlich zu den aktuellen Merkmalsinformationen für den betrachteten Eingabeframe wird ein gemittelter historischer Wert mitgeführt, der in leaky-integrator-Form aktualisiert wird, so dass ältere Informationen geringer gewichtet eingehen. Inwieweit nur die aktuellen Werte oder die gemittelten Werte von höheren Prozessen genutzt werden, kann jeweils dort entschieden werden.

Im Zusammenhang mit der Aufmerksamkeitssteuerung ist wichtig, wann ein Objekt selektiert war und wann nicht. Der Effekt der *inhibition of return* hängt zum Beispiel entscheidend davon ab, wann ein Objekt zuletzt selektiert wurde, damit eine zu frühe erneute Selektion zu vermeiden. Schließlich werden die Ergebnisse höherer Prozesse wie Klassifikation und Erkennung in den Objectfiles gespeichert, um alleine durch sie eine symbolische Beschreibung der aktuellen Szene anhand der wichtigsten Objekte mit ihren Eigenschaften ableiten zu können.

Zusammengefasst lässt sich ein Objectfile wie in Abb. 7.1 visualisieren.

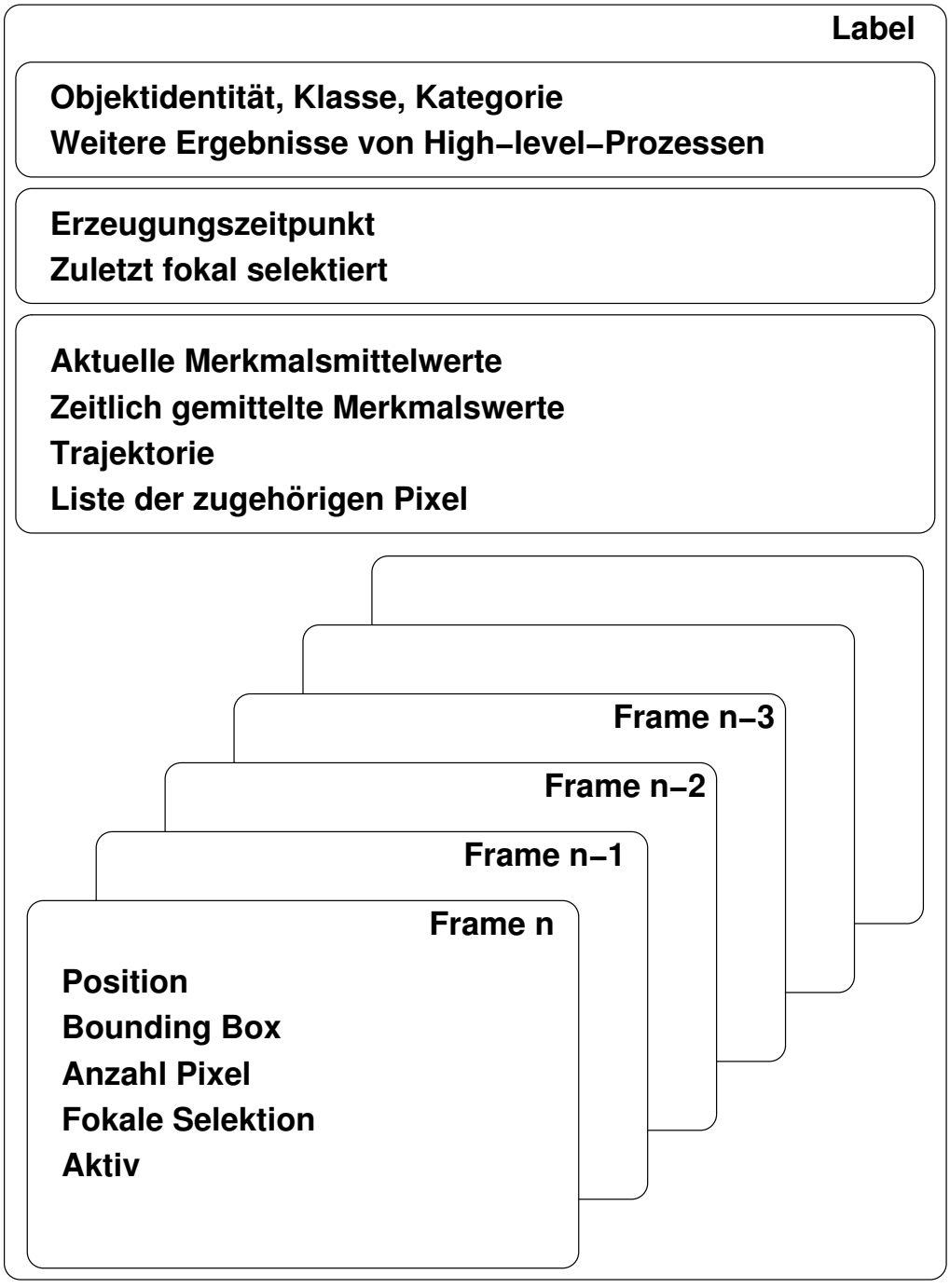


Abbildung 7.1: Schematische Darstellung des Inhalts von Objectfiles.

7.2.3 Korrespondenz von Objectfiles und Aktivitätsclustern

Kahneman et al. [KTG92] identifizieren drei wichtige Operationen auf Objectfiles:

- **Korrespondenzbestimmung**: Überprüfung, ob es sich um ein neues Objekt handelt oder es eine Korrespondenz zu einem bereits existierenden Objectfile gibt? Wenn es eine Korrespondenz gibt, wo befindet sie sich?
- **Review**: Zugriff auf Objekteigenschaften im Objectfile, wenn das zugrunde liegende Objekt aktuell nicht sichtbar ist.
- **Impletion**: Suche nach einer Veränderung oder Bewegung, die eine plausible Verbindung zwischen dem aktuellen Zustand und den Informationen über vorige Zustände herstellt.

Von entscheidender Bedeutung für die zweite Selektionsstufe ist der erste Prozess der Korrespondenzbestimmung. Er stellt sicher, dass sich die Objectfiles tatsächlich auf Objekte und nicht auf statische Orte beziehen. Die Charakteristik der Verfolgung der Bereiche maximaler Salienz durch die Neuronalen Felder ist von der Art der verwendeten Feldstruktur abhängig. So ist im Fall des Systems Neuronaler Felder (Kapitel 6.3.2) durch die Verwendung Neuronaler Felder des globalen Inhibitionstyps sichergestellt, dass jedes Feld jederzeit nur ein einzelnes Aktivitätscluster aufweist, was die Korrespondenzbildung vereinfacht. Für die Modellvarianten mit Neuronalen Feldern lokaler Inhibition (Kapitel 6.3.1 und 6.4) gilt dies nicht. In jedem Fall ist ein Mechanismus notwendig, der die Korrespondenz von Objectfile und Aktivitätsclustern herstellt. Hierzu lassen sich die räumliche Nähe sowie Ähnlichkeiten hinsichtlich der vorkommenden Merkmale verwenden.

System Neuronaler Felder globaler Feldinhibition

Aufgrund der Eigenschaften Neuronaler Felder globaler Feldinhibition kann man davon ausgehen, dass das einzelne Aktivitätscluster, das sich in jedem einzelnen DNF befindet, meist stabil demselben Objekt folgt. Wechselt die Selektion von einem Objekt zu einem anderen, findet typischerweise eine Unterdrückung des Aktivitätsclusters statt. Hieraus leitet sich die ursprüngliche Korrespondenzhypothese her, die im folgenden verifiziert wird. Zur Verifikation dient der räumliche Abstand der Schwerpunkte. Ist dieser kleiner als ein Schwellwert, wird die Hypothese akzeptiert. Der Schwellwert ergibt sich aus dem typischen Radius der Aktivitätscluster zu x_a, dem Bereich in dem eine lokale Maximumssuche stattfindet (s. Kap. 6.2.4).

Für die durch den einfachen Schwellwert nicht zuzuordnenden Objectfiles und Aktivitätscluster wird diejenige Zuordnung getroffen, bei der die summierten Fehler (als Abstände von Objectfileschwerpunkt und Aktivitätsclusterschwerpunkt) minimal sind. Da es sich nur um eine kleine Zahl von Feldern und Objectfiles handelt, stellt die Untersuchung aller Zuordnungskonstellationen kein Problem dar. Wird dabei der doppelte Schwellwert als Distanz überschritten, wird ein neues Objectfile erzeugt und dem Aktivitätscluster zugeordnet, das existierende OF wird als inaktiv markiert.

Einzelnes Neuronales Feld lokaler Feldinhibition

Für die zweidimensionale und dreidimensionale Version Neuronaler Felder lokaler Feldinhibition, also solcher Felder, in denen die Präsenz mehrerer Aktivitätscluster möglich ist, dient als erster Hinweis

zur Korrespondenzbildung die aktuelle Position des Aktivitätsclusters. Da die Grenzen, innerhalb derer die Verfolgung durch Neuronale Felder stattfindet, anhand der Parameter des Neuronalen Feldes bestimmt werden können, wird zunächst versucht, eine Korrespondenz innerhalb des aus diesen Werten resultierenden Radius herzustellen. Als Position wird jeweils der Schwerpunkt der Aktivationsbereiche angesehen. In der Korrespondenzbildung in dreidimensionalen Neuronalen Feldern ist dabei die dritte Dimension anders zu gewichten, da sie nicht aufgrund gemessener Salienz zustande kommt, sondern von der Qualität der Tiefenrekonstruktion abhängt.

Für jedes Aktivitätscluster werden zunächst die Objectfiles ausgewählt, deren Schwerpunkt in 2D-Bildkoordinaten weniger als x_a (s. Kap. 6.2.4) vom 2D-Schwerpunkt des Aktivitätsclusters entfernt ist und deren Entfernungsschwerpunkt höchstens um eine Maximaldistanz vom entsprechenden Schwerpunkt des Aktivitätsclusters entfernt ist. Die Maximaldistanz wurde empirisch auf zwei Tiefenschichten des Neuronalen Feldes festgelegt.

Soweit diese Zuordnungen eindeutig sind (jeweils genau ein OF für ein Aktivitätscluster), werden sie vorgenommen. Die weitere Analyse beschäftigt sich nur noch mit den verbleibenden Aktivitätsclustern und Objectfiles, wobei alle Kombinationen von Zuordnungen auf den Fehler bezüglich der Schwerpunkte und hinsichtlich der Merkmalsdistanz zwischen Objectfile und Aktivitätscluster geprüft werden. Allen Aktivitätsclustern, bei denen der Fehler zu einem OF hinsichtlich beider Kriterien (Ort und Ähnlichkeit) minimal ist, wird dieses Objectfile zugeordnet. Bei allen jetzt noch verbleibenden Fällen werden neue Objectfiles zugeordnet, die verbliebenen als inaktiv vermerkt.

Eine besondere Behandlung erfolgt, wenn sich zwei Aktivitätscluster zu einem einzigen vereinigen. Dies wird im Laufe der Aktualisierung der Neuronalen Felder überprüft. In diesem Falle wird ein neues Objectfile kreiert, das einen Verweis auf die beiden zuvor zugeordneten Objectfiles enthält. Nach einer kleinen Anzahl von Frames (in den durchgeführten Experimenten waren es 4) werden die Merkmalswerte des neuen OF mit den beiden Vorgängern verglichen. Ergibt sich eine eindeutige Zuordnung (die Differenz zu einem beträgt weniger als die Hälfte der Differenz zum anderen), wird diese vorgenommen, andernfalls wird der Verweis entfernt und die beiden älteren Objectfiles als inaktiv eingetragen.

Gemeinsames Vorgehen bei allen Typen Neuronaler Felder

Grundsätzlich wird für jedes Aktivitätscluster, dem kein bestehendes Objectfile zugeordnet werden konnte, ein neues OF angelegt und passend initialisiert. Objectfiles, denen umgekehrt kein Aktivitätscluster mehr zugeordnet werden konnte, werden als inaktiv markiert. Außer dieser Markierung finden keine Aktualisierungen im Objectfile statt, Merkmalsinformationen und sonstige Informationen bleiben erhalten. Vor dem Erzeugen eines neuen Objectfile wird jedoch geprüft, ob unter den inaktiven Objectfiles ein Kandidat zu finden ist, der in Korrespondenz zum Aktivitätscluster steht. Zur Korrespondenzbildung werden primär die Merkmalseigenschaften herangezogen.

Die Operationen auf Objectfiles stellt Abb. 7.2 dar. Von einem Frame zum nächsten werden dabei zuerst die Zuordnungen von Objectfiles zu Aktivitätsclustern innerhalb der vorgegebenen Grenzen aktualisiert. Dies ist im Beispiel für die Objectfiles 1 und 2 möglich, nicht jedoch für das dritte. Dieses wird daraufhin als inaktives Objectfile gespeichert. Für nun nicht zugeordnete Aktivitätscluster werden neue Objectfiles angelegt.

Anhand eines konkreten Beispiels zeigt Abb. 7.3 die Bezüge der Objectfiles zu Objekten.

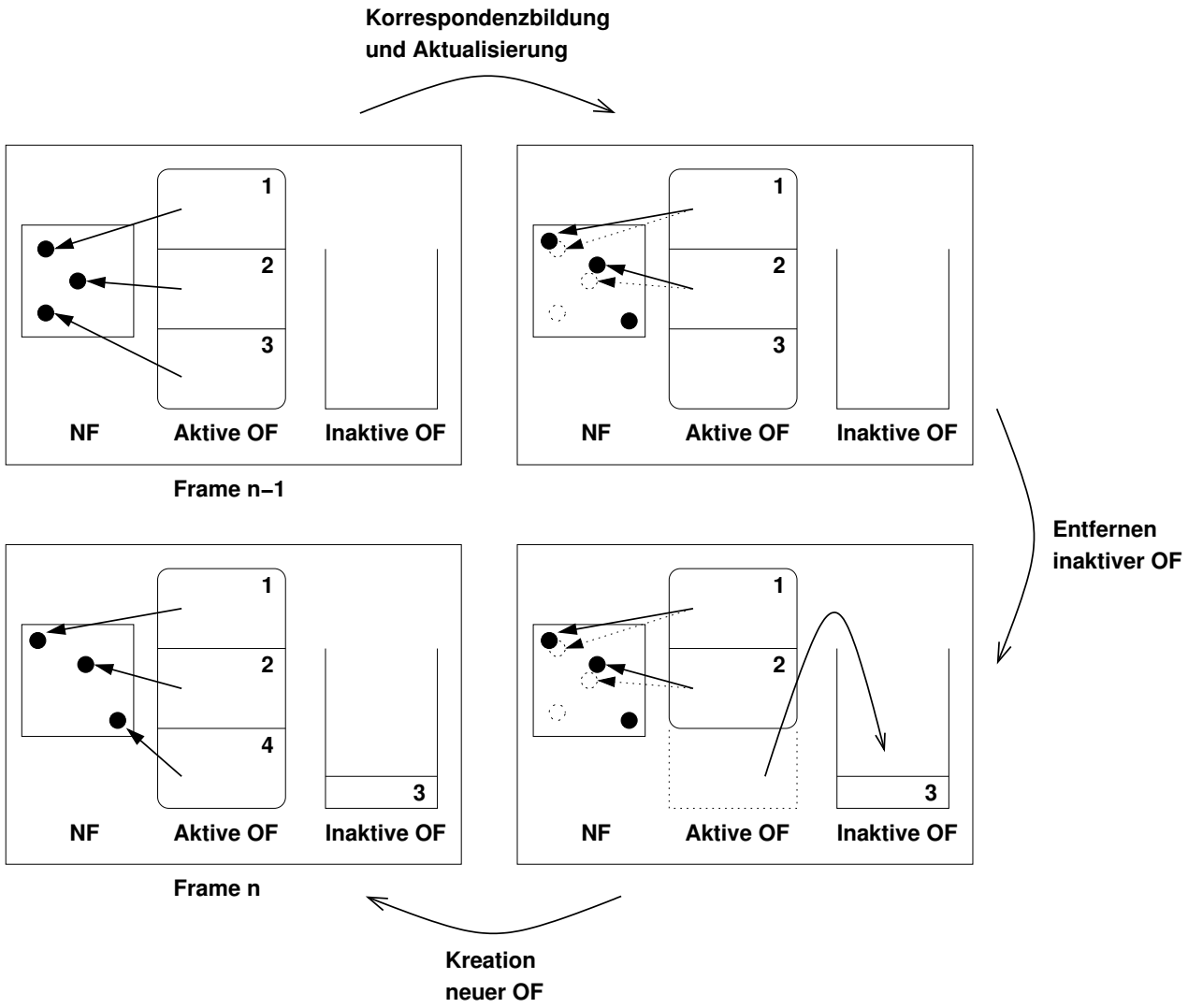


Abbildung 7.2: Schematische Darstellung der Verwendung von Objectfiles (OF) anhand der Aktivitätscluster im Neuronalen Feld (NF): Korrespondenzsuche und Aktualisierung der enthaltenen Informationen, Entfernung inaktiver Objectfiles, Erzeugung neuer Objectfiles.

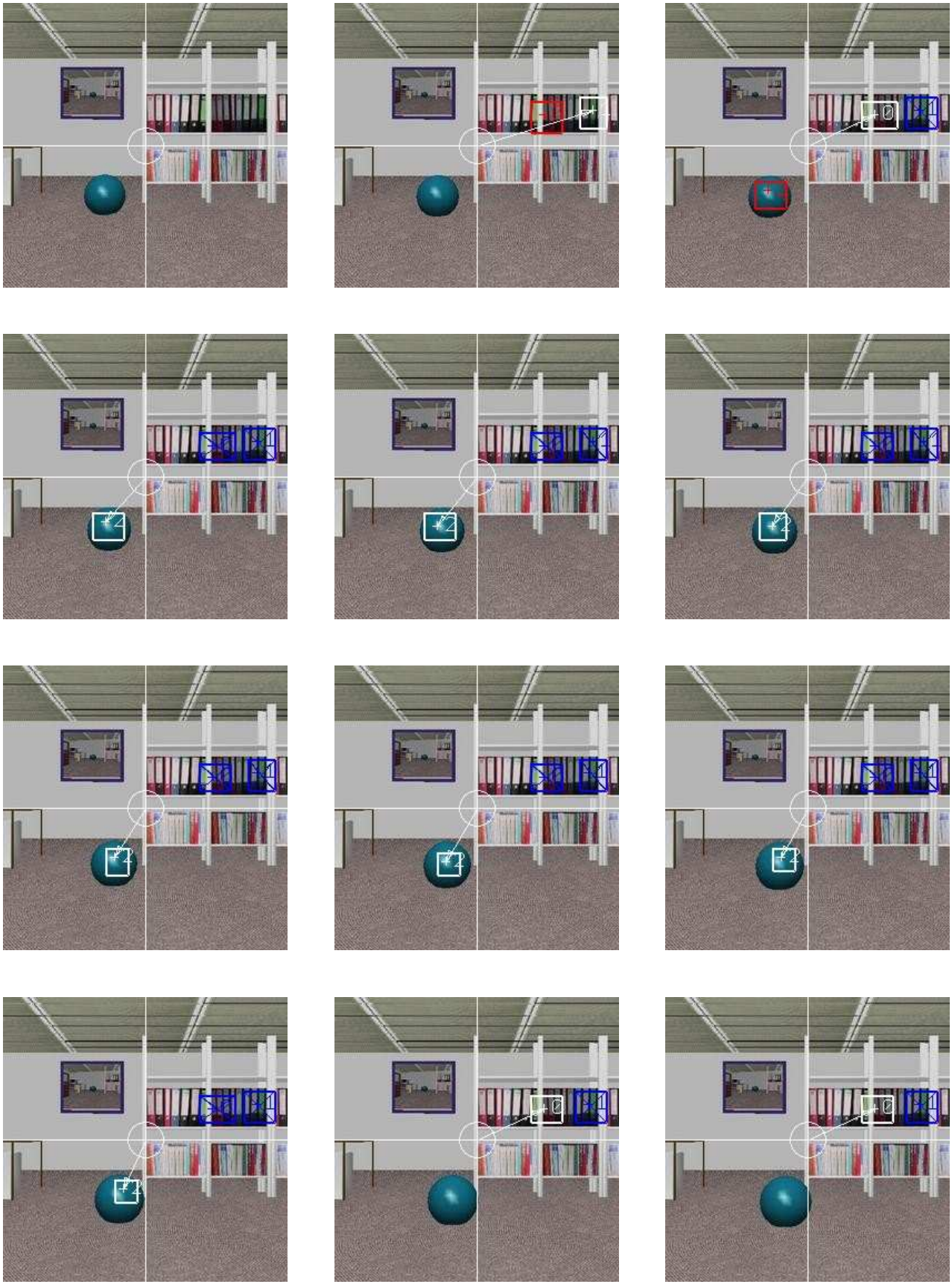


Abbildung 7.3: Bezüge von Objectfiles zu Orten bzw. Objekten in einer dynamischen Beispielszene. Dabei sind die Objectfiles durch ihre Nummer und die Boundingbox der Aktivität markiert. Die Farben deuten den Status an: das weiße Element ist momentan ausgewählt, blaue Elemente wurden zuvor selektiert, rote Elemente sind neu. Es ist zu beachten, dass ein Objectfile dem sich bewegenden Ball folgt.

7.2.4 Aktive und inaktive Objectfiles

Objectfiles lassen sich danach unterscheiden, ob für sie momentan eine Korrespondenz zu einem aktuell sichtbaren Objekt besteht, d.h. ob die Bindung zu einem Aktivitätscluster existiert. Ist dies nicht der Fall, handelt es sich um ein inaktives Objectfile, das zur Selektion nicht zur Verfügung steht. Wird ein Objectfile inaktiv, wird es in eine Stack-ähnliche Struktur eingefügt. Sie dient vor allem dazu, bei neuen Aktivitätsclustern, denen kein aktives Objectfile zuzuordnen ist, eine Korrespondenz zu einem inaktiven Objectfile festzustellen. Ist dies möglich, wird das Objectfile aus dem Speicher entfernt (was, sofern es sich nicht um das oberste Element handelt, von der Stackstruktur abweicht).

Weiterhin wird ein Maximalalter festgelegt und Objectfiles, die länger inaktiv sind, werden ebenfalls entfernt, um den Speicher nicht beliebig wachsen zu lassen. Dieses Maximalalter wird primär durch die Charakteristika von Aufgabe und Umgebung und weiterhin vom Aufwand zur Suche nach Korrespondenzen bestimmt.

7.3 Fokale Selektion

7.3.1 Auswahl von Objectfiles

Die Selektion eines Objectfiles für die Zuweisung fokaler Aufmerksamkeit soll alleine anhand der Informationen in den Objectfiles stattfinden. Diese Einschränkung trägt zur Modularität und Begrenzung der notwendigen Kommunikation und Abhängigkeiten bei. Somit operiert diese Stufe alleine anhand symbolischer Daten und kann vergleichsweise einfach implementiert werden. Dies ist vor allem deswegen von Bedeutung, weil die zweite Stufe sehr viel aufgabenabhängiger und spezifischer arbeitet als die erste Selektionsstufe. Das System erlaubt sowohl die Spezifizierung von allgemeinen Verhaltensmodellen als auch die Implementation stark spezialisierter und aufgabenbezogener Modelle.

Ein Aspekt, der in allen allgemeinen Modellen vorhanden sein soll, ist die *inhibition of return* (s. Kap. 3.2). Sie geschieht anhand der Information, wann die Objectfiles zuletzt mit fokaler Aufmerksamkeit selektiert wurden und priorisiert solche, die lange nicht mehr selektiert wurden. Diese Inhibition of return kann zur Exploration als allgemeine Regel eingesetzt werden, sie kann aber auch mit anderen Mechanismen kombiniert werden. Auch die Aktionen, die mit der Auswahl eines Objectfiles zusammenhängen, wie die Ausführung komplexer Operationen oder die Ausrichtung von Sensoren, ist von solchen Verhaltensmodellen abhängig, die in Kapitel 8 weiter beschrieben werden.

7.3.2 Bestimmung des Fokus der Aufmerksamkeit

Mit der Auswahl eines Objectfiles hängt immer auch eine räumliche Selektion zusammen. Der Fokus der Aufmerksamkeit wird auf die Position des Objectfiles ausgerichtet. Jedoch beschreibt ein Fokus der Aufmerksamkeit nicht nur einen Punkt, auf den etwa eine Kamera ausgerichtet werden könnte. Vielmehr enthält er ein Bildsegment, das aus genau denjenigen Punkten besteht, die zum ausgewählten Objectfile gehören. Dieses Bildsegment stellt eine Hypothese dar für den Bereich des Objektes, auf das sich das OF (Objectfile) bezieht. Auf die Bedeutung von Objekten und Objektformen in der attentiven Selektion weisen Hamker und Gross [HG97] hin. In einem zweistufigen Modell wird bei ihnen jedoch die Selektion nur eines einzelnen Objektes betrieben.

Anstelle der Aktivitätscluster der Neuronalen Felder soll als Ausblick noch eine weitergehende Segmentierung diskutiert werden. Die Form der Aktivitätscluster hängt zum einen natürlich von der Eingabe ab, zum anderen wird sie jedoch auch durch die Charakteristik der (punktsymmetrischen) Gewichtsfunktion beeinflusst. Um eine bessere Abschätzung der zu einem Objekt gehörenden Punkte zu kommen, können folgende Hinweise ausgenutzt werden:

- Form/Bereich des Aktivitätsclusters
- Homogenität der Merkmale (speziell Tiefe) - Die Merkmale wurden ausdrücklich so entworfen, dass sie relevante Objekteigenschaften wiedergeben. Demnach weist eine Homogenität der Merkmale auf einen Zusammenhang der Pixel zu räumlichen Objekten hin.
- Segmentierungen, die für die Merkmalsberechnungen vorgenommen wurden (hier speziell Exzentrizität und Farbe)
- Segmentierungen desselben Objektes (zu bestimmen anhand des Objectfiles) in vorausgehenden Frames

Denkbar wäre ein Seeded Region Growing-Verfahren, das vom (durch Erosion reduzierten) Aktivitätscluster ausgehend das Wachstum von der Homogenität der Merkmale, der Überschreitung von Grenzen der Merkmalssegmentierungen und der Übereinstimmung mit zuvor segmentierten Formen abhängig macht. Grinias und Tziritas [GT98] stellen ein ähnliches Verfahren im Kontext von Bewegungssegmentierung und -verfolgung vor.

7.4 Zusammenfassung und Diskussion

Mit den Objectfiles wird eine aus der Modellbildung der Psychophysik stammende Struktur eingesetzt, um das Ergebnis der ersten Selektionsstufe zu repräsentieren. Sie dient hier wie dort primär der Bindung von objektbasierten und räumlichen Informationen unter dynamischen Bedingungen. Dies erlaubt eine rein symbolische Auswahl eines dieser Objectfiles für fokale Aufmerksamkeit. Es wurden Mechanismen vorgestellt, die diese Bindung aufrechterhalten, wobei die Mechanismen von der verwendeten Struktur Neuronaler Felder abhängen.

Die Objectfiles stellen außerdem einen wichtigen Teil des Weltmodells und damit des Gedächtnis des Systems dar, indem sie kompakt Eigenschaften mehrerer relevanter Objekte in jedem Frame enthalten. In dieser Hinsicht geht die Verwendung über das natürliche Vorbild hinaus. Als Ausblick wurde eine genauere Bestimmung des Segmentes skizziert, das dem selektierten Objekt entspricht. Weiterhin wäre eine Erweiterung der Korrespondenzbildung um die Einbeziehung von Ähnlichkeit denkbar. Diese könnte auf den Merkmalsinformationen ansetzen, die in den Objectfiles enthalten ist.

Kapitel 8

Verhaltensmodelle und Aktives Sehen

Nachdem die bisherige Verarbeitung im Wesentlichen unbeeinflusst vom genauen Kontext des Sehsystems, seiner Aufgabe und seinen Fähigkeiten modelliert wurde, somit eine primär datengetriebene Verarbeitung umsetzte, werden in Verhaltensmodellen die modellgetriebenen Komponenten gekapselt. Sie steuern vor allem die zweite Selektionsstufe. Hier ist es möglich, das System so zu spezifizieren, dass es definierte Aufgaben erfüllt. Auch die Aktivität des Systems durch Ausrichtung der Sensoren liegt in der Verantwortung solcher Verhaltensmodelle, von denen einige allgemeine im Folgenden beschrieben werden. Hervorgehoben werden soll jedoch die wohldefinierte Schnittstelle, die ein einfaches Hinzufügen oder Modifizieren der Verhaltensmodelle erlaubt.

8.1 Ziel

Um einerseits top-down-Einflüsse auf das System zu modellieren, andererseits eine Spezialisierung und Einbindung der Aufmerksamkeitssteuerung in konkrete Aktive Sehsysteme zu ermöglichen, werden sogenannte Verhaltensmodelle verwendet [BMB01]. Sie kapseln den modellgetriebenen und aufgabenabhängigen Einfluss auf den Ablauf der Aufmerksamkeitskontrolle. In diesem Kapitel wird geklärt, an welcher Stelle und über wie geartete Schnittstellen Verhaltensmodelle die Verarbeitung beeinflussen können. Zwei zentrale Aspekte sind dabei die *inhibition of return* und die Ausführung von Blickbewegungen durch ein Aktives Sehsystem.

Die Ausführung wird dann an mehreren Beispielen von Verhaltensmodellen illustriert, die die Modellierung psychophysischer Experimente ebenso wie wichtige Aufgaben Aktiver Sehsysteme abdecken. Verhaltensmodelle enthalten dazu zwei Aspekte: einmal steuern sie vollständig die zweite Selektionsstufe, also die Auswahl eines Objectfiles für fokale Aufmerksamkeit und die damit zusammenhängende Auslösung von Blickbewegungen, zum anderen beinhalten sie alle spezifischen Modifikationen des bisher vorgestellten Modells, die zur Lösung konkreter Aufgaben notwendig sind.

Die Architektur ist dabei so gedacht, dass sich möglichst viele Aufgaben alleine durch Steuerung der zweiten Selektionsstufe erzielen lassen, ohne wesentliche Eingriffe in das restliche System vorzunehmen. Abb. 8.1 gibt die Einordnung der Verhaltensmodelle in das gesamte System wieder.

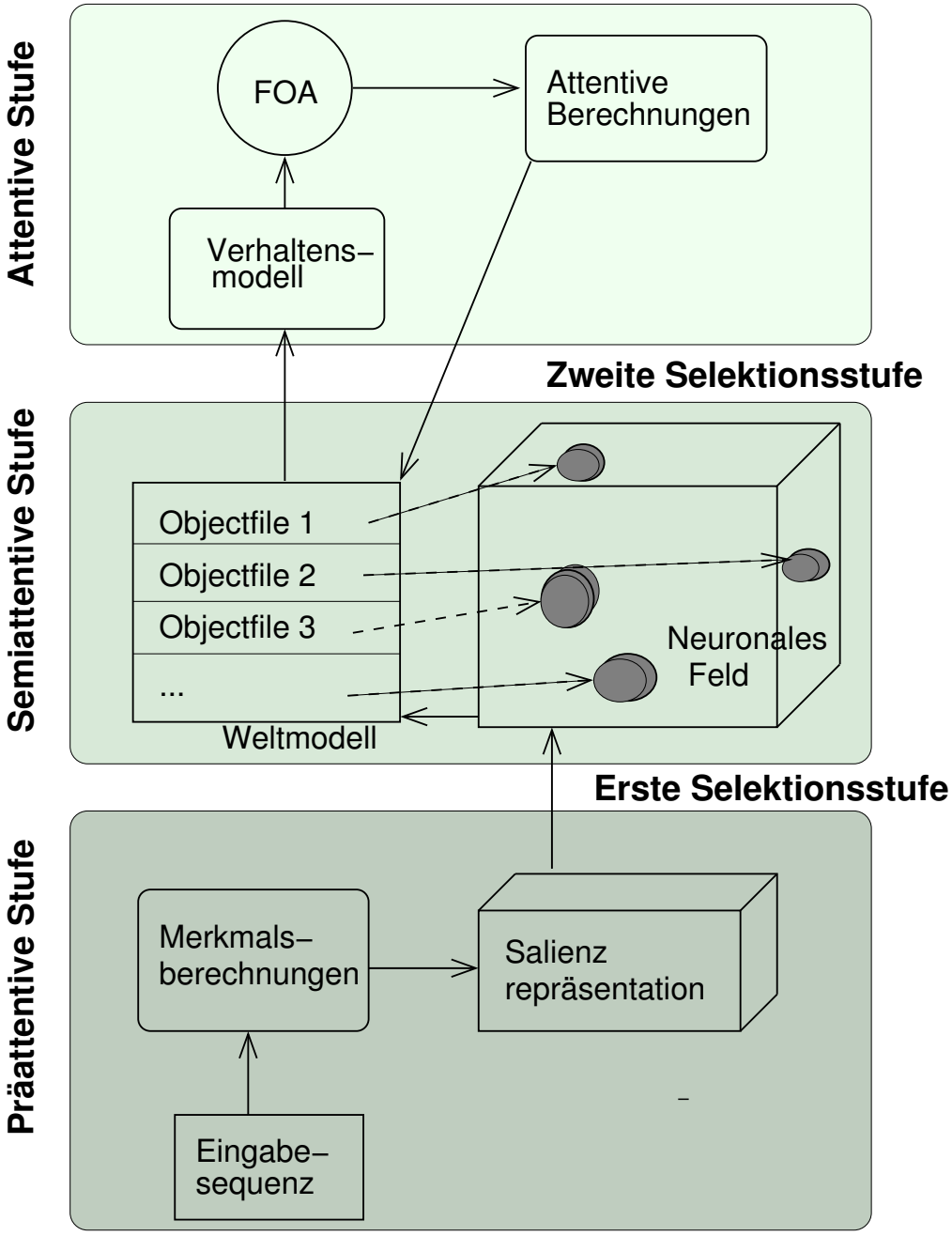


Abbildung 8.1: Überblick über das Aufmerksamkeitsmodell und Einordnung der Verhaltensmodelle.

8.2 Einfluss auf die Selektionsstufen

Für die vorgestellten Verarbeitungsschritte, speziell die beiden Selektionsstufen, sind geeignete Schnittstellen zu definieren, an denen die Kontrolle durch Verhaltensmodelle greifen kann. Beide Selektionsstufen sind unterschiedlich stark durch die Verhaltensmodelle beeinflusst. Während die Arbeit der ersten Selektionsstufe vollständig beschrieben wurde und der Einfluss eines Verhaltensmodells hauptsächlich eine Bestimmung von Parametern oder die Abweichung von einem normalen Ablauf darstellt, ist der Einfluss auf die zweite Selektionsstufe wesentlich stärker. Die zweite Selektionsstufe besteht primär aus Mechanismen, deren genaue Steuerung von einem Verhaltensmodell abhängt.

8.2.1 Erste Selektionsstufe

Ein wichtiger top-down-Einfluss auf die Kontrolle von Aufmerksamkeit, der in vielen Modellen der Psychophysik genannt wird, ist die Gewichtung von Merkmalen. So wird davon ausgegangen, dass bei der Visuellen Suche nach vorher spezifizierten Zielreizen die Merkmale, die den Zielreiz von den Distraktoren unterscheiden, höher gewichtet werden. Zu den Modellen, die eine solche Gewichtung vorsehen, gehört zum Beispiel das Guided Search-Modell von Wolfe [WCF89, Wol94, WG96]. Zu unterscheiden ist dabei, ob eine Gewichtung der Merkmale untereinander oder aber eine Gewichtung bestimmter Merkmalsausprägungen stattfindet. In der psychophysischen Literatur wird das erstere meist als Gewichtung von Dimensionen bezeichnet, während es sich bei letzterem um eine Gewichtung der Merkmale handelt. Es ist dabei eine offene Frage, inwieweit letzteres überhaupt möglich ist, über die Gewichtung von Dimensionen besteht jedoch weitgehend Konsens.

In diesem Modell wird neben der Gewichtung von Dimensionen jedoch auch die Gewichtung von Merkmalsausprägungen zugelassen. Neben einer festgelegten dauerhaften Gewichtung ist es auch möglich, die Gewichte situationsbedingt anzupassen. Das erlaubt zum Beispiel einem mobilen System, während der Bewegung das Merkmal Tiefe insgesamt hoch zu gewichten, um beliebigen Hindernissen ausweichen zu können, während im statischen Fall vielleicht bestimmte Farben und Orientierungen zur Detektion von Menschen höher gewichtet werden. Besonders zu diskutieren ist die Merkmalsgewichtung in der Variante eines Systems Neuronaler Felder, für die implizit bereits eine individuelle Gewichtung der Merkmale für jedes der Felder stattfindet. Hier werden beide Gewichtungssysteme hintereinandergeschaltet, so dass sie sich multiplikativ beeinflussen. Zusätzliche Möglichkeiten zur Gewichtung von Merkmalen finden sich auch in der zweiten Selektionsstufe (s. 8.2.2).

Ein weiterer Einfluss, der in diesem Modell nur in Ansätzen umgesetzt wird, ist die Kontrolle der Ressourcen für die Merkmalsberechnung. Die beiden dafür zentralen Aufgaben sind die Bestimmung der Menge an Ressourcen, die jedem Merkmal in Abhängigkeit von Aufgabe, Reizen und momentanem Zustand zuzuordnen sind, auf der einen Seite, sowie die Umsetzung einer Ressourcenkontrolle für die Merkmalsberechnungen auf der anderen Seite. Letztere geschieht durch geeignete Parametrisierung der Merkmalsberechnungen. Für die Merkmale, die eine Multiskalenberechnung enthalten, also Symmetrie und Tiefe, ist die Auslassung der höchsten Auflösung der am besten geeignete Weg, eine deutlich beschleunigte Berechnung ohne Veränderung der wesentlichen Merkmalscharakteristik zu erreichen. Für die Exzentrizität ist die Reduktion der Dilationszyklen eine Möglichkeit, bei der Farbe bietet sich die Berechnung für ein Bild mit reduzierter Auflösung an.

8.2.2 Zweite Selektionsstufe

Der zentrale Aspekt der zweiten Selektionsstufe, die der Kontrolle durch ein Verhaltensmodell unterliegt, ist die Auswahl eines der Objectfiles für den Fokus der Aufmerksamkeit. Diese Auswahl erfolgt primär anhand

- der Daten innerhalb der Objectfiles (datengetrieben),
- des aktuellen Zustandes und
- des Ziels oder der Aufgabe, die dem System derzeit zugeordnet ist (modellgetrieben).

Sie wird durch ein Verhaltensmodell gesteuert; mehrere solche Modelle für unterschiedliche Aufgaben werden in Abschnitt 8.5 vorgestellt.

Die Gewichtung von Merkmalen als ein wichtiger Aspekt der Kontrolle von Aufmerksamkeit kann sich auch in dieser Stufe auswirken. Da die Objectfiles Informationen über die momentane und zeitlich gemittelte Präsenz der Merkmale im Bereich der OF enthalten, lässt sich die Auswahl eines Objectfiles auch daran ausrichten, wie präsent ein bestimmtes Merkmal ist.

Ein datengetriebener Einfluss, der beide Selektionsstufen miteinbezieht, besteht in der Addition von Salienz im Bereich spezifischer Aktivitätscluster. Dies kann Sinn machen, wenn die Bedeutung eines Objektes durch einen fokalen Prozess bestimmt wurde und insofern unabhängig von den Merkmalen zur Berechnung datengetriebener Salienz wird. Um das Aktivitätscluster aufrecht zu erhalten, ist die Ergänzung der Eingabe in die Neuronalen Felder um eine Aktivation im Bereich des Aktivitätsclusters möglich. Damit werden zwar die Merkmale nicht völlig ausgeschlossen, es erlaubt aber die Stabilisierung der Selektion beim Erscheinen weiterer, noch auffälligerer Objekte. Im Falle von sich bewegenden Objekten müsste ein zusätzlicher einfacher Verfolgungsmechanismus ergänzt werden, da die Verfolgung durch die Neuronalen Felder eben auf der merkmalsbestimmten Auffälligkeit beruht.

8.3 Inhibition of return

Eines der Argumente für die zweistufige Selektion mit Auswahl und Verfolgung mehrerer salienter Einheiten war die Möglichkeit zur objektbasierten Selektion und Inhibition. Insofern spielt die Inhibition of return für kürzlich mit fokaler Aufmerksamkeit versehene Objekte eine Sonderrolle bei den Verhaltensweisen. Sie stellt einen grundlegenden Mechanismus dar, den zu überschreiben für den Menschen offensichtlich selbst dann Aufwand bedeutet, wenn bekannt ist, dass er die Verarbeitung hemmen kann. Daher sollte die Modellierung einer solchen IOR Bestandteil der Verhaltensmodelle sein.

Eine objektbasierte Inhibition of return wird durch ein Verhaltensmodell umgesetzt, das eine Hemmung von Objectfiles durchführt, die kürzlich mit fokaler Aufmerksamkeit versehen wurden. Diese Hemmung kann durch Priorisierung von OF erreicht werden, die lange nicht ausgewählt wurden. In den einzelnen Verhaltensweisen ist die Interaktion dieser Priorisierung mit den anderen Zielen zu gestalten.

Je nach Charakteristik der Umgebung kann es sinnvoll sein, zusätzlich eine raumbasierte IOR durchzuführen. In diesem Fall ist eine leichte Anpassung der ersten Selektionsstufe notwendig. Sie

besteht in der Implementation einer statischen Inhibitionskarte, wie auch andere Modelle sie vorsehen. In dieser Karte werden bei jeder Auswahl eines OF die zugehörigen Pixel mit einer hohen Aktivation versehen. Die gesamte Aktivation der Karte verringert sich im Laufe der Zeit. Sie wirkt inhibitiv auf die *master map of attention*. Im Falle des Systems Neuronaler Felder, wo es keine solche Mastermap gibt, wirkt die Inhibitionskarte inhibitiv auf jede einzelne Eingabe der verschiedenen Neuronalen Felder. Im Falle eines dreidimensionalen Neuronalen Feldes wird entsprechend auch eine dreidimensionale Inhibitionskarte verwendet.

Ob eine solche Karte verwendet werden soll, wie schnell das Abklingen der Aktivität stattfindet und mit welchem Gewicht die Inhibition die Mastermap beeinflusst, hängt vom jeweiligen Verhaltensmodell ab.

8.4 Ausführung von Sakkaden

Eine zentrale Verhaltensweise der offenen Aufmerksamkeit besteht in der Fovealisierung von visuellen Objekten. Diese geschieht beim natürlichen Vorbild grundsätzlich unter vorheriger Zuweisung verdeckter Aufmerksamkeit [DES00], was auch in diesem Modell so umgesetzt werden soll. Die Spezifikation einer Sakkade findet grundsätzlich anhand des ausgewählten Objectfile statt. Eine Sakkade wird durch das aktuelle Verhaltensmodell initiiert.

Das Ziel zur Ansteuerung durch die Kameras wird als Schwerpunkt des Objectfiles definiert. Aus diesen Koordinaten werden die Parameter abgeleitet, die eine entsprechende Fovealisierung durch die Kameras erlauben. Das erste Kamerabild nach Ausführung dieser Bewegung wird verwendet, um eine eventuelle Korrektur der Sakkade durchzuführen. Um die Notwendigkeit einer Korrektur und die entsprechenden Parameter zu bestimmen, wird der Bildbereich des selektierten OF direkt vor der Sakkade gespeichert. Nach der Sakkade erfolgt eine Korrelation des gespeicherten Bereiches in einer Nachbarschaft um das Bildzentrum. Befindet sich das Zentrum außerhalb einer definierten Umgebung des Zentrums, wird sofort eine entsprechende weitere Kamerabewegung initialisiert, ohne weitere Berechnungen (Merkmalsberechnung, Aktualisierung des Neuronalen Feldes) anhand des aktuellen Bildes durchzuführen. Auch für das neue Bild wird die Korrelation durchgeführt, allerdings nicht, um eine weitere Korrektur auszulösen, sondern nur, um Kenntnis über den verbleibenden Fovealisierungsfehler zu erhalten und den Positionsverweis des fovealisierten OF zu aktualisieren.

Nach Ausführung der Sakkade ist es wichtig, die internen Repräsentationen zu aktualisieren, die sich auf Bildkoordinaten beziehen. Dazu gehören als erstes die Neuronalen Felder. Die Aktivität wird so verschoben, dass die Aktivationen im aktualisierten Neuronalen Feld den zu erwartenden Positionen entsprechen. Neuronen, die auf Positionen verweisen, die neu ins Bild gelangen, werden mit dem Ruhewert des Neuronalen Feldes initialisiert. Bei der Korrespondenzbildung zwischen Objectfiles und Aktivitätsclustern, die auf eine Sakkade folgt, werden zwei Modifikationen vorgenommen. Die Schwellwerte zur Entscheidung der Ähnlichkeit werden um einen konstanten Faktor erhöht, um der Ungenauigkeit, die durch die Kamerabewegung erzeugt wird, gerecht zu werden. Weiterhin werden für einen konstanten Zeitbereich alle inaktiven Objectfiles ausgesucht, die sich zuletzt im neu ins Bild gekommenen Bereich befanden. Für dort entstehende Aktivitätscluster werden sie zur Korrespondenzbildung herangezogen.

Die Ausführung der Kamerabewegung wird zusammen mit der Nummer des aktuellen Frames

vermerkt, so dass ein Abgleich der Positionen innerhalb der OF möglich ist. Für die aktiven OF findet eine Anpassung der Positionen statt.

Für die Zukunft wäre die Erstellung eines extraretinalen Speichers für die Aktivität in Neuronalen Feldern denkbar, so dass Bildbereiche, die zuvor sichtbar waren, im Neuronalen Feld nicht neu initialisiert werden müssen, wenn sie wieder ins Bild rücken. Dies macht nur dann Sinn, wenn von einer im Wesentlichen statischen Umgebung ausgegangen wird. In einer statischen Umgebung würde derselbe Ort eine ähnliche Antwort der Salienzberechnung erzeugen und insofern den Nachteil der Initialisierung von Teilen der Neuronalen Felder abschwächen. Da jedoch das System auf die Verarbeitung dynamischer Szenen ausgerichtet ist und der Vorteil auch bei einer Bewegung der Plattform wegfallen würde, wurde hier auf die Umsetzung eines solchen Konzeptes verzichtet.

Eine Sakkade ist verglichen mit einer internen Verlagerung der Aufmerksamkeit mit hohen Kosten assoziiert. Die ungenaue Lokalisierung der Aktivitätscluster kann zu einem Verlust des Kontaktes führen. Die Ausführung der Sakkade (und eventuellen Korrektursakkade) sowie die Anpassung der internen Repräsentation brauchen Zeit, in der die sonstigen Berechnungen nicht durchgeführt werden können. Die Initialisierung von Teilen der Neuronalen Felder führen zu mehr Aktualisierungszyklen, die wiederum die Bearbeitung folgender Frames verzögern. Dieser Abgleich von Problemen und Kosten mit den Vorteilen ist bei der Spezifikation eines Verhaltenmodells zu beachten, so dass Sakkaden nur dann ausgelöst werden, wenn der Vorteil diese Kosten überwiegt.

8.5 Verhaltensmodelle

Die hier beschriebenen Verhaltensmodelle stellen Beispiele dar, die in konkreten Applikationen Verwendung finden können. Es soll deutlich werden, welche Möglichkeiten bestehen und wie gering der Aufwand ist, ein solches Verhaltensmodell zu implementieren. Für jedes Verhalten werden die Parameter mit angegeben, die zur konkreten Spezifizierung des Modells benötigt werden.

8.5.1 Exploration

Als zentrales Verhalten des Systems, das ohne genauere Spezifizierung von Parametern auskommt, ist die Exploration angelegt. Sie drückt das Ziel des Systems aus, bei Abwesenheit speziellerer Vorhaben möglichst viele Informationen über die Umgebung zu sammeln und in einem möglichst vollständigen und aktuellen Weltmodell der wichtigsten Objekte vorzuhalten.

Dies wird anhand der Bestimmung von zwei Eigenschaften der Objectfiles erreicht: der datengetriebenen Salienz und der Dauer seit der letzten Selektion des OF. Anhand dieser Dauer werden die OF in Prioritätsklassen eingeteilt. Die oberste Prioritätsstufe gilt für alle OF, die noch nie fokal selektiert wurden. Die weiteren Stufen werden anhand der Dauer seit dieser Selektion angeordnet, beginnend mit der längsten Zeit. Innerhalb einer Prioritätsstufe werden die OF hinsichtlich der durchschnittlichen Salienz im Bereich der zugehörigen Pixel geordnet. Die Selektion des nächsten OF wird ausgelöst, sobald die attentiven Berechnungen für das aktuelle OF abgeschlossen sind. Kamerabewegungen werden durch dieses Verhalten nicht initiiert.

Die Abb. 8.2 und 8.3 demonstrieren die Operation dieser Verhaltensweise an einem Beispiel. Die Laborszene wurde für 18 Frames beobachtet. Im Bild ist die Eingabe jeweils mit den Bereichen der Objectfiles annotiert. Zusätzlich wird der Fokus der Aufmerksamkeit dargestellt. Es ist im Beispiel

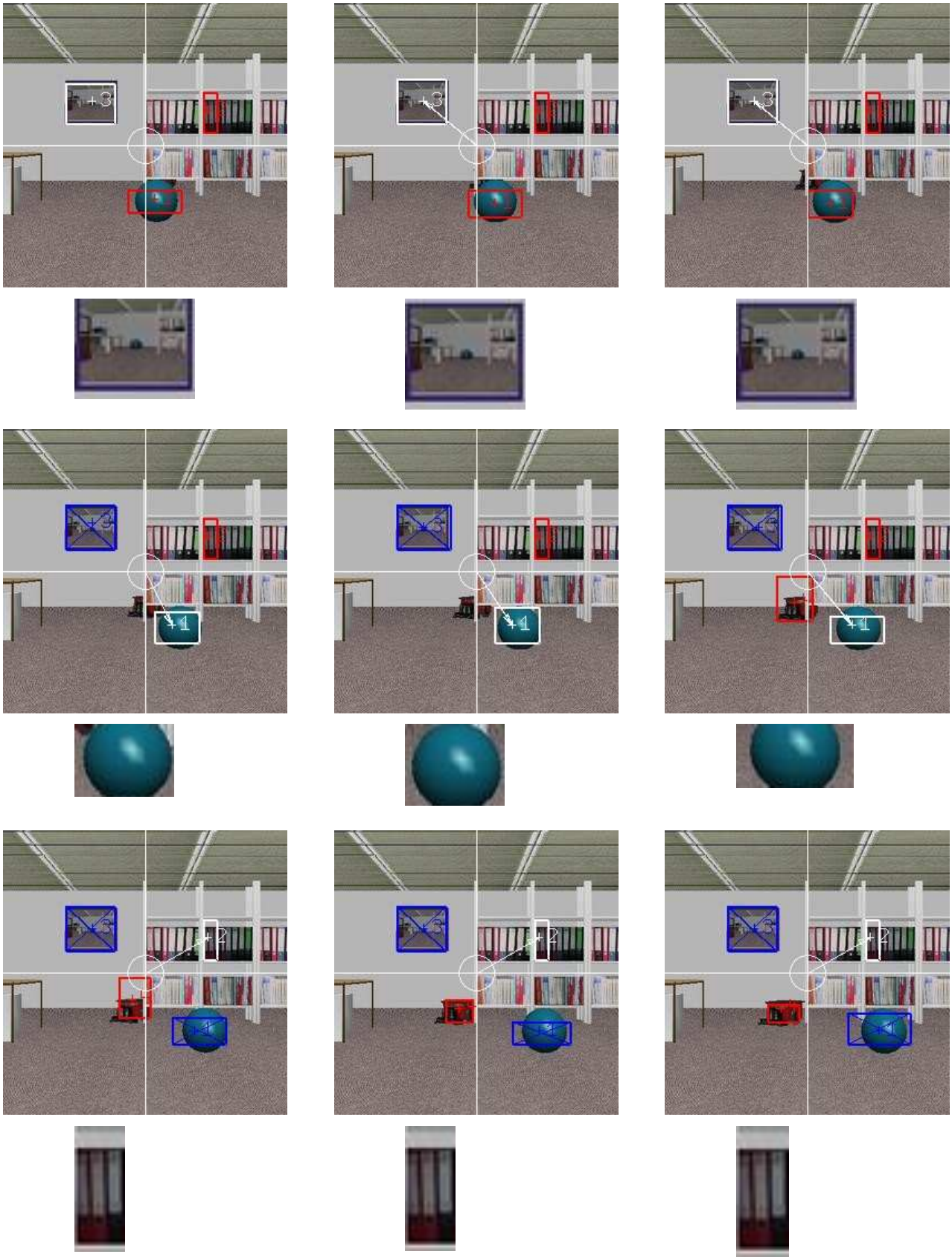


Abbildung 8.2: Demonstration des Verhaltens „Exploration" in einer Beispielszene (Teil 1). In Leserichtung (von links oben nach rechts unten) sind für die ersten 9 Frames die Eingabe und der jeweilige Bereich des FOA abgebildet. In den Eingabeframes sind die Bereiche der Objectfiles farbig und mit ihrer Identität markiert. Weiß steht für das momentan selektierte, rot für noch nie selektierte und blau für zuvor selektierte OF. Die Berechnungen der attentiven Stufe sind hier mit jeweils 3 Frames angesetzt.

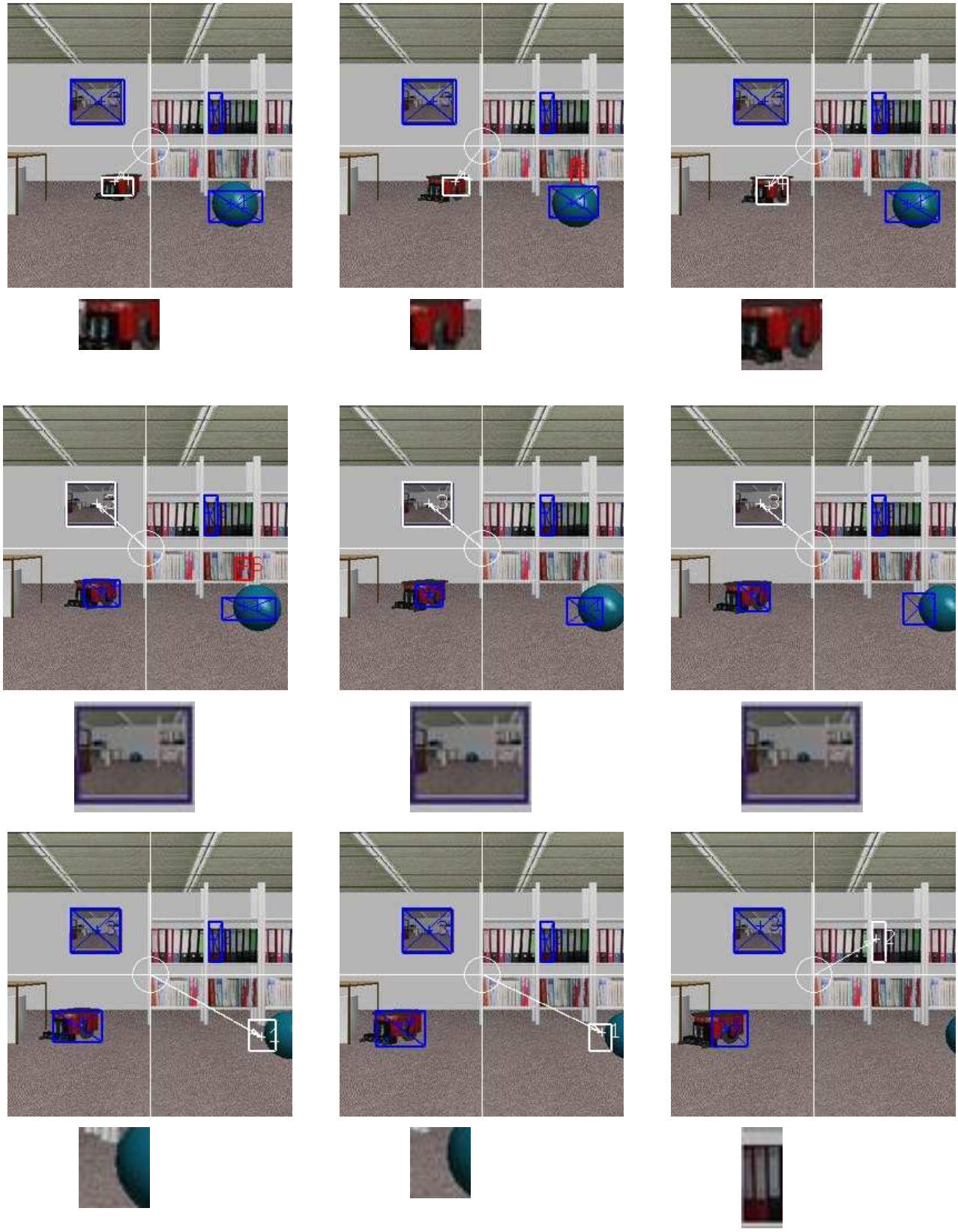


Abbildung 8.3: Fortsetzung von Abb. 8.2: Frames 10 bis 18. Erläuterung siehe dort.

deutlich zu erkennen, dass sich die Aktivitätscluster nicht auf beliebige Flächen, sondern auf relevante Objekte beziehen. Innerhalb der ersten zwölf Frames wurden alle Objekte jeweils für 3 Frames selektiert; erst jetzt erfolgt eine zyklische Selektion derselben Objekte. Ebenfalls deutlich wird die Verfolgung der bewegten Objekte (Ball und Roboter) durch die Aktivitätscluster. Beim letzten Frame erfolgt der Wechsel zu einem neuen Objekt, da das alte Objekt (der Ball) den Blickbereich verlassen hat.

8.5.2 Visuelle Suche

Die Visuelle Suche ist eines der zentralen Paradigmen der experimentellen Psychophysik zur Untersuchung von Aufmerksamkeitsmechanismen (s.a. 3.2.1). Es geht darum, möglichst schnell festzustellen, ob ein vordefiniertes Element im Display vorhanden ist oder nicht. Das Verhalten bezieht sich auf statische Displays, wie sie auch in den psychophysischen Experimenten dominieren.

Entscheidend für die Visuelle Suche ist die Bestimmung des Suchzieles, des Targets. Diese Spezifikation erfolgt anhand der Merkmale, die es von den Ablenkern unterscheiden. Die Erkennung wird auf eine Klassifikation als Target oder Distraktor reduziert. Die erste Merkmalsstufe wird so eingestellt, dass die Merkmale, die das Target auszeichnen, hoch gewichtet werden. Die übrigen Merkmale bleiben unverändert, um eine Lokalisation der Elemente zu vereinfachen, die zur Entscheidung, dass kein Zielreiz anwesend ist, verarbeitet werden müssen.

In der zweiten Selektionsstufe kommen alleine die Merkmale zum Tragen, die zur Unterscheidung von Zielreiz und Ablenker beitragen. Die OF werden anhand der Präsenz dieses Merkmals priorisiert. Wurde das ausgewählte OF als Distraktor klassifiziert, wird es in eine statische Inhibitionskarte eingetragen, um die Selektion eines weiteren Elementes zu ermöglichen.

Der Abbruch der Suche findet in einer Vereinfachung des natürlichen Vorbildes, die von Chun und Wolfe [CW96] ausführlich studiert wurde, statt, wenn eines von zwei Kriterien erfüllt ist. Zuerst wird die Ausprägung der gesuchten Merkmale in den Kandidaten ausgewertet. Liegt diese unter der Hälfte von bereits zurückgewiesenen Kandidaten, wird die Suche mit dem Ergebnis „abwesend" abgebrochen. Das andere Kriterium besteht in einem angepassten Schwellwert für die Anzahl der zu untersuchenden Distraktoren. Dieser Schwellwert hängt von der Rückmeldung nach der Reaktion ab, die die Korrektheit der Reaktion angibt. Der Schwellwert wird immer dann angepasst, wenn die Reaktion „abwesend" erfolgt. War die Reaktion richtig, wird der Schwellwert leicht gesenkt, war sie jedoch fehlerhaft, wird der Schwellwert deutlich erhöht. Dies führt zu einem Kompromiss hinsichtlich Geschwindigkeit und Fehlerrate, wie er auch von den Versuchspersonen erwartet wird (sogenannter *speed-accuracy-tradeoff*).

Die Parameter des Verhaltens bestehen in der Spezifikation der gesuchten Merkmale. Dazu wird für jede Merkmalsausprägung ein Wert zwischen 0 und 1 angegeben, der die Hinweiskraft dieser Ausprägung für die Unterscheidung von Zielreiz und Ablenker angibt. Dies setzt selbstverständlich voraus, dass sich die Objekte anhand der Merkmale unterscheiden lassen. Für das vorgestellte Modell sind Farbe, Tiefe und Orientierung als Merkmale geeignet. Weiterhin muss ein Entscheider existieren, der die endgültige Entscheidung treffen kann, ob es sich bei einem selektierten Objekt um ein Target handelt. Abb. 8.4 gibt den Ablauf dieses Verhaltensmodells wieder.

Das Verhalten wurde auf typische Reize aus Experimenten zur Visuellen Suche angewandt. Im ersten Beispiel wird ein rotes Element unter blauen Ablenkern gesucht (Merkmalssuche). Hier ergibt

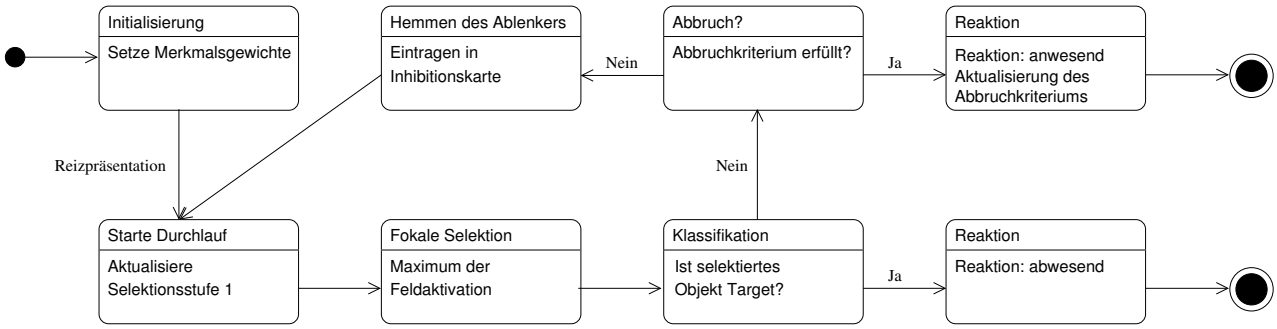


Abbildung 8.4: UML-Zustandsdiagramm zum Verhalten Visuelle Suche.

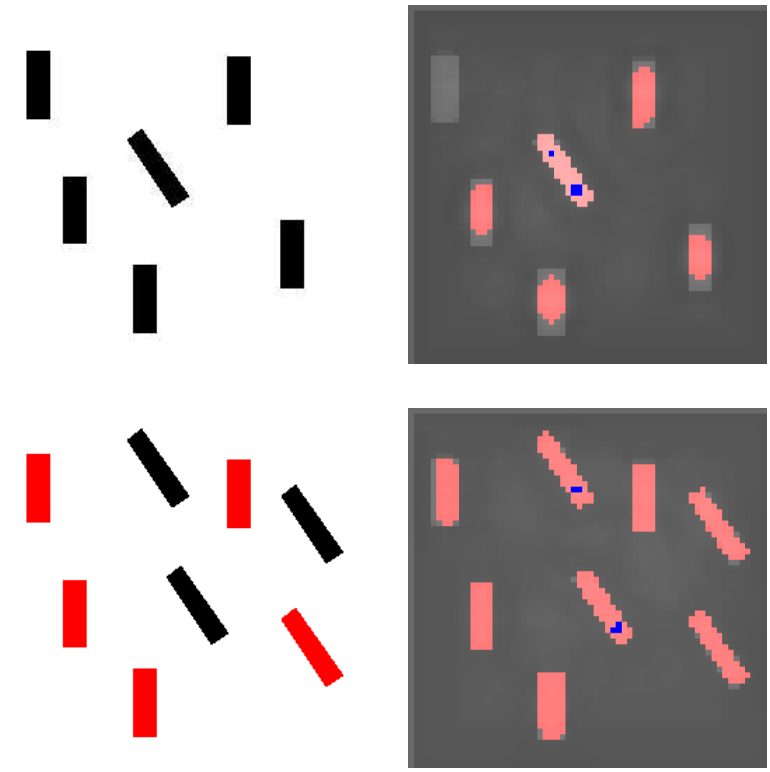


Abbildung 8.5: Zwei Beispiele zur Visuellen Suche: Merkmalssuche (oben) und Konjunktionssuche (unten). Links ist jeweils der Stimulus und rechts die Aktivation des Neuronalen Feldes wiedergegeben. Positive Aktivationen sind farblich rot hervorgehoben, der Maximalwert blau. Zielreiz ist im ersten Fall der schräg orientierte Balken, im zweiten Fall der schräg orientierte rote Balken.

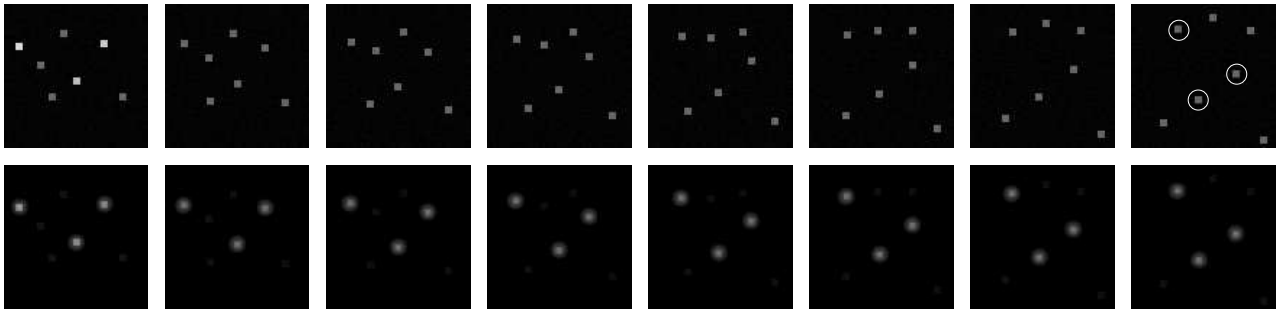


Abbildung 8.6: Durchführung eines Experimentes zum Multi Object Tracking. Es sind die Eingabe in das Neuronale Feld (oben) und die Aktivität des Feldes (unten) dargestellt. Im letzten Frame sind die Zielreize zusätzlich hervorgehoben.

sich der Zielreiz anhand der Gewichtung der Merkmale als auffälligstes Element und wird mit der höchsten Aktivation im Neuronalen Feld sofort selektiert. Im Gegensatz dazu handelt es sich im zweiten Beispiel um eine Konjunktionssuche nach einem roten, geneigten Zielreiz unter roten senkrechten und schwarzen geneigten Ablenkern (Konjunktionssuche). Hier ist es notwendig, die durch das Neuronale Feld selektierten Elemente seriell zu durchsuchen, um den Zielreiz zu finden. Vergleicht man dies mit den Ergebnissen der Psychophysik, wie sie in Kapitel 3.2.1 dargestellt wurden, ist dies genau der vom Menschen bekannte Effekt des Popout in der Merkmalssuche gegenüber einer langsamen seriellen Konjunktionssuche. In dieser entspricht das Verhalten also anderen Modelle, die sich in ihrem Design primär an der Visuellen Suche ausrichten.

8.5.3 Multi Object Tracking

In Anlehnung an die entsprechenden Experimente von Pylyshyn und Storm [PS88] ist dieses Verhaltensmodell angelegt, bei dem in einer Initialisierungsphase einige Elemente des Displays hervorgehoben werden. Es wird ein Detektor benötigt, der diese Hervorhebung auswertet und einmalig in der Salienzrepräsentation vermerkt. Da sich die Objekte in einem solchen Experiment gleichen, ist von gleichen Werten für die Salienzberechnung auszugehen, so dass durch die Hystereseeigenschaft der Neuronalen Felder eine stabile Selektion gewährleistet wird. Es wird keine fokale Selektion der Elemente benötigt, um während des Experimentes die Zielobjekte zu verfolgen. Spezifiziert werden muss hier die Detektion der Hervorhebung, in bisherigen Experimenten wird die Salienz der Reize bei der Initialisierung dazu gezielt erhöht.

Mit diesem Verhalten wurde ein Experiment nach Pylyshyn durchgeführt. Abb. 8.6 zeigt die Eingabe in das Neuronale Feld und die Aktivation des Feldes. Im ersten Frame wurden die Zielreize durch Verdoppelung ihrer Salienz ausgezeichnet. Von da an sind sie von den Distraktoren nicht mehr unterscheidbar. Im letzten Frame wurden sie im Nachhinein hervorgehoben, um den Vergleich zwischen verfolgten Elementen und Zielreizen zu ermöglichen. Es ist zu sehen, dass genau die Zielobjekte verfolgt werden, was die Leistung bei diesen Experimenten erklärt.

8.5.4 Search-and-track

Gibt es für das System nur ein wichtiges Objekt, das zu beobachten ist, setzt man ein Search-and-track-Verhalten ein. Es basiert auf der Visuellen Suche, enthält jedoch einige Modifikationen. Zuerst

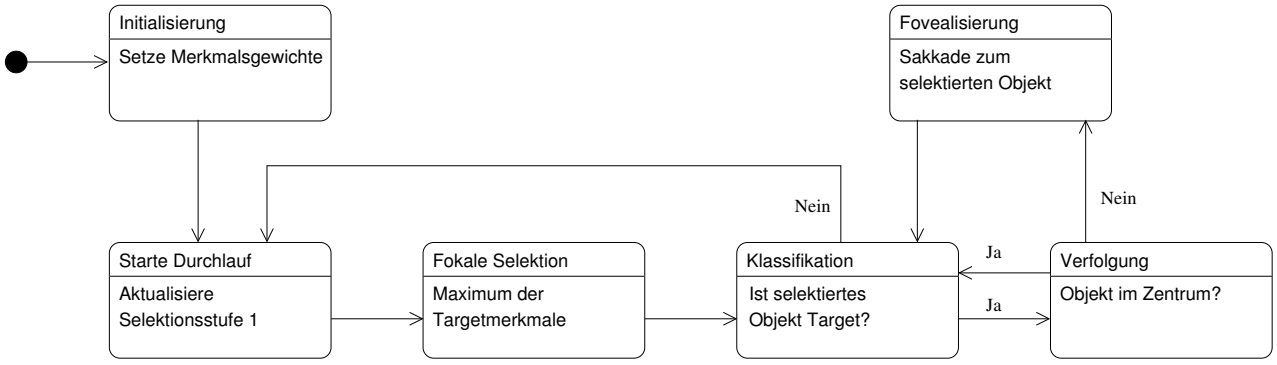


Abbildung 8.7: UML-Zustandsdiagramm zum Verhalten Search-and-track.

fallen die Abbruchkriterien weg, da nicht von einer statischen Eingabe auszugehen ist, sondern die Szene so lange beobachtet wird, bis das Zielobjekt gefunden wurde. Ist dies einmal geschehen, wird eine Sakkade zur Fovealisierung des Zielobjektes ausgelöst. Von nun an erfolgen immer abwechselnd die Verifikation, dass das ausgewählte OF tatsächlich dem gesuchten Objekt entspricht und Sakkaden zu diesem Objekt, sofern es sich aus dem Blickzentrum entfernt hat. Schlägt die Verifikation fehl, wird die Suche neu begonnen. Ist ein Bildbereich komplett abgesucht, was sich darin äußert, dass keines der Objectfiles dem Zielobjekt entspricht, wird eine zufällige Kamerabewegung ausgelöst, so dass die Umgebung untersucht werden kann.

Die festzulegenden Parameter entsprechen denen der Visuellen Suche: Merkmale, die zur Unterscheidung des Zielobjektes von anderen Objekten beitragen und ein Klassifikator für die endgültige Entscheidung. Das Verhaltensmodell ist in Abb. 8.7 dargestellt. Das Verhalten wurde im Experiment, das in Abb. 8.9 und 8.8 gezeigt wird, auf die Szene aus Abb. 8.2 angewandt. Dabei sollte der Roboter gesucht und verfolgt werden. Er zeichnet sich durch Symmetrie und vor allem seine Farbe aus und so wurden die Merkmale höher gewichtet. In den ersten vier Frames ist der Roboter noch nicht weit genug sichtbar und kann so nicht gefunden werden. Danach wird er jedoch gefunden und fixiert. Wann immer der Schwerpunkt des selektierten Aktivitätsclusters das Bildzentrum verlässt, wird die Kamera nachgeführt. Zu beachten ist, dass die Stabilität des OF unter Kamerabewegungen eine adäquate Anpassung der Aktivität in den Neuronalen Feldern voraussetzt.

8.5.5 Weitere Verhaltensweisen

Alarm-System

Im Gegensatz zur Exploration ist die Aufgabe des Alarmsystems, Veränderungen in der Umgebung festzustellen und sie an einen Klassifikator weiterzugeben, der abhängig von diesen Veränderungen Aktionen vollziehen kann. Es wird von einer Initialisierungsphase ausgegangen, in der sich das System an die statische Umgebung anpassen kann, um danach Veränderungen feststellen zu können. Die Umsetzung besteht in der Verwendung einer statischen Inhibitionskarte. Jedes OF, das über einen gewissen Zeitraum keine Bewegung des Schwerpunktes aufweist, wird in diese Karte eingetragen. Für alle anderen OF wird die Bewegung als 2D-Translation approximiert. Die OF werden anhand ihrer Entstehungsdaten in einer LIFO-Strategie selektiert, also immer das zuletzt erstellte Objectfile zuerst.

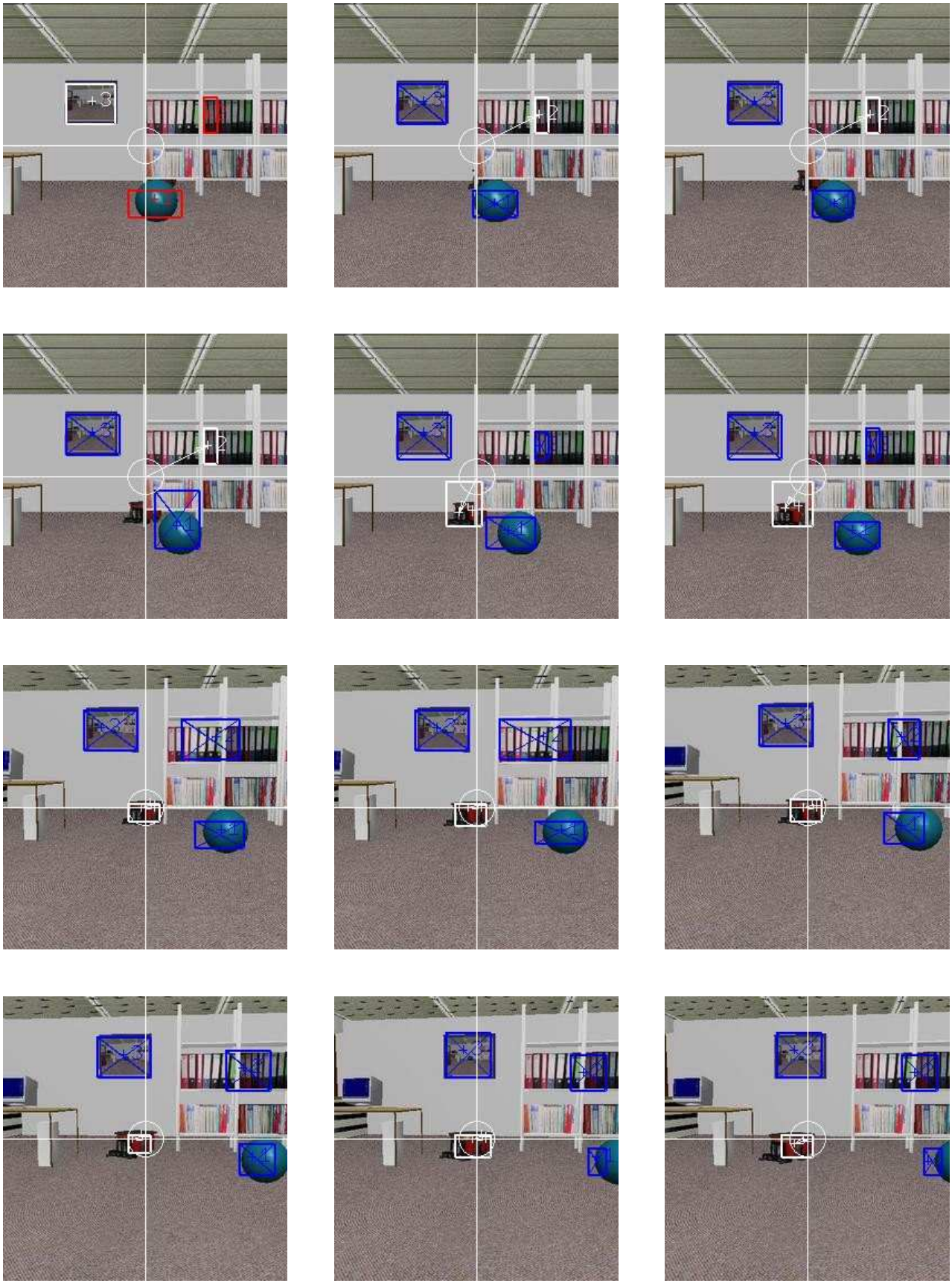


Abbildung 8.8: Demonstration des Verhaltens „Search-and-track" an einer Beispielszene (Teil 1: die ersten 12 von 24 Frames). Die OF sind im Kamerabild mit ihrer Bounding box und Identität annotiert: weiß bezeichnet das momentan selektierte, blau ein zuvor selektiertes und rot ein noch nicht selektiertes OF. Zusätzlich ist der Fixationsbereich durch einen weißen Kreis angegeben.

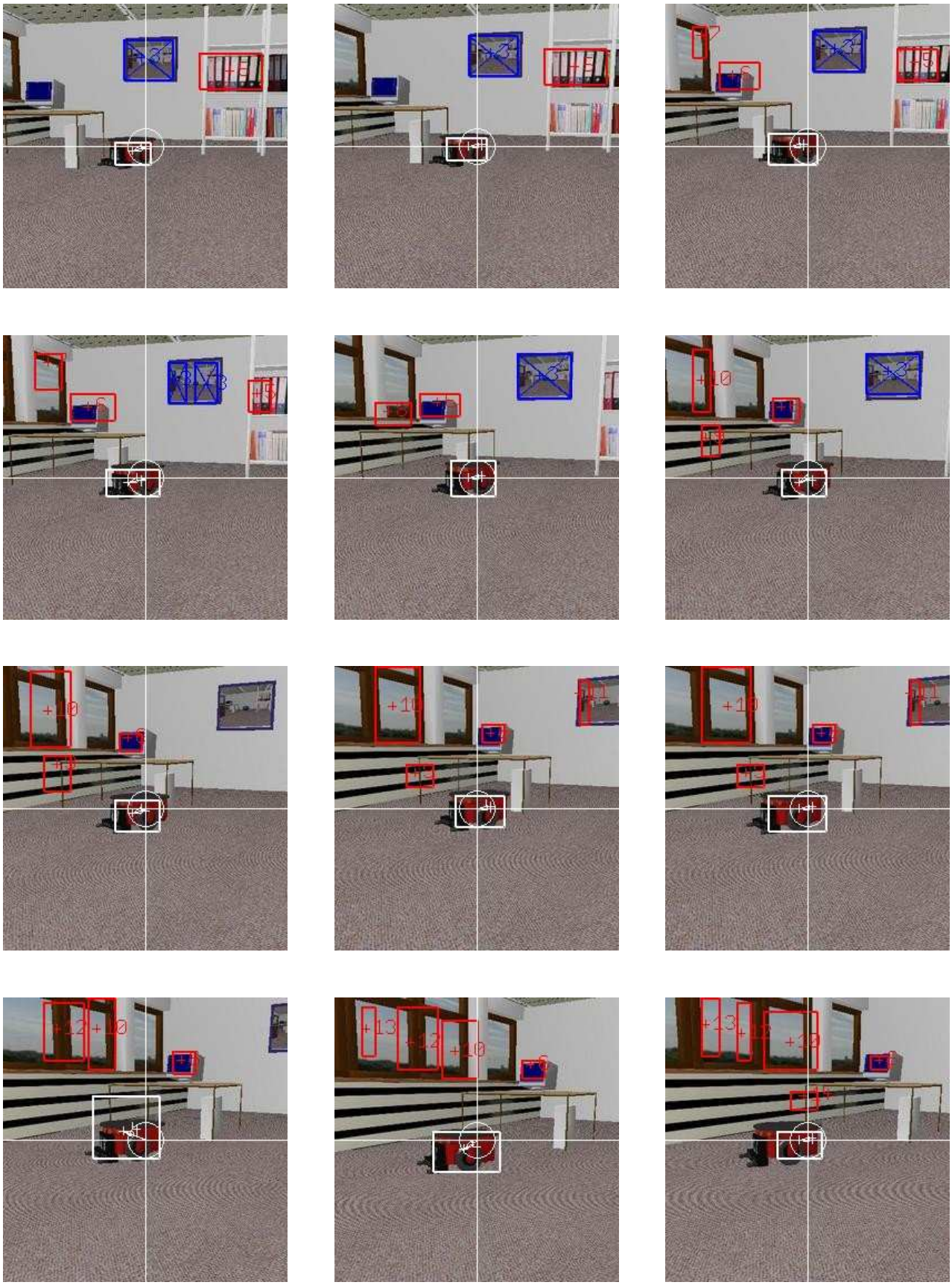


Abbildung 8.9: Demonstration des Verhaltens „Search-and-track" an einer Beispielszene (Teil 2: die letzten 12 von 24 Frames). Erläuterung siehe 8.8.

Eine höhere Priorität bekommen allerdings OF, bei denen die Bewegung des Schwerpunktes von der einmal bestimmten Approximation abweicht. Als Parameter dieses Verhaltens sind die Spezifikation der Klassifikation von Ereignissen und den mit ihnen assoziierten Aktionen anzugeben.

Cueing-Paradigma

Zur Modellierung von Cueing-Experimenten, wie sie Posner [PSD80, Pos80] vorgestellt hat, wird vergleichbar zum Multi object tracking ein Detektor von Hervorhebungen benötigt. Dieser wird genutzt, um seine Ergebnisse in eine zusätzliche Anregungskarte einzutragen, die komplementär zur statischen Inhibitionskarte arbeitet. Die Aktivation wird zur datengetriebenen Salienz addiert, vermindert sich aber im Laufe der Zeit ebenso wie die der Inhibitionskarte. Es sind auch Interpretationen der Hervorhebungen denkbar, etwa indem ein Hinweispfeil als solcher erkannt wird. Daraufhin würde nicht der Bereich, den der Hinweis einnimmt, sondern der, auf den er verweist, in die Anregungskarte eingetragen. Eine solche Erweiterung würde zum Einsatz des Systems in realen Umgebungen beitragen.

Identifikation von Zusammenhängen

Dieses Verhalten führt selbst keine fokale Selektion aus, es ergänzt andere Verhalten um die Bestimmung zusätzlicher Informationen. Es geht dabei darum, in der Bewegung selektierter Elemente eine Struktur zu finden. Dazu werden die Bewegungen der Schwerpunkte der Objectfiles analysiert. Es wird versucht, für Konstellationen aus mehreren OF eine gemeinsame Beschreibung der Bewegung zu finden, z.B. als reine Translation oder nur als Bewegung in der Bildebene. Diese Hypothesen werden dann in den folgenden Frames verifiziert.

Visuelle Routinen

Die Verhaltensmodelle weisen eine Beziehung zu den von Ullman [Ull84] vorgestellten *visual routines* auf. Auch wenn diese als kleine Einheiten zur Lösung definierter Aufgaben im Gegensatz zu Verhaltensweisen zur Steuerung des ganzen Systems gedacht waren, ist es leicht möglich, diese Verhaltensweisen aus solchen einfacheren Visuellen Routinen zusammenzusetzen. Hauptgrund dafür ist die einfache Datenstruktur der OF, die den indizierten Elementen in Ullmans Visuellen Routinen entsprechen.

Lernen

Als Ausblick soll eine Lernkomponente für die Verhaltensmodelle skizziert werden. Im Sinne der Architektur des Modells sollte eine Komponente an dieser Stelle symbolisch operieren. Dies legt die Verwendung Genetischer Programmierung [Koz92, Koz94] nahe, die durch genetische Operationen auf Programmstrukturen die Evolution dieser Strukturen anhand definierter Qualitätskriterien erlaubt. Diese Methodik wurde bereits erfolgreich auf Lernaufgaben [Bac96] und auch im Kontext der Bildverarbeitung [JMD94] angewandt. Das Aufmerksamkeitsmodell ist gerade deswegen für einen solchen Ansatz geeignet, weil die Datenstruktur eine einfache symbolische Manipulation erlaubt, die der Genetischen Programmierung entgegenkommt. Zur Lösung einer konkreten Lernaufgabe wären die OF als Datenstrukturen und einfache Operationen wie Vergleiche von Selektionsdaten, Positions-

und Merkmalsinformationen oder Zugriffe auf die Historie zu definieren und die Bestimmung eines OF als Resultat anzusetzen.

8.6 Zusammenfassung und Diskussion

Eine Vielzahl von technischen Aufgaben wie auch Modellierungen natürlichen Verhaltens in psychophysischen Experimenten lassen sich also durch Verhaltensmodelle nachbilden. Die vorgestellten Modelle mögen dabei als Anregung zur Umsetzung weiterer Verhalten dienen. Speziell mittels einer Kombination und Modifikation mehrerer der hier vorgestellten Modelle lassen sich komplexere Systeme konstruieren. Durch die Verwendung einer symbolischen Repräsentation als zu manipulierender Datenstruktur wird die Implementation der Verhaltensmodelle stark vereinfacht. Trotzdem gibt es natürlich Fälle, in denen auch die frühe Selektionsstufe modifiziert oder ergänzt werden muss. Grundsätzlich erlaubt die Architektur jedoch eine klare Trennung zwischen datengetriebenen und modellgetriebenen Einflüssen auf das Systemverhalten.

Teil III

Evaluation

Kapitel 9

Evaluation von Aufmerksamkeit

Zur Umsetzung eines Modells gehört auch immer die Überlegung, auf welche Art sich Eigenschaften, Angemessenheit und Leistungsfähigkeit des Modells bestimmen lassen. Im Kontext der Modellierung von Aufmerksamkeit gibt es dazu nicht viele Diskussionen, obwohl oder gerade weil die Aufmerksamkeit nicht trivial zu evaluieren ist. Das Grundproblem besteht darin, dass nicht klar ist, wie genau sich ein System aufmerksam verhalten soll, was *richtige* und *falsche* Zuweisung von Aufmerksamkeit ist. Nach einer Diskussion der Problematik werden die verschiedenen Möglichkeiten vorgestellt und mehrere von ihnen auf das vorgestellte System angewandt.

9.1 Möglichkeiten zur Evaluation von Aufmerksamkeitsmodellen

Da es nicht genau ein richtiges attentives Verhalten gibt, das ein System in einer bestimmten Umgebung zeigen sollte, müssen bestimmte Eigenschaften abgeleitet und analysiert werden. Außer dem Gesamtverhalten des Systems lassen sich auch Eigenschaften der Bestandteile oder Module messen und bewerten. Die erste Bewertung der Veranschaulichung des Verhaltens liegt in der subjektiven Analyse an unterschiedlichen Beispielen. Für viele Modelle ist dies die hauptsächliche Evaluationsmethode. Auch wenn die Einschränkungen offensichtlich sind - mangelnde Objektivität, Quantifizierbarkeit und Aussagekraft - ist dies ein relevanter erster Schritt. Wichtig ist dabei, möglichst unterschiedliche Aspekte zu beleuchten.

Jedoch sollte es keinesfalls bei diesem ersten Schritt bleiben. Gibt es Module mit definierten Aufgaben, ist zu überprüfen, ob diese Teilaufgaben einfacher zu bewerten sind als das Gesamtsystem. Im Falle der datengetriebenen Aufmerksamkeit trifft dies auf die Merkmale zu, die bestimmte Eigenschaften der Eingabebilder robust wiedergeben sollen. Die entsprechenden Experimente wurden für das vorgestellte System im Kontext der Entwicklung der Merkmale durchgeführt (s. dazu Kap. 5). Auch wenn diese Analyse keineswegs hinreichend ist, ein System als angemessen zu bewerten, so ist es doch notwendig, dass die Teilsysteme die ihnen zugewiesene Aufgabe angemessen erfüllen.

So wenig es richtige und falsche Aufmerksamkeitszuweisungen gibt, so naheliegend ist es doch, sich am natürlichen Vorbild zu orientieren. Auch für Systeme, die explizit nicht als Modelle natürlicher Aufmerksamkeit erstellt wurden, ist es doch so, dass die Herangehensweise zur selektiven Wahrnehmung ein natürliches Vorbild besitzt.

Für stärker technisch ausgerichtete Modelle gibt es häufig eine konkrete Anwendung, in der die Aufmerksamkeitssteuerung Verwendung finden soll. Es liegt nahe, die Qualität unterschiedlicher An-

sätze durch Bewertung der Leistungsfähigkeit für diese Anwendung zu testen. Dies ist jedoch für eine Bewertung der Qualität als Aufmerksamkeitsmodell insofern problematisch, als nicht bestimmbar ist, welchen Anteil die Aufmerksamkeitssteuerung an der Gesamtlösung hat und inwieweit Aufmerksamkeitsmodell und restliches System eben gerade aufeinander abgestimmt sind.

Schließlich gibt es einige allgemeine Eigenschaften, die sich für ein Aufmerksamkeitsmodell auswerten lassen und die im nächsten Abschnitt behandelt werden sollen.

9.2 Allgemeine Eigenschaften

Zur Bewertung von sogenannten *interest point detectors*, unter die im weiteren Sinne auch die Merkmale fallen, die zur datengetriebenen Berechnung von Salienz beitragen, schlugen Schmid et al. [SMB00] zwei Maße vor:

- die Wiederholbarkeit der Ergebnisse unter geometrischen Transformationen und
- den Informationsgehalt im Sinne von Entropie.

Die Wiederholbarkeit wird in ihrer Arbeit auf unterschiedliche Blickpunkte bezogen, hier leider nur als Wechsel des Szenenausschnitts mit Skalierung und 2D-Rotation, Variation der Helligkeit, jedoch nicht als Veränderung in einer dreidimensionalen Umgebung. Entsprechende Bewertungen wurden bei der Vorstellung der Merkmale in Kapitel 5 durchgeführt und zeigen die Qualität dieses Aspektes der Aufmerksamkeitssteuerung.

Die Anwendung ähnlicher Kriterien auf das bekannte Modell von Itti und Koch [IK00] durch Draper und Lionelle [LD03] zeigt, dass auch dieses Modell nicht immer die erwarteten Invarianzen bietet.

Der Informationsgehalt wird auch von Yamamoto et al. [YYL96] betont. Er verwendet den Gehalt der internen Repräsentation der Szene im Vergleich zur vollständigen Szene als Maß für die Effizienz eines Bildscans. Die Wichtigkeit bestimmter Bildbestandteile wird dabei ignoriert beziehungsweise durch ein subjektiv vorgegebenes Wichtigkeitsmaß bewertet. Eine in diesem Sinne optimale Effizienz erreichen Scanpaths, die Bereiche mit hoher Informationsdichte bevorzugen und eine möglichst breite Abdeckung des Bildes bieten. Insofern weicht das Maß deutlich vom Vorbild der natürlichen Aufmerksamkeit ab.

Ein erweitertes Maß der aktuell gültigen Informationen über eine dynamische Szene ist für dieses System aussagekräftiger. Dazu sollen Aufmerksamkeitssysteme in einer dynamischen Szene mit begrenzten Ressourcen Informationen über Objekte sammeln und jederzeit eine möglichst umfangreiche und akkurate Beschreibung der Szene anhand der Identität und der räumlichen Zuordnung der Objekte vorhalten.

Dazu wurde ein Experiment konzipiert, das auf einfachen Szenen mit einer kleinen Anzahl von statischen und dynamischen Objekten beruht. Die Objekte waren einfache Quadrate von 5 mal 5 Pixeln, die zueinander immer einen Abstand von mindestens 14 Pixeln aufwiesen. Die dynamischen Objekte bewegten sich zwischen zwei Frames um maximal 2 Pixel in x- und y-Richtung. Den Szenen war gleichverteiltes Rauschen mit 50 % der Objektamplitude überlagert. Abb. 9.1 zeigt einige aufeinanderfolgende Frames einer solchen Szene.

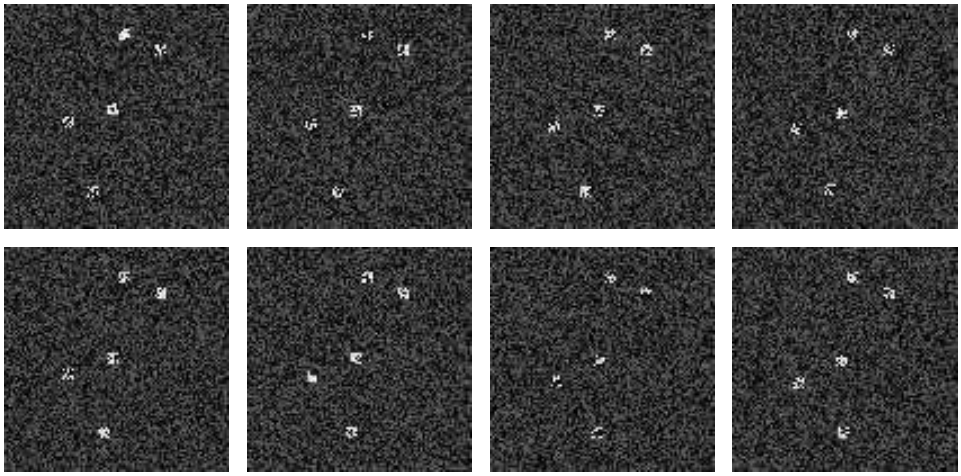


Abbildung 9.1: Acht aufeinanderfolgende Frames des Experimentes zur Exploration.

Diese Szenen sollten so exploriert werden, das dem System zu jedem Zeitpunkt die Identität und Position möglichst vieler Objekte zur Verfügung steht. Die Objekterkennung ist in diesem Falle simuliert; sie gibt zu jedem Pixel korrekt die Identität des Objektes an, das sich dort befindet und wurde mit einem konstanten Zeitaufwand von drei Frames belegt. Untersucht wurde, wieviele Objekte das System im Durchschnitt der Verarbeitung erkannt hatte und wie groß der Positionsfehler war. Bei Fehlern von mehr als 20 Pixeln wurde das Objekt als nicht erkannt klassifiziert, der Positionsfehler wurde dementsprechend nicht bewertet.

Verglichen wurde das vorgestellte Modell mit der Implementation eines klassischen Modells. Dessen Implementation orientiert sich an dem grundlegenden Modell von Koch und Ullman [KU85], das wiederum Grundlage zahlreicher anderer Modelle ist. Zur besseren Vergleichbarkeit der Selektionsmechanismen und zur Abstraktion der spezifischen Aspekte wurde auf die Merkmalsberechnung verzichtet. Vielmehr wurde aus der Eingabeszene eine für beide Modelle identische zentrale Salienzkarte erzeugt. Auf diese Salienzkarte wurde im klassischen Modell eine Maximumssuche angewandt, die aufgrund der Charakteristik der Eingabedaten grundsätzlich eines der Objekte fand. Das Ergebnis wurde dem Fokus der Aufmerksamkeit zugeordnet.

Am Ende der fokalen Bearbeitung, also der simulierten Objekterkennung, wurde das Objekt in einer Inhibitionskarte markiert. Mit dem Wissen um den Abstand der Objekte und ihre Bewegung wurde die Inhibition für dieses Experiment möglichst günstig gestaltet. Es wurde jeweils ein Bereich von acht Pixeln im Quadrat inhibiert, wobei die Inhibition nach jedem Frame um 20 % reduziert wurde. Dies erlaubt es dem Modell, eine lang andauernde Inhibition auch bewegter Objekte vorzunehmen ohne in die Gefahr zu geraten, fälschlicherweise ein Objekt zu inhibieren, das sich erst in diesen Bereich hineinbewegt. Die Identität des Objektes wurde mangels weiterer Hinweise an den Ort gebunden, an dem es erkannt wurde.

Für das vorgestellte Modell wurde die einfache Variante mit einem einzelnen zweidimensionalen Neuronalen Feld lokaler Inhibition gewählt. Obwohl die anderen Varianten bessere Verfolgungs- und Selektionsleistungen zeigen, ist dies die Variante, die hinsichtlich der Salienzrepräsentation am besten mit einem klassischen Modell verglichen werden kann. Um den Berechnungsaufwand der Neuronalen Felder mit in die Simulation eingehen zu lassen, wurde der Aufwand für die Objekterkennung nur

für dieses Modell um einen zusätzlichen Frame erhöht, so dass vier Frames für die Erkennung eines Objektes nötig waren.

Bewertet wurde, wie viele der in der Szene anwesenden Objekte zu jedem Zeitpunkt erkannt waren und welcher Fehler zwischen Schätzung und tatsächlicher Position bestand. Dabei wurde über alle Frames hinweg gemittelt, so dass die Verfahren niemals als Ergebnis die Gesamtzahl der vorhandenen Objekte erreichen konnte. Betrachtet man zum Beispiel die Bedingung mit fünf statischen und fünf dynamischen Objekten, benötigt das vorgestellte Modell alle 40 Frames, um alle Objekte zu erkennen. Als optimales Resultat wären also fünf erkannte Objekte im Durchschnitt der Verarbeitung zu erreichen, für das Standardmodell liegt der Wert aufgrund der schnelleren Objekterkennung bei 6,25.

Das Experiment wurde mit dem Verhaltensmodell "Exploration" durchgeführt. Jedem Datenpunkt liegen 50 Sequenzen zu je 40 Frames zugrunde. Abb. 9.2 zeigt sehr deutlich, dass das klassische System in Szenen ohne dynamische Objekte zwar von der schnelleren Objekterkennung profitiert, jedoch mit der Präsenz dynamischer Objekte in der Leistung sofort gegenüber dem vorgestellten Modell abfällt. Dieses skaliert gut mit der zusätzlichen Anzahl dynamischer Objekte. Hinsichtlich der Positionsfehler verhalten sich die Modelle sehr unterschiedlich. Während die Positionsschätzung des klassischen Modells in allen Bedingungen Fehler zwischen 0,5 Pixeln und 5 Pixeln macht, liegen die Fehler beim vorgestellten Modell immer unter 0,5 Pixeln.

Für das klassische Modell werden die Fehler durch die Präsenz statischer Objekte begrenzt. Sobald mehr dynamische als statische Objekte vorhanden sind, liegt der durchschnittliche Fehler bei mindestens 3 Pixeln. Es ist schließlich noch zu beachten, dass bei einem realen Einsatz der Modelle das Standardmodell mit einer eingeschränkteren Inhibition auskommen müsste, für das neue Modell jedoch leistungsfähigere Architekturen Neuronaler Felder zur Verfügung stehen. Insofern wird der Vorteil gegenüber dem Standardmodell im Experiment eher noch unterschätzt.

Rekapituliert man die in Kap. 3.3.1 erwähnten Forderungen, die Itti und Koch [IK01a] an die Modellierung von Aufmerksamkeit stellen, so ergibt sich für das vorgestellte Modell:

- **Die lokale Salienz ist kontextabhängig:** Durch die Einbeziehung von Merkmalen, die sich bereits auf potenzielle Objektstrukturen beziehen und eine Bewertung der Exklusivität bleibt die Salienzbestimmung nicht auf rein lokale Operationen beschränkt.

- **Eine zentrale topographische Karte akkumuliert die lokale datengetriebene Salienzinformation:** Eine solche Mastermap gibt es, jedoch wird sie in einer Modellvariante sogar auf eine dreidimensionale topographische Salienzrepräsentation erweitert.

- ***Inhibition of return* stellt einen zentralen Prozess dar:** Die IOR wird im Modell für dynamisch bewegte Objekte erweitert und erlaubt so eine IOR nicht nur für statische Orte, sondern auch für bewegte Objekte.

- **Starker Zusammenhang zwischen Augenbewegungen und verdeckter Aufmerksamkeit**: Die Zuweisung verdeckter Aufmerksamkeit stellt im Modell eine Voraussetzung für offene Aufmerksamkeit dar. Die Steuerung erfolgt für beide integriert in ein Verhaltensmodell.

- **Starker Einfluss der Objekterkennung auf die Zuweisung von Aufmerksamkeit:** Dieser Einfluss ist dem Verhaltensmodell und damit dem Interface zu weiteren Systemteilen, wie der Objekterkennung, überlassen.

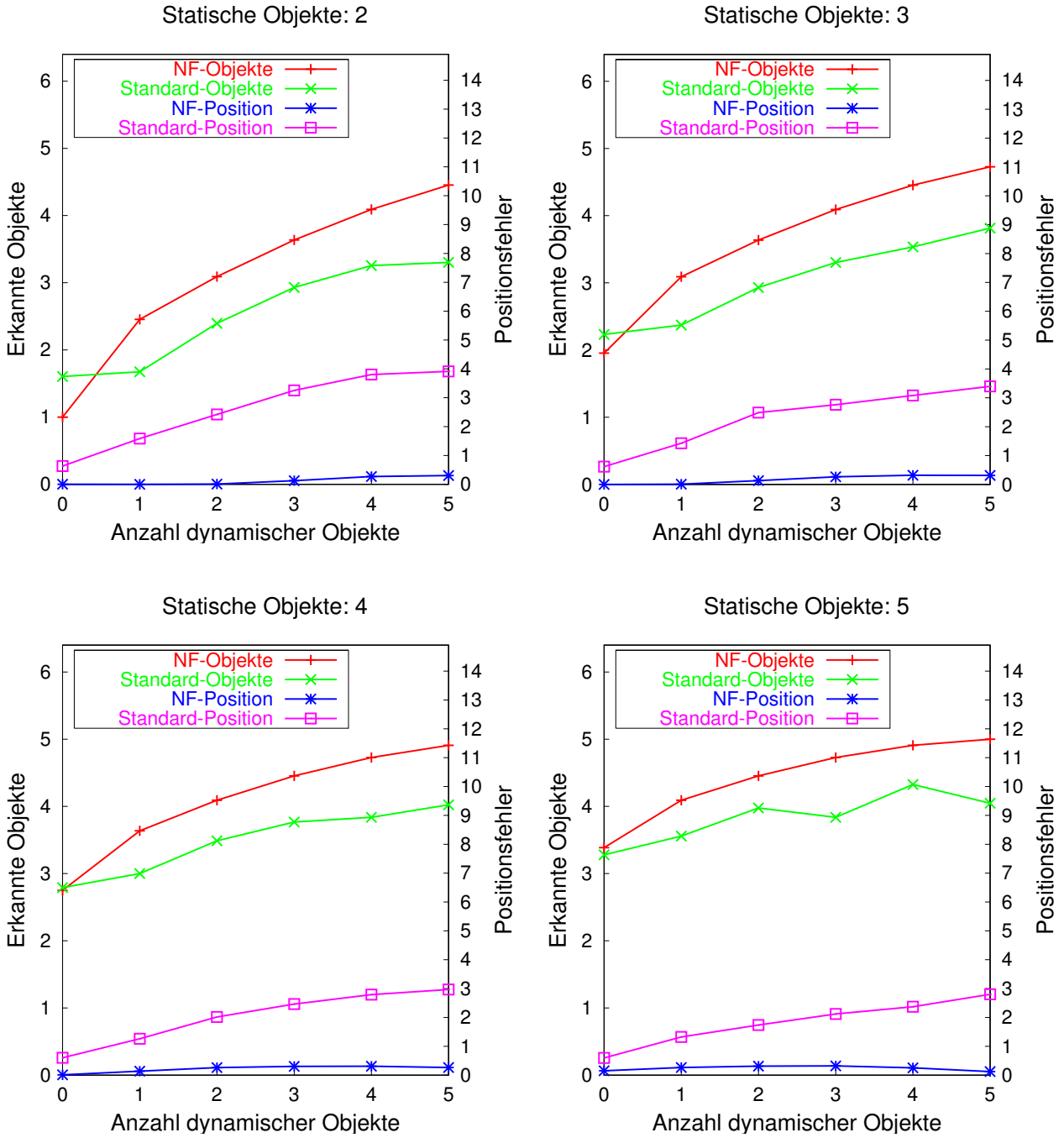


Abbildung 9.2: Vergleich des vorgestellten Modells (NF) mit einem Standardmodell (Standard) bezüglich des Wissens über die aktuelle Szene. Die vier Graphen geben für unterschiedliche Anzahlen statischer Objekte jeweils die von den beiden Modellen erreichten durchschnittlichen Anzahlen erkannter Objekte und den durchschnittlichen Positionsfehler in Abhängigkeit von der Anzahl dynamischer Objekte wieder.

9.3 Vergleich zum natürlichen Vorbild

Das Wissen über natürliche visuelle Aufmerksamkeit ist an vielen Stellen in das Design des Systems eingeflossen und auch dort überprüft worden. Dazu gehört die Definition der Merkmale, die mehrstufige Selektion und speziell die Zuweisung der Aufmerksamkeit an bewegte Objekte.

9.3.1 Diskussion der Angemessenheit

Es gibt ausführliche Diskussionen über Sinn oder Unsinn der Bionik, also der Forschung, die sich mit der Imitation von Lösungen der Natur durch die Technik befasst [Nac02]. Diese sollen hier zugunsten einer spezifischeren Diskussion der Aspekte für das konkrete Problem intelligenter Sehsysteme ausgelassen werden.

Gegner der Heranziehung natürlicher Vorbilder für das Computer-Sehen berufen sich darauf, dass die Implementationsbasis der technischen und der natürlichen Prozesse eine völlig andere ist und keinen Vergleich zulässt. Auch ist das Wissen über die Mechanismen der natürlichen Aufmerksamkeit noch keineswegs gefestigt, vielmehr werfen neuere experimentelle Ergebnisse und speziell auch Daten aus den Neurowissenschaften immer wieder Fragen auf, die keine einhelligen Antworten finden. Offen ist, welches Abstraktionsniveau bei einem Vergleich anzusetzen ist.

Auf der anderen Seite haben die Ingenieurswissenschaften bisher keine zufriedenstellenden technischen Lösungen für intelligente, flexible Sehsysteme erbracht. Die Effizienz, Robustheit und Leistungsfähigkeit der menschlichen Wahrnehmung demonstriert jedoch überzeugend, dass es mindestens eine Lösung gibt. Von den verwendeten Mechanismen zu lernen, Teile und Strukturen zu übernehmen, wenn man verstanden hat, welche Aufgabe sie erfüllen, scheint zumindest so lange naheliegend, wie keine bessere technische Lösung existiert.

Als Konsequenz ist eine Lösung nicht deswegen als schlecht zu bewerten, wenn sie nicht dem natürlichen Vorbild entspricht, die gestellte Aufgabe aber mindestens ebenso gut erfüllt. Ähnlichkeit zur Natur darf der technischen Leistungsfähigkeit nicht entgegenstehen. Doch gerade im Falle der visuellen Aufmerksamkeit, deren Modellierung der empirischen Untersuchung folgt, ist es angezeigt, das natürliche System als Maßstab anzusehen, solange es keine definierte technische Definition gibt.

Während die Ähnlichkeit zum natürlichen Vorbild für technische Systeme nicht das einzige Kriterium sein darf, so stellt es doch ein wichtiges Kriterium dar. Nicht zuletzt sollte beachtet werden, das sich mittlerweile ganze Konferenzen nur mit biologisch motivierten Ansätzen zum Computer-Sehen befassen [LBP00].

9.3.2 Flankerkompatibilitätseffekt

Ein wichtiger und sehr stabiler Effekt, der die Modellierung natürlicher Aufmerksamkeit, gerade hinsichtlich der Vorstellung vom Scheinwerfer der Aufmerksamkeit stark beeinflusst hat, ist der sogenannte Flankerkompatibilitätseffekt (s. 3.2.1). Er besteht im Einfluss von Ablenkern an bekanntermaßen irrelevanten Orten auf Klassifikationsaufgaben und legt nahe, dass auch bei räumlicher Fokussierung von Aufmerksamkeit Prozesse zur Identifikation für mehr als ein Element stattfinden. Wenige technische Modelle befassen sich mit seiner tatsächlichen Modellierung.

Betrachten wir das Verhalten des vorgestellten Aufmerksamkeitsmodells bei den typischerweise verwendeten Displays. Da nur eine kleine Anzahl von Elementen (Buchstaben) präsent ist, wird sich

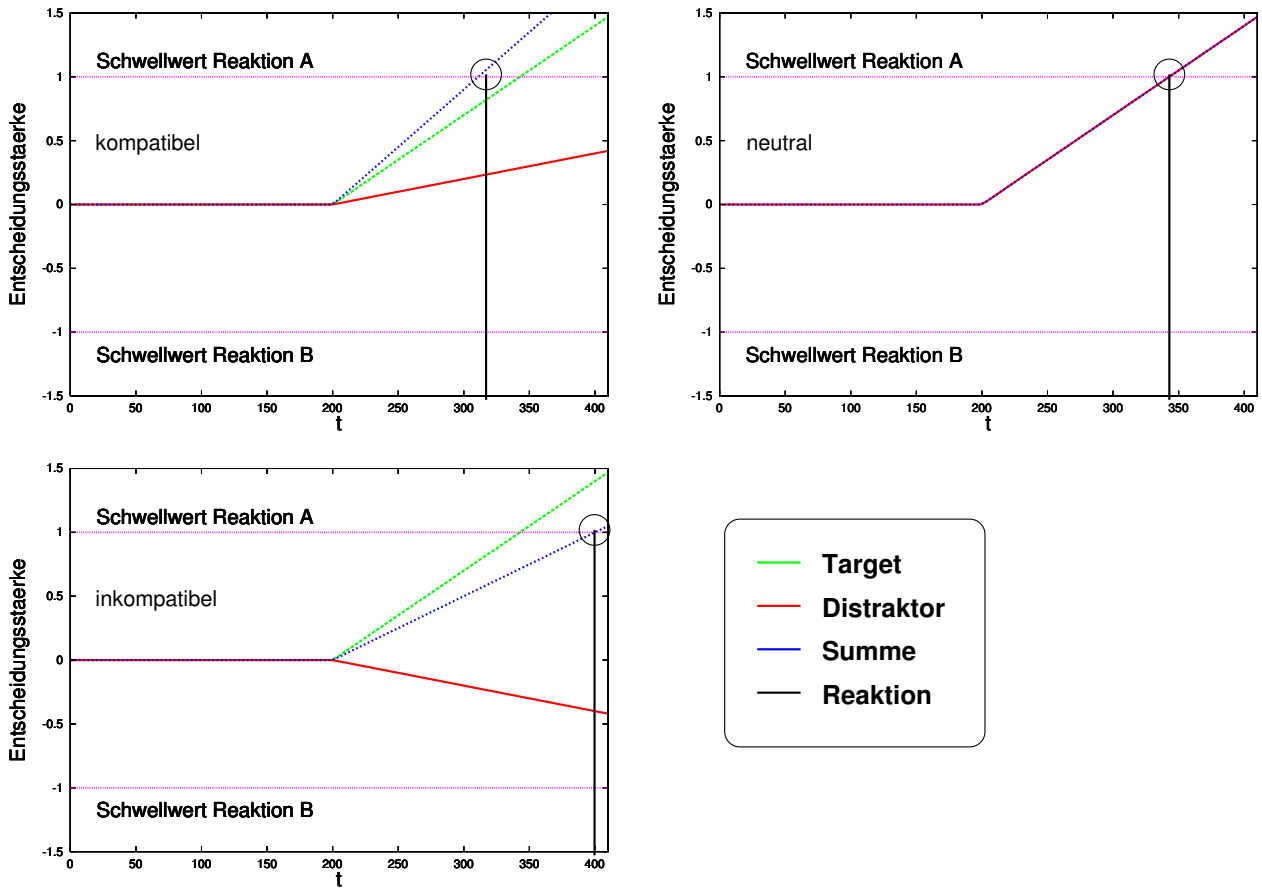


Abbildung 9.3: Effekte der Kompatibilität von Distraktoren auf die Reaktionszeit in Flankerkompatibilitätsexperimenten. Die drei Fälle unterscheiden die Darbietung eines zum Zielreiz kompatiblen, inkompatiblen bzw. neuralen Distraktors. Gezeigt wird die Entwicklung der Entscheidungsstärke für die möglichen Reaktionen in Abhängigkeit der Zeit. Dabei ist der Zeitpunkt markiert ist, an dem die Summe der Aktivationsstärke den Schwellenwert für die Auslösung der Reaktion erreicht.

für die meisten Buchstaben ein Aktivitätscluster und damit ein Objectfile bilden. Fokal wird aufgrund des bekannten Ortes das Zielelement selektiert und erkannt. Da jedoch die Identifikation von Buchstaben sehr stark automatisiert ist und sogar schwer zu unterdrücken ist, wie der Stroop-Effekt [Str35] demonstriert, kann man davon ausgehen, dass für eine solche Erkennung fokale Aufmerksamkeit keineswegs notwendig ist, sondern bereits die Auswahl einiger weniger Elemente durch die erste Selektionsstufe ausreicht.

Nimmt man nun Erkennungsprozesse an, die ein kontinuierliches Maß der Zugehörigkeit zu einer Klasse angeben und dieses Maß im Laufe der Berechnung stärker ausgeprägt wird, sind zwei Mechanismen denkbar, die zu einer Entscheidung führen. Einmal könnte im Falle unterschiedlicher Antworten der Klassifikatoren als zusätzlicher Prozess die Bindung der Klassifikation an einen Ort notwendig werden, oder man geht von einer Gewichtung der Klassifikationen in Abhängigkeit vom Ort und der Auslösung einer Reaktion bei Überschreiten eines Schwellwertes aus. Letzteres Modell wird als *bayesian observer* von Eckstein et al. [ESA02] auch erfolgreich zur Erklärung der Effekte in Posner's Cueing-Paradigma herangezogen.

In beiden Fällen ergibt sich eine schnellere Antwort im kompatiblen Fall als im inkompatiblen

Fall und einen mittleren Wert für die neutrale Bedingung. Für die zweite Variante ist dies in Abb. 9.3 dargestellt. Auch die Verringerung des Flankerkompatibilitätseffektes durch Erhöhung der Distanz wird so plausibel, da die Erkennung des Distraktors hinausgezögert bzw. seine Selektion unwahrscheinlicher wird. Ähnlichkeitseffekte, wie sie Baylis und Driver [BD92] zeigen, ergeben, wenn man annimmt, dass die fokale Selektion eines Objektes eine höhere Gewichtung der zugehörigen Merkmale mit sich bringt.

Es ergibt sich hier ein ähnlicher Prozess, wie ihn Chelazzi [Che99] für die Visuelle Suche annimmt: mehrere Elemente werden parallel selektiert und üben Einfluss auf die Reaktionsvorbereitung aus. Die Reaktion erfolgt erst, wenn sich ein Element gegenüber den anderen durchgesetzt hat oder alle dieselbe Reaktion aktivieren. Dies erklärt auch den Vorteil von Kompatibilität gegenüber einer neutralen Bedingung.

Für derartige Experimente lässt sich sogar eine Hypothese herleiten: bei einer größeren Anzahl von Distraktoren sollte man durch Veränderung der Salienz bestimmter Distraktoren eine Variation des Effektes erreichen, da ihre Selektion in der ersten Selektionsstufe damit mehr oder weniger wahrscheinlich wird.

9.3.3 Frühe und späte Selektion

Weniger als die genaue quantitative Modellierung einzelner Experimentalparadigmen ist die Beachtung unterschiedlicher grundsätzlicher Mechanismen der visuellen Aufmerksamkeit von Bedeutung. Zu diesen gehört die offensichtliche Flexibilität in der Stufe der Selektion, wie sie sich in der ungelösten Diskussion von früher und später Selektion darstellt.

Dieses Modell liefert eine alternative Erklärung der Flexibilität, indem zwei Stufen der Selektion eingesetzt werden, denen zwei Berechnungsstufen folgen, die Aufgaben unterschiedlicher Komplexität durchführen können. Diese zweifache Selektion führt zu einem Verhalten, das je nach Aufgabe und Belastung des Systems mal als frühe, mal als späte Selektion interpretiert werden kann. Ist die Aufgabe im Verhältnis zur momentanen Systembelastung einfach, kann sie auf alle Einheiten angewandt werden, die von der ersten Selektionsstufe ausgewählt wurden. Dies führt zu der Beobachtung, dass Aufgaben ohne Zuweisung fokaler Aufmerksamkeit parallel für mehrere Elemente ausgeführt werden können - ein Ergebnis, das die Theorien später Selektion bestärkt. Ist die Aufgabe dagegen im Verhältnis zur momentanen Systembelastung aufwändig, bleibt sie den Elementen vorbehalten, die durch die zweite Selektionsstufe für fokale Aufmerksamkeit ausgewählt wurden. Dieses lässt sich als frühe Selektion, bei der die Zuweisung fokaler Aufmerksamkeit Voraussetzung für die Ausführung wesentlicher Operationen ist, interpretieren.

9.3.4 Modellierung der Selektivität

Wichtig in der Diskussion der Selektion ist auch die Einheit der Selektion. Während der Ort als klassische und dominierende Einheit gilt, finden sich auch viele Hinweise auf die objektbeeinflusste Selektion. Dies wird im vorgestellten Modell auf zweierlei Weisen ausgedrückt [BM03, BMht]. Die Merkmalsberechnungen der präattentiven Stufe richten sich alle nach Eigenschaften aus, die als Hinweise auf Objekte oder Objektteile dienen. So werden homogene Segmente, Bereiche zusammengehörender Tiefe und Schwerpunkte von Formen mit gleichmäßiger Salienz ausgestattet. Durch die Verarbeitung in Neuronalen Feldern wird die Kontinuität betont, so dass Salienzen, die bewegten

Objekten zugeordnet ist, zusammengefasst wird. Somit sind die durch die erste Selektionsstufe ausgewählten Elemente gute Objektkandidaten. Auch die zweite Selektionsstufe operiert objektbasiert, bildet sie doch als symbolische Strukturen die Objectfiles und operiert auf diskreten Elementen. Besonders der Effekt der objektbasierten Inhibition of return wird hier erfolgreich modelliert.

Ein interessanter Zusammenhang ergibt sich zu den Ergebnissen von Luck und Vogel [LV97] bezüglich der Kapazität des visuellen Arbeitsgedächtnisses. Sie zeigen eine objektbasierte Kapazitätsgrenze, die bei vier Objekten (der Anzahl gleichzeitig aktiver Objectfiles) mit mindestens jeweils vier Merkmalen liegt. Das Modell geht hier aber wohl etwas über die Leistungsfähigkeit des Menschen hinaus, indem es eine Repräsentation mehrerer Objekte inklusive mehrerer Merkmale in dynamischen Szenen aufrecht erhält. Obwohl die einzelnen Aufgaben der Verfolgung mehrerer Objekte [PS88] und des objektbasierten Gedächtnisses mit mehrerer Merkmalen [LV97] lösbar sind, zeigte Saiki [Sai03] aktuell jedoch, dass dies für den Menschen so nicht möglich ist. Dies könnte jedoch auch auf mangelnde Ressourcen für eine dauernde Untersuchung aller verfolgten Objekte auf eine Änderung der Merkmale zurückzuführen sein.

Die neurobiologisch untersuchte Trennung in einen „Wo"- und einen „Was"-Pfad, also die Trennung von Positions- und Identitätsinformationen, schlägt sich im Modell in der Verwendung der Aktivitätscluster in den Neuronalen Feldern und den Objectfiles nieder. Während erstere allein eine Position markieren, enthalten letztere Identitätsinformation, die aber nur indirekt über den Verweis auf ein OF mit dem Ort verbunden ist.

Die Hinweise auf das Vorhandensein mehrerer räumlicher Foki bzw. die parallele Verarbeitung mehrerer visueller Objekte findet ebenfalls ihren Ausdruck in der zweistufigen Selektion, die eine Bearbeitung außerhalb der klassischen Dichotomie von präattentiver paralleler Verarbeitung und attentiver serieller Verarbeitung erlaubt. Im Unterschied zu anderen Modellen visueller Aufmerksamkeit zeigt das Verhalten gerade in dynamischen Szenen sowohl eine Verfolgung einer kleinen Anzahl von Objekten als auch eine Inhibition, die an sich bewegende Objekte gebunden ist. Andere Modelle orientieren sich hier allein an statischen Orten.

Schließlich wurden in Kapitel 8.5 für die experimentellen Paradigmen der Visuellen Suche und der Verfolgung mehrer Objekte Verhaltensmodelle erstellt, die dem natürlichen Vorbild gut entsprechen.

9.4 Einbindung in eine andere Anwendung

Um die Leistung eines Teilsystems zu beurteilen, ist es möglich, es im Kontext einer größeren Anwendung zu bewerten, indem die Leistung des Gesamtsystems als Indikator herangezogen wird. Während man dies einerseits als ultimativen Test der Nützlichkeit eines Systems ansehen kann, ist die Aussagefähigkeit andererseits aber sehr begrenzt. Es ist schwer möglich, den Anteil des Teilsystems an der Gesamtleistung zu bewerten. Unklar ist, inwieweit das System auf die Eigenschaften des zu bewertenden Teilsystems zugeschnitten wurde.

Im Rahmen dieser Arbeit wird aus den genannten Gründen auf eine Einbindung verzichtet, die jedoch eine interessante Weiterentwicklung des Systems darstellen würde. Stattdessen wurde jedoch deutlich gemacht, wie eine Einbindung funktionieren würde und welcher Art die Schnittstelle zwischen der Aufmerksamkeitssteuerung und sonstigen Systemmodulen aussehen würde.

9.5 Verwendung einer Simulationsumgebung zur Evaluation

Experimente mit aktiven Systemen leiden unter der mangelnden Reproduzierbarkeit und Parametrisierbarkeit der Eingabe. Der reale Einsatz in einer Umgebung ist aufwändig und erlaubt weder eine kontrollierte Modifikation von Umgebungseigenschaften noch eine exakte Replikation von Experimenten mit unterschiedlichen Systemparametern und -konfigurationen. Nur durch aufwändige Messungen erhält man *ground truth*-Daten, mit denen die Ergebnisse der Verfahren verglichen werden können. Computergenerierte statische Bilder hingegen sind zwar einfach zu erstellen und zu modifizieren, erlauben jedoch keine dynamische Veränderung und vor allem keine Aktion des Systems, wie eine Kamerabewegung. Die dynamische Veränderung wäre durch dreidimensionale Modellierung und Rendering entsprechender Bilder zwar noch zu erreichen, es fehlt jedoch weiterhin die Interaktion mit dem Kamerasystem. Dies ist allein durch eine Simulationsumgebung machbar, in der simulierte Sehsysteme aktiv sein können und von ihrer Aktivität abhängige Darstellungen einer dreidimensionalen Umgebung erhalten, verarbeiten und eventuell wieder in Aktionen umsetzen.

Ein Simulator kann verwendet werden, um alle zuvor aufgezählten Evaluationsmethoden zu unterstützen und zu erweitern:

- Die exemplarische Betrachtung wird vereinfacht, kontrollierbarer und vergleichbarer.
- Die Überprüfung allgemeiner Eigenschaften wird durch das Vorhandensein von *ground truth* verbessert.
- Der Vergleich zum menschlichen Vorbild lässt sich durch simulierte Experimente in dreidimensionalen Umgebungen ergänzen.
- Die Einbindung in andere Anwendung wird durch die Konfiguration komplexer Systeme aus Sensoren und Aktoren und das schnelle Modifizieren von Parametern erleichtert.

9.5.1 Simulationsumgebungen für Aktive Sehsysteme und Mobile Roboter

Obwohl dreidimensionale Simulationsumgebungen in vielen Bereichen verwendet werden, gibt es nur wenige Beispiele für Aktive Sehsysteme. Für den verwandten Bereich Mobiler Roboter finden sich zwar einige Simulationsumgebungen, diese beruhen jedoch meist auf einer zweidimensionalen Kartenrepräsentation der Umgebung, die nicht geeignet ist, Kamerabilder zur Verarbeitung durch das Sehsystem zu erzeugen. Beispiele für solche Systeme sind [Kon03, Mic96, Act03].

Dreidimensionale Umgebungen und das Rendering entsprechender Kamerabilder erlauben Terzoloulos' Animate Vision [Ter97] oder ein Fahrsimulator [SB98]. Während letzterer zu spezialisiert erscheint, stand der erste nicht als Software zur Verfügung. Das System von Matsumoto et al. [MMII99] verwendet spezialisierte Hardware, während der von Lu und Xie [LX00] vorgestellte Simulator aufgrund der beschränkten Konfigurationsmöglichkeiten nicht zum Einsatz kommen konnte. Allein das aktuelle System Breve [Kle02] scheint die notwendigen Qualitäten mitzubringen, setzt dabei jedoch einen anderen Schwerpunkt hinsichtlich der Modellierung einer größeren Anzahl von Agenten, deren Verhalten durch Skripte und weniger durch externe Applikationen gesteuert wird. Als Konsequenz wurde eine neue Umgebung mit der Bezeichnung Orbital 3D in der AG IMA umgesetzt[1], die im folgenden kurz beschrieben wird.

[1] Für die Implementation des Simulators sind die Herren Andreas Baudry und Michael Bungenstock verantwortlich.

9.5.2 Simulationsrahmenwerk Orbital 3D

Folgende primäre Anforderungen sind an eine Simulationsumgebung zu stellen, die geeignet sein soll, eine Vielfalt von Experimenten unterschiedlicher Art mit einem Aktiven Sehsystem durchzuführen, das sich in einer dreidimensionalen dynamischen Umgebung befindet:

- Die Simulation beruht auf kontrollierbaren, variierbaren dreidimensionalen Umgebungen mit dynamischen Elementen.
- Die simulierten Kameras sind in ihrer Anordnung modifizierbar und während der Simulation steuerbar.
- Das System erlaubt eine Skalierung der Qualität gegenüber der Rechenzeit.
- Die Schnittstelle soll den existierenden Systemen so weit wie möglich ähneln, um den Wechsel zwischen realer Umgebung und Simulation so einfach wie möglich zu gestalten.
- Die Verwendung der Simulationsumgebung sollte keine Einschränkungen für die Hardwareumgebung der Anwendung bedeuten.

Zusätzlich sollen modifizierbare Sensorkomponenten und Aktoren Verwendung finden, um langfristig die Simulation komplexer mobiler aktiver Systeme mit einer Vielzahl von Sensoren zu erlauben. Gleichzeitig erweitert dies die Einsatzmöglichkeiten des Simulators zur Verwendung für unterschiedliche Forschungs- und Lehraufgaben im Kontext des Aktiven Sehens und Mobiler Roboter.

Die beiden wichtigsten Designentscheidungen, die sich daraus ableiteten, waren zum einen die Konzeption einer Webserver-basierten Simulationsumgebung in Java, um den Simulator auf einem anderen Rechner ablaufen lassen zu können als die Anwendung und um eine einfache Kapselung der Schnittstelle zu erhalten. Zum anderen wurde ein komponentenbasierter Ansatz gewählt, der das Hinzufügen und Modifizieren von Sensor- und Aktorkomponenten erlaubt, ohne den Simulator selbst modifizieren zu müssen. Zu den Komponenten zählen auch die Systeme zum Rendern der Bilder, so dass es möglich wurde, die Simulation mit anspruchsvollen, aber langsamen Raytracingsystemen, aber auch mit einfachen, aber schnellen lokalen Beleuchtungsmodellen zu verwenden.

Das Modell ist über die Implementation von Java-Komponenten für Kameras und Sensoren zu erweitern. Die Umgebung wird in einer XML-Format beschrieben, die graphische Repräsentation der Objekte erfolgt wahlweise über POVRay- oder Java3D-Modelle. Abb. 9.4 gibt die Architektur des Simulationsrahmenwerks wieder.

Die Arbeiten von Baudry, Bungenstock, Bitterling und Mertsching [BBBM01, BBM02] stellen den Simulator genauer vor. Seine Verwendung für die Evaluation der Aufmerksamkeitssteuerung ist in [BM02a] beschrieben.

Verwendung von Orbital 3D

Der Simulator Orbital 3D wurde - ohne dass das explizit erwähnt wurde - für alle in dieser Arbeit gezeigten Experimente verwendet, die nicht anhand stark simplifizierten Bildmaterials (geometrische Formen) durchgeführt wurden. Er hat es erlaubt, die Umgebungsparameter zur Überprüfung der Merkmalseigenschaften zu kontrollieren und Experimente mit bewegten Kameras wie in Abb. 8.7 durchzuführen.

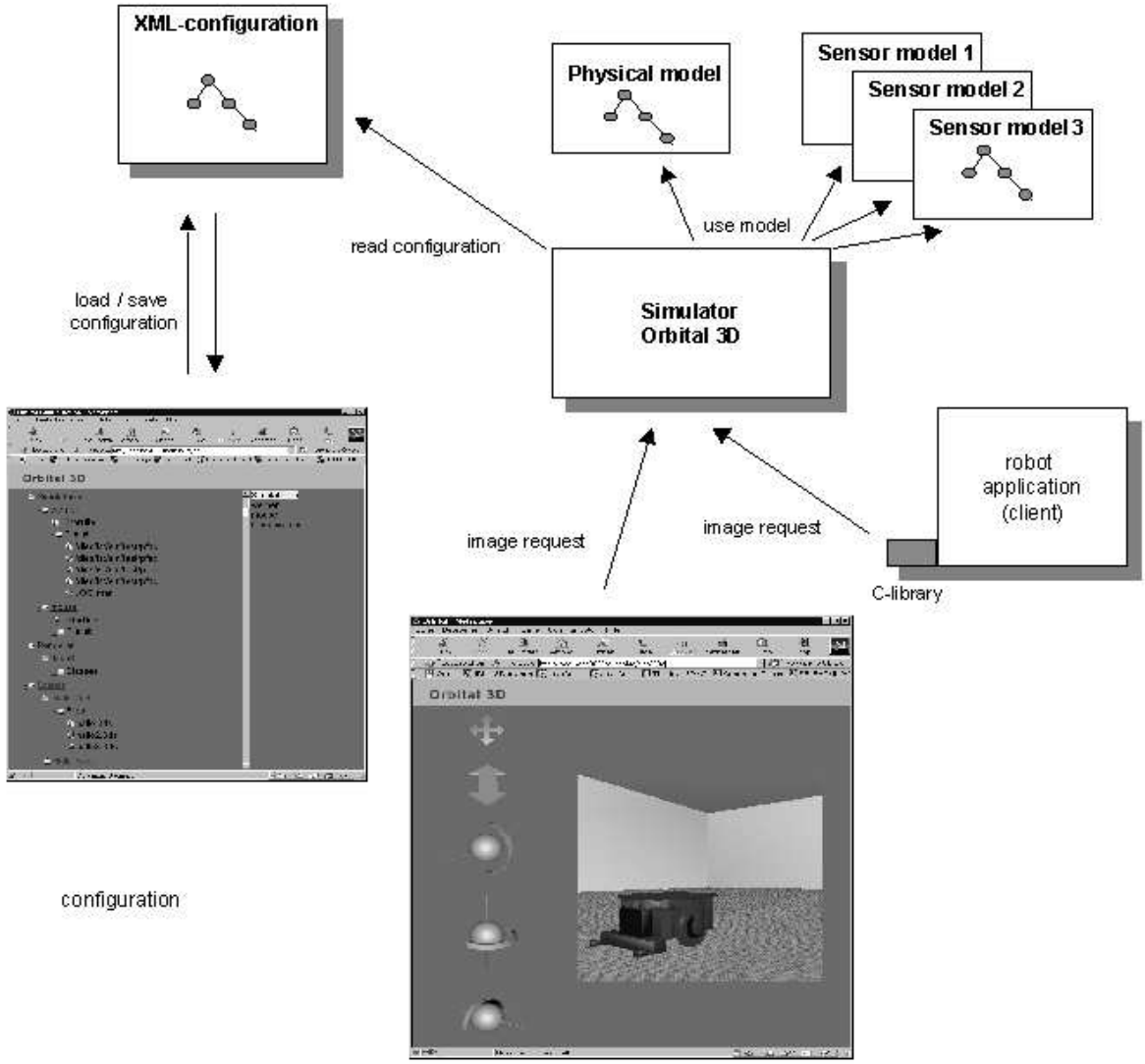


Abbildung 9.4: Verwendung des Simulationsrahmenwerks Orbital 3D.

Langfristig könnte ein solches Simulationsrahmenwerk die Basis für einen Benchmark für Aufmerksamkeitsmodelle darstellen. Durch reproduzierbare und vergleichbare Experimente mit Umgebungen ganz unterschiedlicher Qualitäten wäre die Beobachtung und Analyse verschiedener Aufmerksamkeitsmodelle möglich. Die Umgebungen und Aufgaben könnten ein ganzes Spektrum von der Simulation einfacher psychophysischer Experimente bis hin zu konkreten technischen Aufgaben in simulierten Realweltumgebungen abdecken. Durch die Modifizierbarkeit der Sensoren und Aktoren wäre es möglich, Systeme zu vergleichen, die große technische Unterschiede aufweisen, etwa Stereosysteme gegenüber solchen mit einzelnen Kameras.

Kapitel 10

Zusammenfassung und Ausblick

Das Ziel der Arbeit bestand in der Entwicklung eines Modells zur visuellen Aufmerksamkeit, das als Bestandteil zur Verwendung in einem Aktiven Sehsystem geeignet wäre. Besondere Beachtung sollte dabei die Komplexität der Umgebung erhalten, so dass sich das System den drei räumlichen Dimensionen und der Dynamik der Umgebung angepasst darstellt und objektbasierte Mechanismen verwendet.

Objektbasierte Ansätze und die Berücksichtigung von Tiefeninformation fanden gleich im ersten Teil des Modells - der Berechnung lokaler Salienz - Beachtung. Die umgesetzten Merkmale gehen über die klassischen Filteroperationen vieler Modelle (Kapitel 4.1.2) hinaus und trotz des in dieser Stufe begrenzten Aufwandes wird die Auffälligkeit so berechnet, dass sie sich möglichst auf visuelle Objekte oder Objektteile bezieht. Dazu werden verschiedene Cues (Kanten- und Flächeninformationen, Farbe und Stereodisparität) ausgenutzt, um eine Vielfalt von Objekten unter unterschiedlichen Bedingungen zu bewerten.

Die Geschwindigkeit und Leistungsfähigkeit wird unter anderem durch Verwendung von Multiskalenansätzen erreicht. Charakteristika und Qualität der Verfahren wurden durch umfangreiche Experimente nachgewiesen. Zur Integration der Merkmale wurden verschiedene Verfahren diskutiert und in Zusammenhang mit dem Entwurf der ersten Selektionsstufe bewertet. Eine der Integrationsmöglichkeiten verwendet Tiefeninformationen, was die Integration und Lokalisierung von Salienz im dreidimensionalen Raum erlaubt. Eine solche Repräsentation weisen nur die Modelle von Maki [MUE96, Mak96], Braumann [Bra01] und Ouerhani und Hügli [OH00] auf, die sich jedoch hinsichtlich der allgemeinen Modellierung von Aufmerksamkeit als weitaus unvollständiger darstellen.

Die Aufteilung der Selektion in zwei Stufen ist ein vollständig neuer Ansatz, der entsprechend umfangreich begründet wurde. Er leitet sich gleichzeitig aus der Beachtung des natürlichen Vorbildes und aus Erfordernissen des technischen Systems ab. Speziell die dynamische Natur der Umgebung mit potenziell mehreren sich bewegenden Objekten stellte eine Herausforderung dar, die sich durch klassische Verfahren nicht lösen ließ. Sowohl die Selektion und Inhibition von Objekten als auch die notwendige Aktualisierung des Weltmodells konnten nicht befriedigend mit bestehenden Systemen gewährleistet werden. Die Aufteilung der Selektion in zwei Stufen jedoch, die eine zusätzliche Berechnungsstufe zwischen rein präattentiver paralleler Berechnung und rein attentiver serieller Berechnung mit sich bringt, erlaubt eine Lösung dieser Probleme. Gegenüber anderen Aufmerksamkeitsmodellen, die eine Anwendung auf dynamische Szenen beinhalten, wie [Kop96, BBC+97, MNE00], zeichnet sich das vorgestellte Modell durch eine Berücksichtigung der Umgebungsdynamik nicht nur für die

Merkmalsberechnung oder Formung des Fokus, sondern einen angepassten Selektionsmechanismus aus.

Die erste Selektionsstufe wählt dabei anhand der datengetriebenen Berechnung lokaler Salienz subsymbolisch eine kleine Anzahl von auffälligen Elementen aus einer geeigneten Repräsentation aus. Diese Auswahl sollte robust mit einer räumlich-zeitlichen Integration stattfinden. Weiterhin war eine Verfolgung der ausgewählten Bereiche hoher Salienz notwendig. Als Modell, das diese beiden Aufgaben in sich vereint, boten sich Dynamische Neuronale Felder nach Amari [Ama77] an. Zur Anpassung an die Salienzrepräsentation waren Modifikationen notwendig, die zu neuen Architekturen der Neuronalen Felder führten. Die Eignung dieser Strukturen, zu denen auch ein dreidimensionales Neuronales Feld gehört, für robuste Selektion und modellfreie Verfolgung wurde experimentell nachgewiesen. Auf diese Weise konnten die wichtigen Aufgaben der ersten Selektionsstufe und der zusätzlichen Berechnungsstufe erfolgreich gelöst und integriert werden.

Das Wissen der Aufmerksamkeitssteuerung über die Umgebung wurde wesentlich durch sogenannte Objectfiles modelliert, die von Modellen der natürlichen Aufmerksamkeit inspiriert sind. Die Objectfiles stellen die erste symbolische Repräsentation der selektierten Elemente dar und erlauben so eine einfachere Manipulation. In Zusammenhang mit den Verfolgungseigenschaften der Neuronalen Felder und Prozessen zur Erhaltung der Korrespondenz von Objectfile und Aktivitätscluster wird so unter den Bedingungen serialisierter Objekterkennung und dynamischer Umgebung ein möglichst aktuelles Weltmodell gesichert, das nicht nur die Identitäten der wichtigsten Objekte, sondern auch ihre aktuellen Orte enthält. Die mit ressourcenintensiven Prozessen extrahierten Informationen lassen sich so besser an bewegte Objekte binden und bleiben länger gültig.

In der zweiten Selektionsstufe, die nun auf rein symbolischer Ebene operiert, hat die klassische Selektion eines einzelnen Fokus der Aufmerksamkeit für die Anwendung komplexer Operationen zu erfolgen. Auch die Fovealisierung eines Objektes durch Ansteuerung von Kameras liegt in der Verantwortung dieser Stufe. Die Steuerung durch ein Verhaltensmodell erlaubt die einfache Konfiguration des Systems für verschiedene Aufgaben. An dieser Stelle wird der modellgetriebene Einfluss auf das System gebündelt und eine einfache Schnittstelle für die Einbindung in komplexere Systeme zur Verfügung gestellt. Mehrere Beispiele für solche Verhaltensmodelle aus den Bereichen der Modellierung natürlicher visueller Aufmerksamkeit und aus praktischen Anwendungen des Systems wurden entwickelt und demonstriert. Die Implementation einer objektbasierten Inhibition bewegter Objekte innerhalb der Verhaltensmodelle stellte dabei einen weiteren Schritt hin zu objektbasierter Aufmerksamkeit dar.

Die beiden Kriterien, mit denen sich ein Modell visueller Aufmerksamkeit messen und bewerten lässt, sind die Ähnlichkeit zum natürlichen Vorbild und die Effizienz und Leistungsfähigkeit als technisches System. Beide wurden diskutiert und wegen der erkannten Problematik in der Evaluation von visueller Aufmerksamkeit wurden Möglichkeiten entwickelt, diese Evaluation umfassend und gründlich zu gestalten. Dazu gehört auch die Verwendung eines Simulationsrahmenwerks, das die kontrollierbare und reproduzierbare Ausführung von Experimenten in dynamischen dreidimensionalen Umgebungen erlaubt.

Als Modell natürlicher Aufmerksamkeit zeichnet sich die zweistufige Selektion besonders in der Beachtung von Aspekten aus, die über das einfache Scheinwerfermodell hinausgehen. Dazu gehören die objektbasierte Selektion und Inhibition, die man zusätzlich zu rein räumlichen Aspekten findet, die

gleichzeitige Verarbeitung und Verfolgung mehrerer Objekte und die Modellierung des Einflusses von Distraktoren. Auch zur klassischen Diskussion von früher und später Selektion konnte ein innovativer Beitrag geleistet werden.

Das Modell bietet so eine deutliche Alternative zu den vielen Modellen, die in ihrer Architektur dem Aufmerksamkeitssystem von Itti und Koch [IKN98, IK00] entsprechen, ohne deswegen Kompromisse hinsichtlich der biologischen Plausibilität, der Vollständigkeit als Aufmerksamkeitsmodell oder der technischen Verwendbarkeit zu machen.

Auf die diskutierte Einbindung des Aufmerksamkeitsmodells in ein Aktives Sehsystem wurde zugunsten der ausführlichen Analyse der neuartigen Struktur verzichtet. Eine solche Einbindung, die wegen der definierten einfachen Schnittstelle und der Möglichkeit zur Verwendung der unterschiedlichen Verhaltensmodelle ohne wesentliche Anpassungen der Aufmerksamkeitssteuerung stattfinden kann, würde den nächsten wichtigen Schritt in der Etablierung des Modells darstellen. Damit einhergehen würde die Entwicklung zusätzlicher Merkmalsberechnungen - gerade Merkmale, die auf der Veränderung der Umgebung beruhen, würden eine Bereicherung des Modelles darstellen. Die Verwendung in einem komplexeren System könnte auch die Erstellung spezifischerer Verhaltensmodelle und die Integration mehrerer vorgestellter Modelle in ein System aus mehreren Verhalten, zwischen denen ein flexibler, der aktuellen Situation angepasster Wechsel stattfindet, mit sich bringen.

Weiteres Entwicklungspotenzial ist in der attentiven Segmentierung von Objekten zu sehen. Hier wäre zuerst die Bildung von Segmenten anhand der Aktivitätscluster in den Neuronalen Feldern zu nennen, die unter Verwendung der Featureinformationen und der Historie des Segmentes beste Voraussetzungen mitbringt, mit begrenztem Aufwand eine gute Segmentierung für die auffälligsten Objekte der Szene vorzunehmen. In Anlehnung an die Unterscheidung von präattentivem Clustering und attentiver Segmentierung nach Trick und Enns [TE97] könnte zusätzlich noch eine aufwändigere Segmentierung als attentiver Prozess implementiert werden.

Schließlich wäre eine Echtzeitimplementierung des Systems von Interesse, die den Rechenaufwand je Eingabebild von mehreren Sekunden in der derzeitigen Implementation so stark reduziert, dass eine Interaktion mit einer dynamischen Umgebung möglich wird. Dazu kann es nötig sein, die Neuronalen Felder durch ein klassisches Verfolgungsverfahren ergänzt um ein geeignetes Selektionsmodul zu ersetzen.

Literaturverzeichnis

[Act03] ActiveMedia Robotics, Basic Suite. http://www.amigobot.com/. 2003

[AF01a] Arbel, T. ; Ferrie, F.P.: Entropy-based gaze planning. In: *Image and Vision Computing* 19 (2001), S. 779–786

[AF01b] Arbel, T. ; Ferrie, F.P.: On The Sequential Accumulation of Evidence. In: *International Journal of Computer Vision* 43 (2001), Nr. 3, S. 205–230

[Ahm91] Ahmad, S.: *VISIT: An efficient compuational model of human visual attention*, University of Illinois, Dissertation, 1991

[Ahr00] Ahrns, I.: *Ortsvariantes aktives Sehen für die partielle Tiefenrekonstruktion*, Universität Ulm, Dissertation, 2000

[All90] Allport, A.: Visual attention. In: Posner, M.I. (Hrsg.): *Foundations of Cognitive Science*. Cambridge, MA : MIT Press, 1990

[Alo93] Aloimonos, Y.: Active vision revisited. In: Aloimonos, Y. (Hrsg.): *Active perception*. 1993

[Alo94] Aloimonos, Y.: What I have learned. In: *Computer Vision, Graphics, and Image Processing* 60 (1994), Nr. 1, S. 77–87

[Ama77] Amari, S.-I.: Dynamics of pattern formation in lateral inhibition type neural field. In: *Biological Cybernetics* 27 (1977), S. 77–87

[AN99] Ahrns, I. ; Neumann, H.: Space-Variant Dynamic Neural Fields for Visual Attention. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition (ICPR)*, 1999, S. 313–318

[AO91] Ahmad, S. ; Omohundro, S.: Efficient visual search: A connectionist solution. In: *Proceddings of the 13th Annual Conference of the Cognitive Science Society*, 1991, S. 293–298

[Ask03] Ask A Biologist. http://askabiologist.asu.edu/. 2003

[AWB87] Aloimonos, Y. ; Weiss, I. ; Bandopadhay, A.: Active vision. In: *Proceedings of the first International Conference on Computer Vision*, 1987, S. 35–54

[Bac96] Backer, G.: Learning with missing data using Genetic Programming. In: *The 1st Online Workshop on Soft Computing (WSC1).* http://www.bioele.nuee.nagoya-u.ac.jp/wsc1/ : Nagoya University, Japan, 8 1996

[Bac98] Backer, G.: Attentional processes in Computer Vision on the basis of neural networks. In: Nauck, D. (Hrsg.) ; Krell, G. (Hrsg.) ; Kruse, R. (Hrsg.) ; Michaelis, B. (Hrsg.): *Neural Networks in Applications NN'98*, 1998, S. 9–16

[Baj85] Bajcsy, R.: Active perception vs. passive perception. In: *Proc. 3rd Workshop on Computer Vision: Representation and Control*, 1985, S. 55–59

[Baj88] Bajcsy, R.: Active perception. In: *Proceedings IEEE* Bd. 76, 1988, S. 996–1005

[Baj95] Bajcsy, R.: From Active Perception to Active Cooperation - Fundamental Processes of Intelligent Behavior. In: Zangenmeister, W.H. (Hrsg.) ; Stiehl, H.S. (Hrsg.) ; Freksa, C. (Hrsg.): *Visual Attention and Cognition*. W.H. Zangenmeister AND H.S. Stiehl AND C. Freksa, 1995, S. 309–321

[Bal91] Ballard, D.H.: Animate Vision. In: *Artificial Intelligence* 48 (1991), Nr. 1, S. 57–86

[Bal98] Baluja, S.: Using Expectation to Guide Processing: A Study of Three Real-World Applications. In: Jordan, M.I. (Hrsg.) ; Kearns, M.J. (Hrsg.) ; Solla, S.A. (Hrsg.): *Advances in Neural Information Processing Systems 10*, 1998

[Bar95] Barkley, R.A.: Is there an attention deficit in ADHD? In: *The ADHD Report* 3 (1995), S. 1–3

[BBB+98] Boehme, H.-J. ; Brakensiek, A. ; Braumann, U.-D. ; Krabbes, M. ; Gross, H.-M.: Neural Networks for Gesture-Based Remote Control of a Mobile Robot. In: *Proc. IEEE World Congress on Computational Intelligence WCCI '98 - IJCNN '98* Bd. 1. Anchorage, 1998, S. 372–377

[BBBM01] Bungenstock, M. ; Baudry, A. ; Bitterling, J. ; Mertsching, B.: Development of a Simulation Framework for Mobile Robots. In: *Proceedings of the EUROIMAGE ICAV3D 2001*, 2001, S. 89–92

[BBC+97] Barile, J. ; Bishay, M. ; Cambron, M. ; Watson, R. ; Peters, R.A. ; Kawamura, K.: Color-Based Initialisation for Human Tracking with a Trinocular Camera System. In: *Proc. of the Int. Conf. on Robotics and Manufacturing*, 1997

[BBH03] Brown, M.Z. ; Burschka, D. ; Hager, G.D.: Advances in Computational Stereo. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2003)

[BBM02] Baudry, A. ; Bungenstock, M. ; Mertsching, B.: Architectrue of a 3D-Simulation Environment for Active Vision Systems and Mobile Robots. In: *1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT), Padua, 2002*, 2002

[BD92] Baylis, G.C. ; Driver, J.: Visual parsing and response competition: The effect of grouping factors. In: *Perception and Psychophysics* 51 (1992), S. 145–162

[BD93] Baylis, G.C. ; Driver, J.: Visual attention and objects: Evidence for hierarchical coding of locations. In: *Journal of Experimental Psychology: Human Perception and Performance* 19 (1993), Nr. 3, S. 451–470

[BD98] Bruckhoff, C. ; Dahm, P.: Neural Fields for Local Path Planning. In: *Proceedings of the International Conference on Intelligent RObotic Systems (IROS 98)*, 1998, S. 1431–1436

[BD02] Blaser, E. ; Domini, F.: The conjuction of feature and depth information. In: *Vision Research* 42 (2002), Nr. 3, S. 273–279

[BE01] Björkman, M. ; Eklundh, J.-O.: Visual Cues for a Fixating Active Agent. In: Klette, R. (Hrsg.) ; Peleg, S. (Hrsg.) ; Sommer, G. (Hrsg.): *Robot Vision, International Workshop RobVis 2001*, Springer, 2001 (Lecture Notes in Computer Science), S. 1–9

[BET95] Bülthoff, H. ; Edelman, S. ; Tarr, M.: How are three-dimensional objects represented in the brain. In: *Cerebral Cortex* 5 (1995), Nr. 3, S. 247–260

[BEU94] Brunnström, K. ; Eklundh, J.-O. ; Uhlin, T.: Active Fixation for Scene Exploration. In: *Int. J. of Computer Vision* Bd. 17, 1994, S. 137–162

[BG83] Buchsbaum, G. ; Gottschalk, A.: Trichromacy, opponent colours coding and optimum colour information transmission in the retina. In: *Proceedings of the Royal Society (London) B* 220 (1983), S. 89–113

[BGG96] Bruce, V. ; Green, P.R. ; Georgeson, M.A.: *Visual Perception: Physiology, Psychology, and Ecology*. 3. Psychology Press, 1996

[BH99] Balkenius, C. ; Hulth, N.: Attention as Selection-for-Action: A Scheme for Active Perception. In: *Proc. EUROBOT 1999*, 1999

[BHM97] Bollmann, M. ; Hoischen, R. ; Mertsching, B.: Integration of static and dynamic scene features guiding visual attention. In: Paulus, E. (Hrsg.) ; Wahl, F.M. (Hrsg.): *Mustererkennung 1997*, 1997, S. 483–490

[Bie85] Biederman, I.: Human image understanding: Recent research and a theory. In: *Computer Vision, Graphics, and Image Processing* 32 (1985), S. 29–73

[Bie87] Biederman, I.: Recognition by components: A theory of human image understanding. In: *Psychological Review* 94 (1987), S. 115–145

[BJM98] Bollmann, M. ; Justkowski, C. ; Mertsching, B.: Utilizing Color Information for the Gaze Control of an Active Vision System. In: Rehrmann, V. (Hrsg.): *4. Workshop Farbbildverarbeitung*, 1998, S. 73–79

[BJT90] Barron, J.L. ; Jepson, A.D. ; Tsotsos, J.K.: The Feasibility of Motion and Structure from Noisy Time-Varying Image Velocity Information. In: *International Journal of Computer Vision* 5 (1990), Nr. 3, S. 239–269

[BL98] Bolduc, M. ; Levine, M.D.: A Review of Biologically Motivated Space-Variant Data Reduction Models for Robotic Vision. In: *Computer Vision and Image Understanding* 69 (1998), Nr. 2, S. 170–184

[BM95] Bollmann, M. ; Mertsching, B.: Vergleich zweier Farbkonstanzalgorithmen für die Bildanalyse. In: Rehrmann, V. (Hrsg.): *1. Workshop Farbbildverarbeitung*, 1995, S. 52–55

[BM00] Backer, G. ; Mertsching, B.: Integrating time and depth into the attentional control of an active vision system. In: Baratoff, G. (Hrsg.) ; Neumann, H. (Hrsg.): *Dynamische Perzeption. Workshop der GI-Fachgruppe 1.0.4 Bildverstehen, Ulm, November 2000*, 2000, S. 69–74

[BM02a] Backer, G. ; Mertsching, B.: Evaluation of Attentional Control in Active Vision Systems using a 3D Simulation Framework. In: *Journal of the WSCG - 10th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision* Bd. 10, 2002, S. 32–39

[BM02b] Backer, G. ; Mertsching, B.: Using neural field dynamics in the context of attentional control. In: *Proceedings of the ICANN 2002*, 2002, S. 1237–1242

[BM03] Backer, G. ; Mertsching, B.: Two Selection Stages Provide Efficient Object-based Attentional Control for Dynamic Vision. In: Paletta, L. (Hrsg.) ; Humphreys, G.W. (Hrsg.) ; Fisher, R.B. (Hrsg.): *Proc. of the International Workshop on Attention and Performance in Computer Vision, WAPCV 2003, Graz*, 2003, S. 9–16

[BMB01] Backer, G. ; Mertsching, B. ; Bollmann, M.: Data- and Model-Driven Gaze Control for an Active-Vision System. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), Nr. 12, S. 1415–1429

[BMD95] Bollmann, M. ; Mertsching, B. ; Drüe, S.: Entwicklung eines Gegenfarbmodells für das Neuronale-Active-Vision-System NAVIS. In: Sagerer, G. (Hrsg.) ; Posch, S. (Hrsg.) ; Kummert, F. (Hrsg.): *Mustererkennung 1995*, 1995, S. 456–463

[BMht] Backer, G. ; Mertsching, B.: Object-based computations for multiple levels of visual attention. In: *Computer Vision and Image Understanding* (eingereicht)

[BMR94] Burr, D.C. ; Morrone, M.C. ; Ross, J.: Selective suppression of the magnocellular visual pathway during saccadic eye movements. In: *Nature* 371 (1994), S. 511–513

[Bol00] Bollmann, M.: *Entwicklung einer Aufmerksamkeitssteuerung für ein aktives Sehsystem*, FB Informatik, Universität Hamburg, Dissertation, 2000

[BP93] Burkell, J. ; Pylyshyn, Z.W.: Indexing multiple loci in the visual field: Evidence for simultaneous facilitation in visual search / Centre for Cognitive Science, University of Western Ontario. 1993 (Cogmem 64). – Forschungsbericht

[BP97a] Baluja, S. ; Pomerleau, D.: Dynamic Relevance: Vision-Based Focus of Attention Using Artificial Neural Networks. In: *Artificial Intelligence* 97 (1997), S. 381–395

[BP97b] Baluja, S. ; Pomerleau, D.: Expectation-based Selective Attention for Visual Monitoring and Control of a Robot Vehicle. In: *Robotics and Autonomous Systems* 22 (1997), S. 329–344

[BP97c] Burkell, J. A. ; Pylyshyn, Zenon W.: Searching through subsets: a test of the visual indexing hypothesis. In: *Spatial Vision* 11 (1997), Nr. 2, S. 145–258

[BPH00] Blaser, E. ; Pylyshyn, Z.W. ; Holcombe, A.O.: Tracking an object through feature space. In: *Nature* 408 (2000), November, S. 196–199

[Bra01] Braumann, U.-D.: *Multi-Cue-Ansatz für ein dynamisches Auffälligkeitssystem zur visuellen Personenlokalisation*, Fakultät für Informatik und Automatisierung der TU Ilmenau, Dissertation, 2001

[Bro58] Broadbent, D.E.: *Perception and Communication*. London : Pergamon Press, 1958

[BS96] Bicho, E. ; Schöner, G.: The dynamic approach to autonomous robotics demonstrated on a low-level vehicle platform. In: *Proceedings of the fourth international symposium on Intelligent Robotic Systems (SIRS'96)*, 1996

[BSP01] Badenas, J. ; Sanchiz, J.M. ; Pla, F.: Motion-based segmentation and region tracking in image sequences. In: *Pattern Recognition* 34 (2001), S. 661–670

[BT94] Behrmann, M. ; Tipper, S.P.: Object-based attentional mechanisms: Evidence from patients with unilateral visual neglect. In: Umilta, C. (Hrsg.) ; Moscovitch, M. (Hrsg.): *Attention and Performance XV*. Cambridge, MA : MIT Press, 1994, S. 351–375

[Bun90] Bundesen, C.: A theory of visual attention. In: *Psychological Review* 97 (1990), S. 523–547

[Cav99] Cave, K.R.: The FeatureGate model of visual attention. In: *Psychological Research* 62 (1999), S. 182–194

[CBB+98] Corradini, A. ; Braumann, U.-D. ; Brakensiek, A. ; Krabbe, M. ; Boehme, H.-J. ; Gross, H.-M.: Visual Person Localization with Dynamic Neural Fields: Towards a Gesture Recognition System. In: *Proceedings fo the 10. Workshop Italiano sulle Reti Neurali (WIRN98), Vietri Sul Mare 1998*, 1998, S. 201–206

[CBBG98] Corradini, A. ; Braumann, U.-D. ; Boehme, H.-J. ; Gross, H.-M.: 3D Neural Fields and Steerable Filters for Contour-Based Person Localization. In: *Proc. Workshop on Virtual Intelligence - Dynamic Neural Networks, VI-DYNN '98*, 1998

[CBBS94] Crowley, J.L. ; Bedrone, J. M. ; Bekker, M. ; Schneider, M.: Integration and Control of Reactive Visual Processes. In: *Proceedings of the European Conference on Computer Vision ECCV '94*, 1994, S. 47–58

[CC95] Crowley, J.L. (Hrsg.) ; Christensen, H.I. (Hrsg.): *Vision as Process*. Springer, 1995

[CF92] Clark, J.J. ; Ferrier, N.J.: Attentive Visual Servoing. In: *Active Vision*. A. Blake AND A. Yuille, 1992, S. 137–154

[CF01] Callari, F.G. ; Ferrie, F.P.: Active Object Recognition: Looking for Differences. In: *International Journal of Computer Vision* 43 (2001), Nr. 3, S. 198–204

[CGL98] Carpenter, G.A. ; Grossberg, S. ; Lesher, G.W.: The What-and-Where Filter. In: *Computer Vision and Image Understanding* 69 (1998), 1, Nr. 1, S. 1–22

[Che53] Cherry, E.: Some experiments on the recognition of speech, with one and with two ears. In: *Journal of the Acoustical Society of America* 25 (1953), S. 975–979

[Che99] Chelazzi, L.: Serial attention mechanisms in visual search: A critical look at the evidence. In: *Psychological Research* 62 (1999), S. 195–219

[CIE03] CIE. Commision Interntationale De L'Eclairage. http://www.cie.co.at/cie/home.html. 2003

[CLT01] Cavanagh, P. ; Labianca, A.T. ; Thornton, I.M.: Attention-based visual routines: Sprites. In: *Cognition* 80 (2001), S. 47–60

[Coh93] Cohen, A.: Asymmetries in visual search for conjunctive targets. In: *Journal of Experimental Psychology: Human Perception and Performance* 19 (1993), Nr. 4, S. 775–797

[Cor90] Corbetta, M.: Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? In: *Proc. Natl. Acad. Sci.* 95 (1990), 2, S. 831–838

[CR75] Corbalis, M. ; Roldan, C.: Detection of symmetry as a function of angular rotation. In: *Journal of Experimental Psychology: Human Perception and Performance* 1 (1975), Nr. 3, S. 221–230

[CRD94] Colombo, C. ; Rucci, M. ; Dario, P.: Attentive behavior in an anthropomorphic robot vision system. In: *Robotics and Autonomous Systems* 12 (1994), S. 121–131

[CRD96] Colombo, C. ; Rucci, M. ; Dario, P.: Integrating Selective Attention and Space-Variant Sensing in Machine Vision. In: *Image Technology*. J.L.C. Sanz, 1996

[CT92] Culhane ; Tsotsos, J.K.: An attentional prototype for early vision. In: *ECCV '92. The 2nd European Conference on Computer Vision*, Springer, 1992

[CTW+00] Casey, B.J. ; Thomas, K.M. ; Welsh, T.F. ; Badgaiyan, R.D. ; Eccard, C.H. ; Jennings, J.R. ; Crone, E.A.: Dissociations of response conflict, attentional selection, and expectancy with functional magnetic resonance imaging. In: *Proc. Natl. Acad. Sci.* 97 (2000), Nr. 15, S. 8728–8733

[CVAC00] Culham, J.C. ; Verstraten, F.A.J. ; Ashida, H. ; Cavanagh, P.: Independent Aftereffects of Attention and Motion. In: *Neuron* 28 (2000), S. 607–615

[CW78] Campbell, F.W. ; Wurtz, R.H.: Saccadic omission: why we do not see a gray-out during saccadic eye movement. In: *Vision Research* 18 (1978), S. 1297–1303

[CW96] Chun, M.M. ; Wolfe, J.M.: Just say no: How are visual searches terminated when there is no target present? In: *Cognitive Psychology* 30 (1996), S. 39–78

[DBJ98] Dahm, P. ; Bruckhoff, C. ; Joublin, F.: A Neural Field Approach to Robot Motion Control. In: *Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics (SMC'98)*, 1998, S. 3460–3465

[DCTO97] Dickinson, S.J. ; Christensen, H.I. ; Tsotsos, J.K. ; Olofsson, G.: Active Object Recognition Integrating Attention and Viewpoint Control. In: *Computer Vision and Image Understanding* 67 (1997), Nr. 3, S. 239–260

[DD63] Deutsch, J.A. ; Deutsch, D.: Attention: Some theoretical considerations. In: *Psychological Review* 70 (1963), S. 80–90

[Dec00] Deco, G.: A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system. In: Baratoff, G. (Hrsg.) ; Neumann, H. (Hrsg.): *Dynamische Perzeption*, 2000, S. 45–50

[Dec01] Deco, G.: Biased Competition Mechanism for Visual Attention in a Multimodular Neurodynamical System. In: Wermter, S. (Hrsg.) ; Austin, J. (Hrsg.) ; Willshaw, D. (Hrsg.): *Emergent Neural Computation Architectures*. 2001, S. 114–126

[DES00] Ditterich, J. ; Eggert, T. ; Straube, A.: The role of the attention focus in the visual information processing underlying saccadic adaptation. In: *Vision Research* 40 (2000), S. 1125–1134

[Dic92] Dickmanns, E.D.: Expectation-based dynamic scene understanding. In: Blake, A. (Hrsg.) ; Yuille, A. (Hrsg.): *Active Vision*. 1992, Kapitel 18, S. 303–336

[Dic98] Dickmanns, D.: *Rahmensystem für visuelle Wahrnehmung veränderlicher Szenen durch Computer*. Shaker Verlag, 1998

[DPC98] Driscoll, J.A. ; Peters, R.A. ; Cave, K.R.: A Visual Attention Network for a Humanoid Robot. In: *Proceedings of the 1998 IEEE/RSJ International Conference on Intelligent Robotic Systems (IROS'98), Victoria, B.C.*, 1998

[Dun80] Duncan, J.: The locus of interference in the perception of simultaneous stimuli. In: *Psychological review* 87 (1980), S. 272–300

[DW99] Dickmanns, E.D. ; Wünsche, H.J.: Dynamic Vision for Perception and Control of Motion. In: Jähne, B. (Hrsg.) ; Haussecker, H. (Hrsg.) ; Geissler, P. (Hrsg.): *Handbook of Computer Vision and Applications* Bd. 3. London : Academic Press, 1999, S. 569–622

[Ebn01] Ebner, M.: Evolving Color Constancy for an Artificial Retina. In: Miller, J. (Hrsg.) ; Tomassini, M. (Hrsg.) ; Lanzi, P.L. (Hrsg.) ; Ryan, C. (Hrsg.) ; Tettamanzi, A.G.B. (Hrsg.) ; Langdon, W.B. (Hrsg.): *Genetic Programming: 4th European Conference, EuroGP 2001* Bd. 2038. Berlin, Heidelberg : Springer, 2001, S. 11–22

[Ege77] Egeth, H.: Attention and preattention. In: Bower, G.H: (Hrsg.): *The Psychology of Learning and Motivation* Bd. 11. New York : Academic Press, 1977, S. 277–320

[EH73] Eriksen, C.W. ; Hoffman, J.E.: The extent of processing of noise elements during selective encoding from visual displays. In: *Perception and Psychophysics* 14 (1973), S. 155–160

[EJ86] Eriksen, C. W. ; James, J. S.: Visual attention within and around the focus of attention: A zoom lens model. In: *Perception and Psychophysics* 40 (1986), Nr. 4, S. 225–240

[EM87] Eriksen, C.W. ; Murphy, T.D.: Movement of attentional focus across the visual field: A critical look at the evidence. In: *Perception and Psychophysics* 42 (1987), Nr. 3, S. 299–305

[Enn90] Enns, J.T. (Hrsg.): *The development of attention*. Elsevier Science, 1990

[ES95] Engels, C. ; Schöner, G.: Dynamic fields endow behavior-based robots with representations. In: *Robotics and Autonomous Systems* 14 (1995), S. 55–77

[ESA02] Eckstein, M.P. ; Shimozaki, S.S. ; Abbey, C.K.: The footprints of visual attention in the Posner cueing paradigm revealed by classification images. In: *Journal of Vision* 2 (2002), S. 25–45

[Eve95] Everett, H.R.: *Sensors for Mobile Robots - Theory and Application*. A K Peters, 1995

[EY85] Eriksen, C. W. ; Yeh, Y. Y.: Allocation of attention in the visual field. In: *Journal of Experimental Psychology: Human Perception and Performance* 11 (1985), Nr. 5, S. 583–597

[FA95] Fermüller, C. ; Aloimonos, Y.: Vision and Action. In: *Image and Vision Computing* 13 (1995), Nr. 10, S. 725–744

[Fau92] Faugeras, O.D.: What can be seen in three dimensions with an uncalibrated camera rig? In: *Proceedings of the second European Conference on Computer Vision (ECCV)*, 1992, S. 563–578

[FB83] Fischer, B. ; Boch, R.: Saccadic eye movements after extremely short reaction times in the monkey. In: *Brain Research* 260 (1983), S. 21–26

[FBG01] Findlay, J.M. ; Brown, V. ; Gilchrist, I.D.: Saccade target selection in visual search: the effect of information from the previous fixation. In: *Vision Research* 41 (2001), S. 87–95

[Fel97] Fellenz, W.A.: *Ein neuromorphes System für die datengetriebene Szenenanalyse*, U-GH Paderborn, Dissertation, 1997

[FH96] Fellenz, W.A. ; Hartmann, G.: Preattentive grouping and Attentive selection for early visual computation. In: *13th ICPR 1996, International 25 - 30, 1996 Conference on Pattern Recognition Technical University, Wien, August*, 1996

[Fis98] Fischer, B.: Attention in Saccades. In: Wright, R.D. (Hrsg.): *Visual attention*. Oxford University Press, 1998, S. 289–305

[FKPS95] Firby, R.J. ; Kahn, R.E. ; Prokopowitz, P.N. ; Swain, M.J.: An Architecture for Vision and Action. In: Mellish, C. (Hrsg.): *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, S. 72–79

[FPCR96] Fayman, J. ; Pirjanian, P. ; Christensen, H.I. ; Rivlin, E.: Exploiting Redundancy of Purposive Modules in the Context of Active Vision / Technion, Haifa, Israel. 1996. – Forschungsbericht

[FRNS03] Frintrop, S. ; Rome, E. ; Nuechter, A. ; Surmann, H.: An Attentive, Multi-modal Laser Ëye". In: Crowley, J. (Hrsg.) ; Piater, J.H. (Hrsg.) ; Vincze, M. (Hrsg.) ; Paletta, L. (Hrsg.): *Proceedings of the 3rd International Conference on Computer Vision Systems, ICVS 2003*, 2003, S. 202–211

[FRR99] Fislage, M. ; Rae, Robert ; Ritter, H.: Using Visual Attention to Recognize Human Pointing Gestures in Assembly Tasks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999

[FRW94] Folk, C. ; Remington, R.W. ; Wright, J.: The structure of attentional control: Contigengent attentional capture by apparent motion, abrupt onset and color. In: *Journal of Experimental Psychology: Human Perception and Performance* 20 (1994), Nr. 2, S. 317–329

[FSP93] Fisher, B.D. ; Schmidt, W.C. ; Pylyshyn, Z.W.: Multiple abrupt onset cues produce illusory line motion. In: *Investigative Ophthalmology and Visual Science* 34 (1993)

[FTIC92] Fujita, I. ; Tanaka, K. ; Ito, M. ; Cheng, K.: Columns for visual features of objects in monkey inferotemporal cortex. In: *Nature* 360 (1992), S. 343–346

[Gab46] Gabor, D.: Theory of communication. In: *Journal IEEE* 93 (1946), S. 429–457

[Gib79] Gibson, J.J.: *The ecological approach to visual perception*. Boston, MA : Houghton Mifflin, 1979

[Gie99] Giese, M. A.: *Dynamic Neural Field Theory for Motion Perception*. Kluwer Academic, 1999

[GJM92] Giefing, G.-J. ; Janssen, H. ; Mallot, H.A.: Saccadic object recognition with an active vision system. In: *10th European Conference on Artificial Intelligence*, John Wiley and Sons, 1992, S. 803–805

[GKG98] Gottlieb, J.P. ; Kusunoki, M. ; Goldberg, M.E.: The representation of visual salience in monkey parietal cortex. In: *Nature* 391 (1998), S. 481–484

[GKWW94] Granlund, G. H. ; Knutsson, H. ; Westelius, C.-J. ; Wiklund, J.: Issues in Robot Vision. In: *Image and Vision Computing* 12 (1994), Nr. 3, S. 131–148

[GSH96] Giese, M.A. ; Schöner, G. ; Hock, H.S.: Neural Field Dynamics for Motion Perception. In: Malsburg, C. von d. (Hrsg.) ; Seelen, W. von (Hrsg.) ; Vorbrüggen, J.C. (Hrsg.) ; Sendhoff, B. (Hrsg.): *Artificial Neural Networks - ICANN 96*, Springer, Berlin, 1996, S. 335–340

[GT98] Grinias, I. ; Tziritas, G.: Motion Segmentation and Tracking using a Seeded Region Growing Method. In: *European Signal Processing Conference*, 1998

[GVG02] Gysen, V. ; Verfaillie, K. ; Graef, P. D.: Transsaccadic perception of translating objects: effects of landmark objects and visual field position. In: *Vision Research* 42 (2002), S. 1785–1796

[Ham00] Hamker, F. H.: Distributed competition in directed attention. In: Baratoff, G. (Hrsg.) ; Neumann, H. (Hrsg.): *Dynamische Perzeption*, 2000, S. 39–44

[Har87] Harrington, S.: *Computer graphics: A programming approach.* McGraw-Hill, 1987

[HAV98] Hillyard, S.A. ; Anllo-Vento, L.: Event-related brain potentials in the study of visual selective attention. In: *Proc. Natl. Acad. Sci.* 95 (1998), 2, S. 781 787,

[HB96] Horswill, I. ; Barnhart, C.: Unifying segmentation, tracking, and visual search / Northwestern University. 1996. – Forschungsbericht

[HB99] Hübner, R. ; Backer, G.: Perceiving spatially inseparable objects: evidence for feature-based object selection not mediated by location. In: *Journal of Experimental Psychology: Human Perception and Performance* 23 (1999), Nr. 4, S. 948–961

[HBD99] Hartmann, G. ; Büker, U. ; Drüe, S.: A Hybrid Neuro-Artificial Intelligence-Architecture. In: Jähne, B. (Hrsg.) ; Haussecker, H. (Hrsg.) ; Geissler, P. (Hrsg.): *Handbook of Computer Vision and Applications* Bd. 3. London : Academic Press, 1999, S. 153–196

[HCT] Hassoumi, N. ; Chiva, E. ; Tarroux, P.: A neural model of preattentional and attentional visual search. – Unveröffentlichtes Manuskript

[Hel96] Helmholtz, H. von: *Handbuch der Physiologischen Optik*. Bd. 2. Voss: Hamburg, 1896

[HF00] Hooge, I.T.C. ; Frens, M.A.: Inhibition of saccade return (ISR): spatio-temporal properties of saccade programming. In: *Vision Research* 40 (2000), S. 3415–3426

[HG97] Hamker, F.H. ; Gross, H.-M.: Region selection with dynamic neural maps. In: *Proceedings of the International Conference on Artificial Neural Network (ICANN'97)*, 1997, S. 919–924

[HG01] Heinze, A. ; Gross, H.-M.: Anticipation-Based Control Architecture for a Mobile Robot. In: Dorffner, G. (Hrsg.) ; Bischof, H. (Hrsg.) ; Hornik, K. (Hrsg.): *ICANN 2001*, 2001, S. 899–905

[HKKM97] Heslenfeld, D. J. ; Kenemans, J. L. ; Kok, A. ; Molenaar, P.C.M.: Feature processing and attention in the human visual system: an overview. In: *Biological Psychology* 45 (1997), S. 183–215

[HM93] Humphreys, G.W. ; Müller, H.J.: Search via recursive rejection (SERR): A connectionist model of visual search. In: *Cognitive Psychology* 25 (1993), S. 43–110

[HMS93] Hikosaka, O. ; Miyauchi, S. ; Shimojo, S.: Focal visual attention produces motion sensation in lines. In: *Investigative Ophthalmology and Visual Science* 32 (1993)

[HMS99] Hoischen, R. ; Mertsching, B. ; Springmann, S.: Object Tracking in Image Sequences Based on Parametric Features. In: *ÖVE Verbandszeitschrift Elektrotechnik und Informationstechnik (e & i)* (1999), Nr. 6, S. 390–394

[HN95] He, Z.J. ; Nakayama, K.: Visual attention to surfaces in three-dimensional space. In: *Proc. Natl. Acad. Sci.* 92 (1995), November, S. 11155–11159

[HNR98] Heidemann, G. ; Nattkemper, T. ; Ritter, H.: Farbe und Symmetrie für die datengetriebene Generierung prägnanter Fokuspunkte. In: Rehrmann, V. (Hrsg.): *4. Workshop Farbbildverarbeitung. Koblenz (Fölbach) 1998*, 1998, S. 65–72

[HS81] Horn, B.K.P. ; Schunk, B.G.: Determinig optical flow. In: *Artificial Intelligence* 17 (1981), S. 185–203

[HS99] Haussecker, H. ; Spies, H.: Motion. In: Jähne, B. (Hrsg.) ; Haussecker, H. (Hrsg.) ; Geissler, P. (Hrsg.): *Handbook of Computer Vision and Applications* Bd. 2. 1999, S. 309–396

[Hun87] Hunt, R.W.G.: *Measuring Colour*. Chichester: Ellis Horwood, 1987

[HW84] Heisenberg, M. ; Wolf, R.: *Studies of brain function: vision in Drosophila*. Bd. 12. Springer-Verlag, 1984

[HW98] Horowitz, T.S. ; Wolfe, J.M.: Visual search has no memory. In: *Nature* 394 (1998), S. 575–577

[IB98a] Isard, M. ; Blake, A.: Condensation - conditional density propagation for visual tracking. In: *International Journal of Computer Vision* 29 (1998), Nr. 1, S. 5–28

[IB98b] Isard, M. ; Blake, A.: ICONDENSATION: Unifying low-level and high-level tracking in a Stochastic Framework. In: *Proceedings of the ECCV'98*, 1998, S. 893–908

[IC01] Intriligator, J. ; Cavanagh, P.: The Spatial Resolution of Visual Attention. In: *Cognitive Psychology* (2001)

[IEJ01] Igel, C. ; Erlhagen, W. ; Jancke, D.: Optimization of Neural Field Models. In: *Neurocomputing* 36 (2001), Nr. 1-4, S. 225–233

[IK00] Itti, L. ; Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. In: *Vision Research* 10-12 (2000), 6, S. 1489–1506

[IK01a] Itti, L. ; Koch, C.: Computational Modelling of Visual Attention. In: *Nature Reviews: Neuroscience* 2 (2001), 3

[IK01b] Itti, L. ; Koch, C.: Feature Combination Strategies for Saliency-Based Visual Attention Systems. In: *Journal of Electronic Imaging* 10 (2001), Nr. 1

[IKN98] Itti, L. ; Koch, C. ; Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), Nr. 11, S. 1254–1259

[Ill86] Illumination, International C. Colorimetry. Central Bureau of the CIE, Vienna, Austria, CIE Pub. 15.2. 1986

[Itt00] Itti, L.: *Models of Bottom-Up and Top-Down Visual Attention*, California Institute of Technology, Dissertation, 2000

[Jäh97] Jähne, B.: *Digitale Bildverarbeitung*. Springer, 1997

[JB97] Jiand, X. ; Bunke, H.: *Dreidimensionales Computersehen - Gewinnung und Analyse von Tiefenbildern*. Berlin, Heidelber, New York : Springer Verlag, 1997

[JC79] Julesz, B. ; Chang: Symmetry perception and spatial frequency channels. In: *Perception* 8 (1979), S. 711–718

[JD99] Joliceour, P. ; Dell'Acqua, R.: Attentional and structural constraints on visual encoding. In: *Psychological Research* 62 (1999), S. 154–164

[Jäg95] Jägersand, M.: Saliency Maps and Attention Selection in Scale and Spatial Coordinates: An Information Theoretic Approach. In: *Proceedings of the 5th International Conference on Computer Vision 1995*, 1995, S. 195–202

[JJFK01] Jirsa, V.K. ; Jantzen, K.J. ; Fuchs, A. ; Kelso, J.A.S.: Neural field dynamics on the folded three-dimensional cortical sheet and its forward EEG and MEG. In: Insana, M.F. (Hrsg.) ; Leahy, R.M. (Hrsg.): *Information Processing in Medical Imaging*, 2001, S. 286–299

[JKS95] Jain, R. ; Kasturi, R. ; Schunck, B.G.: *Machine Vision*. McGraw-Hill, 1995

[JMD94] Johnson, M.P. ; Maes, P. ; Darrel, T.: Evolving Visual Routines. In: Brooks, R. (Hrsg.) ; Maes, P. (Hrsg.): *Artifical Life IV*, 1994, S. 198–209

[JMNS96] Jähne, B. ; Massen, R. ; Nickolay, B. ; Scharfenberg, H.: *Technische Bildverarbeitung - Maschinelles Sehen*. Springer, 1996

[JMR96] Johnston, J.C. ; McCann, R.S. ; Remington, R.W.: Selective Attention Operates at Two Processing Loci. In: Kramer, A. (Hrsg.) ; Logan, G. (Hrsg.): *Essays in Honor of Charles Eriksen*. American Psychological Association, 1996, S. 439–458

[Jul71] Julesz, B.: *Foundations of Cyclopean Perception*. Chicago and London : Chicago University Press, 1971

[KA79] Kishimoto, K. ; Amari, S.-I.: Existence and stability of local excitations in homogenous neural fields. In: *Journal of Mathematical Biology* 7 (1979), S. 303–318

[KADB95] Kowler, E. ; Anderson, E. ; Dosher, B. ; Blaser, E.: The Role of Attention in the Programming of Saccades. In: *Vision Research* (1995), Nr. 13, S. 1897–1916

[KC01] Kim, M.S. ; Cave, K.R.: Perceptual grouping via spatial selection in a focused-attention task. In: *Vision Research* 41 (2001), S. 611–624

[KH95] Kramer, A.F. ; Hahn, S.: Splitting the Beam: Distribution of Attention Over Noncontiguous Regions of the Visual Field. In: *Psychological Science* 6 (1995), Nr. 6, S. 381–386

[KJ91] Kramer, A.F. ; Jacobson, A.: Perceptual organization and focused attention: The role of objects and proximity in visual processing. In: *Perception and Psychophysics* 50 (1991), Nr. 3, S. 267–284

[KK94] Kaski, S. ; Kohonen, T.: Winner-Take-All Network for Physiological Models of Competitive Learning. In: *Neural Networks: Official Journal of INNS, ENNS, and JNNS* (1994)

[Kle80] Klein, R.: Does oculomotor readiness mediate cognitive control of visual attention? In: Nickerson, R.S. (Hrsg.): *Attention and Performace VIII*. Hillsdale, NJ : Erlbaum, 1980, S. 259–276

[Kle02] Klein, J.: A 3D simulation environment for the simulation of decentralized systems and artificial life. In: *Proceedings of Artificial Life VIII, the 8th International Conference on the Simulation and Synthesis of Living Systems*, 2002

[KM96] Kehrer, L. ; Meinecke, C.: Perzeptive Organisation visueller Muster: Die Segmentierung von Texturen. In: *Enzyklopädie der Psychologie: Wahrnehmung - Kognition*. Hogrefe, Göttingen, 1996

[KMM94] Konen, W.K ; Maurer, T. ; Malsburg, C. von d.: A Fast Dynamic Link Matching Algorithm for Invariant Pattern Recognition. In: *Neural Networks* 7 (1994), Nr. 6/7, S. 1019–1030

[Kof35] Koffka, K.: *Principles of Gestalt Psychology*. New York : Harcourt, 1935

[Kon03] Konolige, K.: *Saphira Software Manual Version 6.1*. http://www.ai.sri.com/ konolige/saphira/: ActiveMedia, 2003

[Kop96] Kopecz, K.: Neural Field Dynamics Provide Robust Control of Attentional Resources. In: Mertsching, B. (Hrsg.): *Aktives Sehen in technischen und natürlichen Systemen*, 1996, S. 137–144

[Koz92] Koza, J.R.: *Genetic Programming - On the Programming of Computers by Means of Natural Selection*. Cambridge : MIT Press, 1992

[Koz94] Koza, J.R.: *Genetic Programming II - Automatic Discovery of Reusable Programs*. Cambridge : MIT Press, 1994

[KR69] Kaufman, L. ; Richards, W.: Spontaneous Fixation Tendencies for Visual Forms. In: *Perception and Psychophysics* 5 (1969), Nr. 2, S. 85–88

[KR99] Kundur, S.R. ; Raviv, D.: Novel Active Vision-Based Visual Threat Cue for Autonomous Navigation Tasks. In: *Computer Vision and Image Understanding* 73 (1999), February, Nr. 2, S. 169–182

[KT84] Kahneman, D. ; Treisman, A.: Changing views of attention and automaticity. In: Parasuraman, R. (Hrsg.) ; Davies, D.A. (Hrsg.): *Varieties of attention*. New York : Academic press, 1984

[KTG92] Kahneman, D. ; Treisman, A. ; Gibbs, B.J.: The reviewing of object files: object-specific integration of information. In: *Cognitive Psychology* 24 (1992), Nr. 2, S. 175–210

[KU85] Koch, C. ; Ullman, S.: Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. In: *Human Neurobiology* 4 (1985), S. 219–227

[KU00] Kastner, S. ; Ungerleider, L.G.: Mechanisms of Visual Attention in the Human Cortex. In: *Annual Review of Neuroscience* 23 (2000), S. 315–341

[KW96] Kolb, B. ; Whishaw, I.Q.: *Neuropsychologie*. Spektrum Akademischer Verlag, 1996

[KWT87] Kass, M. ; Witkin, A. ; Terzopoulos, D.: Snakes: Active Contour Models. In: *Proceedings of the International Conference on Computer Vision ICCV 1987*, 1987

[KWW97] Kramer, A.F. ; Weber, T.A. ; Watson, S.E.: Object-based attentional selection - grouped arrays or spatially invariant representations? Comment on Vecera and Farah(1994). In: *Journal of Experimental Psychology: General* 126 (1997), S. 3–13

[LB89] LaBerge, D. ; Brown, V.: Theory of Attentional Operation in Shape Identification. In: *Psychological Review* 96 (1989), Nr. 1, S. 101–124

[LB98] Luck, S.J. ; Beach, N.J.: Visual Attention and the Binding Problem: A Neurophysiological Perspective. In: Wright, R.D. (Hrsg.): *Visual Attention*. Oxford University PRess, 1998, S. 455–478

[LBP00] Lee, S.-W. (Hrsg.) ; Bülthoff, H.H. (Hrsg.) ; Poggio, T. (Hrsg.): *Biologically Motivated Computer Vision (Lecture Notes in Computer Science)*. Bd. 1811. Springer, 2000

[LCWB97] LaBerge, D. ; Carlson, R.L. ; Williams, J.K. ; Bunney, B.G.: Shifting Attention in Visual Space: Tests of Moving-Spotlight Models Versus an Activity-Distribution Model. In: *Journal of Experimental Psychology: Human Perception and Performance* 23 (1997), Nr. 5, S. 1380–1392

[LD03] Lionelle, A. ; Draper, B.A.: Evaluation of Selective Attention under Similarity Transforms. In: Paletta, L. (Hrsg.) ; Humphreys, G.W. (Hrsg.) ; Fisher, R.B. (Hrsg.): *Journal of the International Workshop on Attention and Performance in Computer Vision, WAPCV 2003, Graz*, 2003, S. 31–38

[Lea94] Leavers, V.F.: Preattentive computer vision - Towards a 2-stage computer vision system for the extraction of qualitative descriptors and the cues for focus of attention. In: *Image and Vision Computing* 12 (1994), Nr. 9, S. 583–599

[LF98] Luck, S.J. ; Ford, M.A.: On the role of attention in visual perception. In: *Proc. Nat. Acad. Sci. xxx* Bd. 95, 1998, S. 825–830

[Lie99] Lieder, T.: *Integration von Tiefen- und Bewegungsschätzung bei der Analyse von Stereobildfolgen*, AG IMA, FB Informatik, Universität Hamburg, Diplomarbeit, 1999

[LL00] Lee, S.-I. ; Lee, S.-Y.: Robust pattern recognition based on selective attention at feature space. In: *International Conference on Neural Information Processing*, 2000, S. 440–444

[LMS98] Lieder, T. ; Mertsching, B. ; Schmalz, S.: Using depth information for invariant object recognition. In: Posch, S. (Hrsg.) ; Ritter, H. (Hrsg.): *Dynamische Perzeption*, St. Augustin (Infix), 1998, S. 9–16

[Log95] Logie, R.H.: *Visuo-spatial working memory*. Lawrence Erlbaum, 1995

[Log96] Logan, G.D.: The CODE Theory of Visual Attention: An Integration of Space-Based and Object-Based Attention. In: *Psychological Review* 103 (1996), Nr. 4, S. 603–649

[LS96] Livingstone, D. ; Spacek, L.: A Behavioural Vision System for Search and Motion Tracking / University of Essex, Dept. of Computer Science. 1996 (CSM-268). – Forschungsbericht

[LTGE02] Laing, C.R. ; Troy, W.C. ; Gutkin, B. ; Ermentrout, G.B.: Multiple bumps in a neuronal model of working memory. In: *SIAM Journal Appl. Math.* 63 (2002), Nr. 1, S. 62–97

[LV97] Luck, S.J. ; Vogel, E.K.: The capacity of visual working memory for features and conjunctions. In: *Nature* (1997), Nr. 390, S. 279–281

[LX00] Lu, J.Z. ; Xie, M.: Simulation of Vision-Guided Vehicle. In: *Sixth International Conference on Control, Automation, Robotics and Vision, Singapore*, 2000

[MAJ00] MacAskill, M. R. ; Anderson, T.J. ; Jones, R.D.: Suppression of displacement in severely slowed saccades. In: *Vision Research* 40 (2000), S. 3405–3413

[Mak96] MAKI, A.: *Stereo Vision in Attentive Scene Analysis*, KTH Stockholm, Dissertation, 1996

[Mal98] MALLOT, H.A.: *Sehen und die Verarbeitung visueller Information.* Vieweg, Braunschweig, 1998

[Mal99] MALLOT, H.A.: Stereopsis - Geometrical and Global Aspects. In: JÄHNE, B. (Hrsg.) ; HAUSSECKER, H. (Hrsg.) ; GEISSLER, P. (Hrsg.): *Handbook of Computer Vision and Applications* Bd. 2. Academic Press, 1999, S. 485–504

[Mar82] MARR, D.: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco : W.H. Freeman, 1982

[MBHS99] MERTSCHING, B. ; BOLLMANN, M. ; HOISCHEN, R. ; SCHMALZ, S.: The Neural Active Vision System NAVIS. In: JÄHNE, B. (Hrsg.) ; HAUSSECKE, H. (Hrsg.) ; GEISSLER, P. (Hrsg.): *Handbook of Computer Vision and Applications Vol. 3 (Systems and Applications).* Academic Press, 1999, S. 543–568

[MBM+95] MURRAY, D.W. ; BRADSHAW, K.J. ; MCLAUCHLAN, P.F. ; REID, I.D. ; SHARKEY, P.M.: Driving Saccade to Pursuit using Image Motion. In: *Proceedings of the 5th International Conference on Computer Vision 1995*, 1995

[MBM02] MASSAD, A. ; BABOS, M. ; MERTSCHING, B.: Perceptual Grouping in Grey Level Images by Combination of Gabor Filtering and Tensor Voting. In: *International Conference on PAttern Recognition (ICPR2002), Quebec, Canada*, 2002

[MD85] MORAN, J. ; DESIMONE, R.: Selective attention gates visual processing in the exstrastriate cortex. In: *Science* 229 (1985), S. 782–784

[ME97] MOORE, C. M. ; EGETH, H.: Perception without Attention: Evidence of Grouping under Conditions of Inattention. In: *Journal of Experimental Psychology: Human Perception and Performance* 23 (1997), Nr. 2, S. 339–352

[MEHS00] MALINOV, I.V. ; EPELBOIM, J. ; HERST, A.N. ; STEINMAN, R.M.: Characteristics of saccades and vergence in two kinds of sequential looking tasks. In: *Vision Research* 40 (2000), S. 2083–2090

[Mer96] MERTSCHING, B. (Hrsg.): *Aktives Sehen in technischen und biologischen Systemen* infix, 1996

[MF01] MCSORLEY, E. ; FINDLAY, J.M.: Visual search in depth. In: *Vision Research* 41 (2001), Nr. 25-26, S. 3487–3496

[MH01] MALINOWSKI, P. ; HÜBNER, R.: The effect of familiarity on visual-search performance: Evidence for learned basic features. In: *Perception and Psychophysics* 63 (2001), S. 458–463

[MHS+01] Mangun, G.R. ; Hinrichs, H. ; Scholz, M. ; Müller-Gärtner, H.W. ; Herzog, H. ; Krause, B.J. ; Tellman, L. ; L.Kemna ; Heinze, H.J.: Integrating electrophysiology and neuroimaging of spatial selective attention to simple isolated visual stimuli. In: *Vision Research* 41 (2001), S. 1423–1435

[MI01] Miau, F. ; Itti, L.: A Neural Model Combining Attentional Orienting to Object Recognition: Preliminary Explorations on the Interplay Between Where and What. In: *Proceedings IEEE Engineering in Medicine and Biology Society (EMBS)*, 2001

[Mic96] Michel, O.: *Khephera Simulator 2.0 - User Manual*, 1996

[Mil91] Miller, J.O.: The flanker compatibility effct as a function of visual angle, attention focus, visual transients, and perceptual load: A search for boundary conditions. In: *Perception and Psychophysics* 49 (1991), S. 270–288

[Mil93] Milanese, R.: *Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation*, University of Geneva, Dissertation, 1993

[MK01] Melcher, D. ; Kowler, E.: Visual scene memory and the guidance of saccadic eye movements. In: *Vision Research* 41 (2001), S. 3597–3611

[MM95] McLauchlan, P.F. ; Murray, D.W.: A unifying framework for structure and motion recovery from image sequences. In: Grimson, E. (Hrsg.): *Proceedings 5th IEEE Int. Conf. Computer Vision*. Cambridge, MA, 1995, S. 314–320

[MMGH03] Müller, M.M. ; Malinowski, P. ; Gruber, T. ; Hillyard, S.A.: Sustained division of the attentional spotlight. In: *Nature* 424 (2003), Nr. 6946, S. 309–312

[MMII99] Matsumoto, Y. ; Miyazaki, T. ; Inaba, M. ; Inoue, H.: View Simulation System : A Mobile Robot Simulator using VR Technology. In: *Proceedings of 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'99)*. Kyongju, Korea, 1999, S. 936–941

[MMS98a] Massad, A. ; Mertsching, B. ; Schmalz, S.: Combining multiple views and temporal associations for 3-D object recognition. In: *Proceedings of the ECCV'98*, 1998, S. 699–715

[MMS98b] Massad, A. ; Mertsching, B. ; Schmalz, S.: 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, 1998, S. 2074–2078

[MNE96] Maki, A. ; Nordlund, P. ; Eklundh, J.-O.: A Computational Model of Depth-Based Attention. In: *Proc. 13th Int. Conf. on Pattern Recognition* Bd. 4, 1996, S. 734–738

[MNE00] Maki, A. ; Nordlund, P. ; Eklundh, J.-O.: Attentional Scene Segmentation: Integrating Depth and Motion. In: *Computer Vision and Image Understanding* 78 (2000), S. 351–373

[MPI01] Miau, F. ; Papageorgiou, C. ; Itti, L.: Neuromorphic algorithms for computer vision and attention. In: *Proceedings SPIE 46 Annual International Symposium on Optical Science and Technology*, 2001

[MS96] Mozer, M.C. ; Sitton, M.: Computational modeling of spatial attention. In: Pashler, H. (Hrsg.): *Attention*. UCL Press, 1996

[MS99] Mertsching, B. ; Schmalz, S.: Active Vision Systems. In: Jähne, B. (Hrsg.) ; Haussecker, H. (Hrsg.) ; Geissler, P. (Hrsg.): *Handbook of Computer Vision and Applications* Bd. 3. Academic Press, 1999, S. 197–219

[MSN00] McPeek, R.M. ; Skavenski, A.A. ; Nakayama, K.: Concurrent processing of saccades in visual search. In: *Vision Research* 40 (2000), August, Nr. 18, S. 2499–2516

[MUE96] Maki, A. ; Uhlin, T. ; Eklundh, J.-O.: A direct disparity estimation technique for depth segmentation. In: *Proc. 5th IAPR Workshop on Machine Vision Applications*, 1996, S. 530–533

[Mun66] Munsell, A.E.O.: *Munsell Book of Color*. Baltimore, MD : Munsell Color Co., 1929-1966

[MW94] McKee, S.P. ; Watamaniuk, S.N.J.: The psychophysics of motion perception. In: Smith, A.T. (Hrsg.) ; Snowden, R.J (Hrsg.): *Visual Detection of Motion*. Academic Press, 1994

[MW01] Moore, C.M. ; Wolfe, J.M.: Getting beyond the serial/parallel debate in visual search: a hybrid approach. In: Shapiro, K. (Hrsg.): *The limits of attention*. Oxford University Press, 2001, S. 178–198

[MWG+94] Milanese, R. ; Wechsler, H. ; Gil, S. ; Bost, J.M. ; Pun, T.: Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In: *Proceedings, of the IEEE Conference on Computer Vision and Pattern Recognition (Seattle, 1994)*, 1994, S. 781–785

[MY88] Miyahara, M. ; Yoshida, Y.: Mathematical Transform of (R, G, B) Color Data to Munsell (H, V, C) Color Data. In: *Visual Communication and Image Processing* 1001 (1988), S. 650–657

[MZ01] Mitchell, J. ; Zipser, D.: A model of visual-spatial memory across saccades. In: *Vision Research* 41 (2001), S. 1575–1592

[Nac02] Nachtigall, W.: *Bionik*. 2. Berlin : Springer, 2002

[Nak96] Nakayama, K.: Binocular visual surface perception. In: *Proceedings of the National Academy of Sciences, USA* 93 (1996), S. 634–639

[Nei67] Neisser, U.: *Cognitive Psychology*. Appleton-Century-Crofts, New York, 1967

[NKR93] Niebur, E. ; Koch, C. ; Rosin, C.: An oscillation-based model for the neuronal basis of attention. In: *Vision Research* 18 (1993), S. 2789–2802

[NS86] Nakayama, K. ; Silverman, G. H.: Serial and parallel processing of visual feature conjunctions. In: *Nature* 320 (1986), March, Nr. 6059

[OAV95] Olshausen, B.A. ; Anderson, C.H. ; Van Essen, D.C.: A Multiscale Dynamic Routing Circuit for Forming Size- and Position-Invariant Object Representations. In: *Journal of Computational Neuroscience* 2 (1995), Nr. 1, S. 45–62

[ODCR00] O'Regan, J.K. ; Deubel, H. ; Clark, J.J. ; Rensink, R.A.: Picture Changes During Blinks: Looking Without Seeing and Seeing Without Looking. In: *Visual Cognition* 7 (2000), S. 191–211

[ODR+99] Owen, G.S. ; Domik, G. ; Rhyne, T.-M. ; Brodlie, K.W. ; Santos, B.S. Hypervis - teaching scientific visualization using hypermedia. http://www.siggraph.org/education/materials/HyperVis/hypervis.htm. 1999

[OEA93] Olshausen, B.A. ; Essen, D.C. V. ; Anderson, C.H.: A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. In: *Journal of Neuroscience* 13 (1993), S. 4700–4719

[OH00] Ouerhani, N. ; Hügli, H.: Computing Visual Attention from Scene Depth. In: *Proceedings of the 15th International Conference on Pattern Recognition, ICPR 2000* Bd. 1, IEEE Computer Society Press, September 2000, S. 375–378

[Ost35] Osterberg, G.: Topography of the layer of rods and cones in the human retina. In: *Acta Ophtalmologicy* 6 (1935), S. 1–102

[PAH+00] Paulus, D. ; Ahlrichs, U. ; Heigl, B. ; Denzler, J. ; Hornegger, J. ; Zobel, M. ; Niemann, H.: Active Knowledge-Based Scene Analysis. In: *Videre: Journal of Computer Vision Research* 1 (2000), Nr. 4, S. 5–29

[Pas93] Pashler, H.: Dual task interference and elementary mental mechanisms. In: Meyer, D. (Hrsg.) ; Kornblum, S. (Hrsg.): *Attention and Performance* Bd. 14. Cambridge, MA : MIT Press, 1993, S. 245–264

[Pas98] Pashler, H.E.: *The Psychology of Attention*. MIT Press, 1998

[PBF+94] Pylyshyn, Z.W. ; Burkell, J. ; Fisher, B. ; Sears, C.R. ; Schmidt, W. ; Trick, L.: Multiple parallel access in visual attention. In: *Canadian Journal of Experimental Psychology* 48 (1994), Nr. 2, S. 260–283

[PC84] Posner, M.I. ; Cohen, Y.: Components of visual orienting. In: Bouma, H. (Hrsg.) ; Bouwhuis, D. (Hrsg.): *Attention and Performance X*. London: Erlbaum, 1984, S. 531–556

[PDR+00] Paulus, D. ; Drexler, C. ; Reinhold, M. ; Zobel, M. ; Denzler, J.: Active Computer Vision System. In: Cantoni, V. (Hrsg.) ; Guerra, C. (Hrsg.): *Computer Architectures for Machine Perception*. Los Alamitos, CA : IEEE Computer Society, 2000, S. 18–27

[PE99] Pessoa, L. ; Exel, S.: Attentional Strategies for Object Recognition. In: Mira, J. (Hrsg.) ; Sachez-Andres, J.V. (Hrsg.): *Proceedings of the IWANN, Alicante, Spain 1999* Bd. 1606, Springer, 1999, S. 850–859

[Pit00] Pitas, I.: *Digital Image Processing Algorithms and Applications*. John Wiley and Sons, 2000

[PKE98] Pauly, M. ; Kopecz, K. ; Eckhorn, R.: Model of a Fixation Control Network Performs Saccades, Smooth Pursuit, and Provides the Basis for Segmentation of Unknown Objects in Dynamic Real-World Scenes. In: *Proceedings Workshop Dynamische Perzeption*, 1998

[PKE99] Pauly, M. ; Kopecz, K. ; Eckhorn, R.: Gaze control with neural networks: A unified approach for saccades and smooth pursuit. In: Mira, J. (Hrsg.) ; Sanchez-Andres, J. (Hrsg.): *Engeneering Applications of Bio-Inspired artificial neural networks - Proceedings of the IWANN '99 Conference* Bd. 1, Springer, 1999, S. 113–122

[PKL+94] Priese, L. ; Klieber, J. ; Lakmann, R. ; Rehrmann, V. ; Schian, R.: New Results on Traffic Sign Recognition. In: *Intelligent Vehicles Symposium 1994, Paris*, 1994, S. 249–254

[PLN02] Parkhurst, D. ; Law, K. ; Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. In: *Vision Research* 42 (2002), Nr. 1, S. 107–123

[Pos80] Posner, M.I.: Orienting of attention. In: *Quarterly Journal of Experimental Psychology* 32 (1980), S. 3–25

[Pos94] Postma, E.O.: *SCAN: A Neural Model of Covert Attention*, Rijksuniversiteit Limburg, Wageningen, Dissertation, 1994

[PR83] Pollen, D.A. ; Ronner, S.F.: Visual cotrical neurons as localized spatial frequency filters. In: *IEEE Transactions on Systems, Man, and Cybernetics* 13 (1983), S. 907–916

[PRSW00] Pomplun, M. ; Reingold, E.M. ; Shen, J. ; Williams, D.E.: The Area Activation Model of Saccadic Selectivity in Visual Search. In: Gleitman, L.R. (Hrsg.) ; Joshi, A.K. (Hrsg.): *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 2000, S. 375–380

[PS88] Pylyshyn, Z. W. ; Storm, R.W.: Tracking multiple independent targets: evidence for a parallel tracking mechanism. In: *Spatial Vision* 3 (1988), Nr. 3

[PSD80] Posner, M.I. ; Snyder, C.R. ; Davidson, B.J.: Attention and the detection of signals. In: *Journal of Experimental Psychology* 109 (1980), Nr. 2, S. 160–174

[Pyl98] Pylyshyn, Z.W.: Visual Indexes in Spatial Vision and Imagery. In: Wright, R.D. (Hrsg.): *Visual Attention*. Oxford University Press, 1998 (Vancouver Studies in Cognitive Science 8), S. 215–231

[Pyl99] Pylyshyn, Z.W.: Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. In: *Behavioral and Brain Sciences* 22 (1999), Nr. 3, S. 341–423

[Pyl00] Pylyshyn, Z.W.: Situating Vision in the World. In: *Trends in Cognitive Sciences* 4 (2000), Nr. 5

[Ras02] Rasmussen, C.: Combining Laser Range, Color, and Texture Cues for Autonomous Road Following. In: *IEEE International Conference on Robotics and Automation, ICRA-02*, 2002

[Rat95] Ratan, A.L.: The Role of Fixation and Visual Attention in Object Recognition / MIT, Artificial Intelligence Laboratory. 1995 (1529). – Forschungsbericht. A.I. Technical Report

[Ray98] Rayner, K.: Eye movements in reading and information processing. Twenty years of research. In: *Psychological Bulletin* 124 (1998), S. 372–422

[Ray01] Raymond, J.E.: Perceptual links and attentional blinks. In: Shapiro, K. (Hrsg.): *The limits of attention*. Oxford University Press, 2001, S. 217–228

[RB95] Rao, R. ; Ballard, D.H.: An Active Vision Architecture based on Iconic Representations. In: *Artificial Intelligence* 78 (1995), S. 461–505

[Ren02] Rensink, R.A.: Change Detection. In: *Annual Review of Psychology* 53 (2002), S. 245–277

[RJ96] Rentschler, I. ; Jüttner, M.: Foveales und extrafoveales Formensehen - wie verschiebungsinvariant ist die Mustererkennung im Gesichtsfeld? In: Mertsching, B. (Hrsg.): *Aktives Sehen in technischen und biologischen Systemen*. Sankt Augustin : infix, 1996, S. 15–22

[ROC97] Rensink, R.A. ; O'Regan, J.K. ; Clark, J.J.: To see or not to see: The need for attention to perceive changes in scenes. In: *Psychological Science* 5 (1997), S. 368–373

[Roc98] Rock, I.: *Wahrnehmung - Vom visuellen Reiz zum Sehen und Erkennen*. Spektrum, Akad. Verlag, 1998

[Roy81] Royer, F.: Detection of symmetry. In: *Journal of Experimental Psychology: Human Perception and Performance* 7 (1981), Nr. 6, S. 1186–1210

[RR00] Rivlin, E. ; Rotstein, H.: Control of a Camera for Active Vision: Foveal Vision, Smooth Tracking and Saccade. In: *International Journal of Computer Vision* 39 (2000), Nr. 2, S. 81–96

[RS95] Reece, D.A. ; Shafer, S.A.: Control of perceptual attention in robot driving. In: *Artificial Intelligence* 78 (1995), S. 397–430

[RSA92] Raymond, J.E. ; Shapiro, K.L. ; Arnell, K.M.: Temporary supression of visual processing in an RSVP task: an attentional blink. In: *Journal of Experimental Psychology: Human Perception and Performance* 18 (1992), S. 849–860

[RV79] Rovamo, J. ; Virsu, V.: An estimation and application of the human cortical magnification factor. In: *Experimental Brain Research* 37 (1979), S. 1–20

[RWY95] Reisfeld, D. ; Wolfson, H. ; Yeshurun, Y.: Context Free Attentional Operators: The Generalized Symmetry Transform. In: *International Journal of Computer Vision* (1995)

[Sai03] Saiki, J.: Feature binding in object-file representations of multiple moving items. In: *Journal of Vision* 3 (2003), S. 6–21

[San88] Sanger, T.D.: Stereo Disparity Computation Using Gabor Filters. In: *Biological Cybernetics* 59 (1988), S. 405–418

[SB93] Sarkar, S. ; Boyer, K.: Perceptual organization in computer vision: A review and a proposal for a classificatory structure. In: *IEEE Transactions On Systems, Man, and Cybernetics* 23 (1993), Nr. 2, S. 382–399

[SB94] Sarkar, S. ; Boyer, K.: *Computing Perceptual Organization in Computer Vision*. World Scientific, 1994

[SB98] Salgian, G. ; Ballard, D.H.: Visual Routines for Vehicle Control. In: Kriegman, D. (Hrsg.) ; Hager, G. (Hrsg.) ; Morse, S. (Hrsg.): *The Confluence of Vision and Control*. Springer, 1998

[SBDH00] Stemmer, R. ; Brockers, R. ; Drüe, S. ; Hartmann, G.: Erkennung und Vermessung komplexer Demontageobjekte. In: Baratoff, G. (Hrsg.) ; Neumann, H. (Hrsg.): *Dynamische Perzeption*, infix, 2000 (Proceedings in Artificial Intelligence), S. 147–152

[SBSJ71] Sperling, G. ; Budiansky, J. ; Spivak, J.G. ; Johnson, M.C.: Extremely rapid visual search: The maximum rate of scanning letters for the presence of a numeral. In: *Science* 174 (1971), S. 307–311

[Sch99] Schneider, W.X.: Visual-spatial working memory, attention, and scene-representation: A neuro-cognitive theory. In: *Psychological Research* 62 (1999), S. 220–236

[Sch00] Schmalz, S.: *Entwurf und Evaluierung von Strategien zur 2D/3D-Objekterkennung in aktiven Sehsystemen*, FB Informatik, Universität Hamburg, Dissertation, 2000

[SDE96] Schöner, G. ; Dose, M. ; Engels, C.: Dynamics of behavior: Theory and applications for autonomous robot architectures. In: *Robotics and Autonomous Systems* 16 (1996), S. 213–245

[SF03] Sun, Y. ; Fisher, R.: Object-based Visual Attention for Computer Vision. In: *Artificial Vision* 146 (2003), Nr. 1, S. 77–123

[SFP98] Schmidt, W.C. ; Fisher, B.D. ; Pylyshyn, Z.W.: Multiple location access in vision: Evidence from a line-motion illusion. In: *Journal of Experimental Psychology: Human Perception and Performance* 24 (1998), Nr. 2, S. 505–525

[SG95] Singer, W. ; Gray, C.M.: Visual feature integration and the temporal correlation hypothesis. In: *Annual Review of Neuroscience* 18 (1995), S. 555–586

[SJ91] Saarinen, J. ; Julesz, B.: The speed of attentional shifts in the visual field. In: *Proc. Natl. Acad. Sci.* 88 (1991), March, S. 1812–1814

[SK01] Shor, R. ; Kiryati, N.: Towards Segmentation from Multiple Cues: Symmetry and Color. In: al., R. K. (Hrsg.): *Multi-Image Analysis* Bd. 2032, 2001, S. 145–152

[SKC00] Stasse, O. ; Kuniyoshi, Y. ; Cheng, G.: Development of a Biologically Inspired Real-Time Visual Attention System. In: *BMCV 2000*, 2000, S. 150–159

[SL97] Simons, D.J. ; Levin, D.T.: Change Blindness. In: *Trends in Cognitive Sciences* 1 (1997), Nr. 7, S. 261–267

[SMB00] Schmid, C. ; Mohr, R. ; Bauckhage, C.: Evaluation of Interest Point Detectors. In: *International Journal of Computer Vision* 37 (2000), Nr. 2, S. 151–172

[SP75] Smith, C. ; Pokorny, J.: Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm. In: *Vision Research* 15 (1975), S. 161–171

[SP99] Scholl, B.J. ; Pylyshyn, Z.W.: Tracking Multiple Items Through Occlustion: Clues to Visual Objecthood. In: *Cognitive Psychology* 38 (1999), S. 259–290

[SP00] Sears, C.R. ; Pylyshyn, Z.W.: Multiple Object Tracking and Attentional Processing. In: *Canadian Journal of Experimental Psychology* 54 (2000), Nr. 1, S. 1–14

[Spe60] Sperling, G.: The information available in brief visual presentations. In: *Psychological Monographs: General and Applied* 74 (1960), Nr. 498, S. 1–29

[SS77] Shiffrin, R.M. ; Schneider, W.: Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. In: *Psychological Review* 84 (1977), S. 127–190

[SS96] Shih, S. ; Sperling, G.: Is there feature-based attentional selection in visual search. In: *Journal of Experimental Psychology: Human Perception and Performance* 22 (1996), S. 758–779

[SSP+99] Sathian, K. ; Simon, T.J. ; Peterson, S. ; Patel, G.A. ; Hoffman, J.M. ; Grafton, S.T.: Neural Evidence Linking Visual Object Enumeration and Attention. In: *Journal of Cognitive Neuroscience* 11 (1999), Nr. 1, S. 36–51

[Str35] Stroop, J.R.: Studies of interference in serial verbal reactions. In: *Journal of Experimental Psychology* 18 (1935), S. 643–662

[Sty97] Styles, E.A.: *The Psychology of Attention*. Psychology Press, 1997

[SVT98] Sharma, G. ; Vrhel, M.J. ; Trussell, H.J.: Color Imaging for Multimedia. In: *Proceedings of the IEEE* 86 (1998), Nr. 6, S. 1088–1108

[SW95] Sperling, G. ; Weichselgartner, E.: Episodic Theory of the Dynamics of Spatial Attention. In: *Psychological Review* 102 (1995), Nr. 3, S. 503–532

[TA79] Takeuchi, A. ; Amari, S.-I.: Formation of topographic maps in columnar microstructures in nerve fields. In: *Biological Cybernetics* 35 (1979), S. 63–72

[TAK98] Theeuwes, J. ; Atchley, P. ; Kramer, A.F.: Attentional Control Within 3-D Space. In: *Journal of Experimental Psychology: Human Perception and Performance* 24 (1998), Nr. 5, S. 1476–1485

[TCW+95] Tsotsos, J. K. ; Culhane, S.M. ; Wai, W.Y.K. ; Lai, Y.H. ; Davis, N. ; Nuflo, F.: Modelling visual attention via selective tuning. In: *Artificial Intelligence* 78 (1995), 10, Nr. 1-2, S. 507–545

[TDM95] Trapp, R. ; Drüe, S. ; Mertsching, B.: Korrespondenz in der Stereoskopie bei räumlich verteilten Merkmalsrepräsentationen im Neuronalen Active-Vision-System NAVIS. In: Sagerer, G. (Hrsg.) ; Posch, S. (Hrsg.) ; Kummert, F. (Hrsg.): *Mustererkennung 1995*, 1995, S. 492–499

[TDW91] Tipper, S. P. ; Driver, J. ; Weaver, B.: Object-centered Inhibition of Return of Visual Attention. In: *Quarterly Journal of Experimental Psychology* 43A (1991), May, Nr. 2, S. 289–298

[TE97] Trick, L.M. ; Enns, J.T.: Clusters precede Shapes in Perceptual Organization. In: *Psychological Science* 8 (1997)

[Ter97] Terzopoulos, D.: Animat Vision: Active Vision in Artificial Animals. In: *Videre: Journal of Computer Vision Research* 1 (1997), Nr. 1

[TG80] Treisman, A. ; Gelade, G.: A feature integration theory of attention. In: *Cognitive Psychology* 12 (1980), S. 97–136

[The93] Theeuwes, J.: Visual selective attention: A theoretical analysis. In: *Acta Psychologica* 83 (1993), S. 93–154

[TK94] Theeuwes, J. ; Kooi, J.L.: Parallel search for a conjunction of shape and contrast polarity. In: *Vision Research* 34 (1994), Nr. 22, S. 3013–3016

[TLB92] Tipper, S. P. ; Lortie, C. ; Baylis, G.C.: Selective Reaching: Evidence for Action-Centered Attention. In: *Journal of Experimental Psychology: Human Perception and Performance* 18 (1992), Nr. 4, S. 891–905

[TM94] Theimer, W. ; Mallot, H.A.: Phase-based binocular vergence control and depth reconstruction using active vision. In: *CVGIP: Image Understanding* 60 (1994), Nr. 3, S. 343–358

[Tov96] Tovee, M.J.: *An introduction to the visual system*. Cambridge University Press, 1996

[TP94] Trick, L.M. ; Pylyshyn, Z.W.: Why Are Small and Large Numbers Enumerated Differently? A Limited-Capacity Preattentive Stage in Vision. In: *Psychological Review* 101 (1994), Nr. 1, S. 80–102

[Tra96] Trapp, R.: Entwurf einer Filterbank auf der Basis neurophysiologischer Erkenntnisse zur orientierungs- und frequenzselektiven Dekomposition von Bilddaten / Heinz-Nixdorf Institut und Universität-GH Paderborn. 1996. – Forschungsbericht

[Tre60] Treisman, A.: Contextual cues in selective listening. In: *Quarterly Journal of Experimental Psychology* 12 (1960), S. 242–248

[Tre91] Treisman, A.: Representing visual objects. In: Meyer, D. (Hrsg.) ; Kornblum, S. (Hrsg.): *Attention and Performance* Bd. 14. Hillsdale, NJ: Erlbaum, 1991

[Tre93] Treisman, A.: The perception of features and objects. In: Baddeley, A. (Hrsg.) ; Weiskrantz, L. (Hrsg.): *Attention: Selection, awareness, and control*. Oxford : Clarendon Press, 1993, S. 5–35

[Tre98] Treisman, A.: The Perception of Features and Objects. In: Wright, R.D. (Hrsg.): *Visual Attention*. Oxford University Press, 1998, S. 26–54

[TS90] Treisman, A. ; Sato, S.: Conjunction search revisited. In: *Journal of Experimental Psychology: Human Perception and Performance* 16 (1990), Nr. 3, S. 459–478

[Tso93] Tsotsos, J. K.: An inhibitory beam for attentional selection. In: Harris, L. (Hrsg.) ; Jenkins, M. (Hrsg.): *Spatial visions in humans and robots*. 1993

[TW98a] Takacs, B. ; Wechsler, H.: A Dynamic and Multiresolution Model of Visual Attention and Its Application to Facial Landmark Detection. In: *Computer Vision and Image Understanding* 70 (1998), Nr. 1, S. 63–73

[TW98b] Tipper, S.P. ; Weaver, B.: The Medium of Attention: Location-based, Object-Centered, or Scene-bases. In: Wright, R.D. (Hrsg.): *Visual Attention*. Oxford University Press, 1998, S. 77–107

[TYN97] Tani, J. ; Yamamoto, J. ; Nishi, H.: Dynamical Interactions between Learning, Visual Attention, and Behavior: An Experiment with a Vision-Based Mobile Robot. In: *Fourth European Conference on Artificial Life*, MIT Press, 1997, S. 309–317

[Ull84] Ullman, S.: Visual Routines. In: *Cognition* 18 (1984), S. 97–159

[UM82] Ungerleider, L.G. ; Mishkin, M.: Two cortical visual systems. In: Ingle, D.J. (Hrsg.) ; Goodale, M.A. (Hrsg.) ; Mansfield, R.J.W. (Hrsg.): *Analysis of visual behavior*. MIT Press, 1982, S. 549–586

[VB97] Vecera, S.P. ; Behrmann, M.: Spatial Attention Does Not Require Preattentive Grouping. In: *Neuropsychology* 11 (1997), Nr. 1, S. 30–43

[VCL00] Verstraten, F.A.J. ; Cavanagh, P. ; Labianca, A.T.: Limits of attentive tracking reveal temporal properties of attention. In: *Vision Research* 40 (2000), S. 3651–3664

[Vei97] Veit, D.: *Existenz und Stabilität von lokalen Anregungen in homogenen neuronalen Feldern*, Seminararbeit, 1997

[VF94] Vecera, S.P. ; Farah, M.: Does visual attention select objects or locations. In: *Journal of Experimental Psychology: General* 123 (1994), S. 146–160

[VM98] Viswanathan, L. ; Mingolla, E.: Attention in Depth: Disparity and Occlusion Cues facilitate Multi-Element Visual Tracking / Boston University. 1998 (CAS/CNS-TR-98-012). – Forschungsbericht

[VM99] Viswanathan, L. ; Mingolla, E.: Dynamics of Attention in Depth: Evidence from Multi-element Tracking / Boston University. 1999 (CAS/CNS-TR-99-010). – Forschungsbericht

[VM01] Voss, N. ; Mertsching, B.: Design and Implementation of an Accelerated Gabor Filter Bank Using Parallel Hardware. In: *11th International Conference on Field-Programmable Logic and Applications (FPL 2001), Belfast*, 2001

[VS91] Vincent, L. ; Soille, P.: Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (1991), Nr. 6, S. 583–598

[Wan95] Wandell, B.A.: *Foundations of Vision*. Sunderland MA (Sinauer Associates), 1995

[Wan99] Wang, D.L.: Object selection based on oscillatory correlation. In: *Neural Networks* 12 (1999), S. 579–592

[War01] Ward, R.: Visual attention moves no faster than the eyes. In: Shapiro, K. (Hrsg.): *The limits of attention*. Oxford University Press, 2001, S. 199–216

[Was99] Wasson, G.: *Design of Representation Systems for Autonomous Agents*, University of Virginia, Dissertation, 1999

[WB97] Wolfe, J.M. ; Bennett, S.C.: Preattentive Object Files: Shapeless Bundles of Basic Features. In: *Vision Research* 37 (1997), S. 25–43

[WCF89] Wolfe, J.M. ; Cave, K. R. ; Franzel, S. L.: Guided Search: An alternative to the Feature Integration model for visual search. In: *Journal of Experimental Psychology: Human Perception and Performance* 15 (1989), S. 419–433

[Wer23] Wertheimer, M.: Untersuchungen zur Lehre von der Gestalt. In: *Psychologische Forschung* (1923), Nr. 4, S. 301–350

[WG96] Wolfe, J.M. ; Gancarz, G.: Guided Search 3.0: A Model of Visual Search Catches Up With Jay Enoch 40 Years Later. In: Lakshminarayanan, V. (Hrsg.): *Basic and Clinical Applications of Vision Science*. Kluwer Academic, 1996, S. 189–192

[Wil98] Wilhelm, T.: *Untersuchungen zu neuronalen Felddynamiken*, Technische Universität Ilmenau, Studienarbeit, 1998

[Wol94] Wolfe, J.M.: Guided Search 2.0: A revised model of visual search. In: *Psychonomic Bulletin and Review* 1 (1994), Nr. 2, S. 202–238

[Wol96] Wolfe, J.M.: Visual Search. In: Pashler, H. (Hrsg.): *Attention*. University College London Press, 1996

[Wol98] Wolfe, J.M.: What can 1,000,000 trials tell us about visual search? In: *Psychological Science* 9 (1998), Nr. 1

[Wol00] Wolfe, J.M.: Visual Attention. In: DeValois (Hrsg.): *Seeing*. 2nd ed. Academic Press, 2000, S. 335–386

[WP99] Wagner, T. ; Plankensteiner, P.: Industrial Object Recognition. In: Jähne, B. (Hrsg.) ; Haussecker, H. (Hrsg.) ; Geissler, P. (Hrsg.): *Handbook of Computer Vision and Applications* Bd. 3. Academic Press, 1999, S. 297–314

[WT02] Wheeler, M.E. ; Treisman, A.: Binding in short-term visual memory. In: *Journal of Experimental Psychology: General* 131 (2002), S. 48–64

[Xu02] Xu, Y.: Limitations of object-based feature encoding in visual short-term memory. In: *Journal of Experimental Psychology: Human Perception and Performance* 28 (2002), S. 458–468

[Yes97] Yeshurun, Y.: Attentional Mechanisms in Compter Vision. In: Contoni, V. (Hrsg.) ; Levialdi, S. (Hrsg.) ; Roberto, V. (Hrsg.): *Artificial Vision: Image Description, Recognition and Communication*. Academic Press, 1997, S. 43–52

[YKLH01] Yoon, K.J. ; Kweon, I.S. ; Lee, C.H. ; Hur, J.S.: Landmark design and real-time landmark tracking using color histogram for mobile robot localization. In: *32th International Symposium on Robotics*, 2001

[You95] Young, M.P.: Open questions about the neural mechanisms of visual pattern recognition. In: Gazzaniga, M.S. (Hrsg.): *The cognitive neurosciences*. Cambridge, MA : MIT Press, 1995, S. 463–474

[YYL96] Yamamoto, H. ; Yeshurun, Y. ; Levine, M.D.: An Active Foveated Vision System: Attentional Mechanisms and Scan Path Convergence Measures. In: *Computer Vision and Image Understanding* 63 (1996), Nr. 1, S. 50–65

[Zab90] Zabrodsky, H.: Symmetry - A Review / Department of Computer Science, Hebrew University of Jerusalem. 1990 (90-16). – Forschungsbericht

[Zek93] Zeki, S.: *A vision of the Brain*. Oxford : Blackwell Scientific, 1993

[Zuc87] Zucker, S.W.: The diversity of perceptual grouping. In: Arbib, M.A. (Hrsg.) ; Hanson, A.R. (Hrsg.): *Vision, Brain and Cooperative Computation*. MIT Press, 1987

Index