Technische Universität Ilmenau

Fachgebiet Neuroinformatik

98684 Ilmenau (Thüringen)
PF. 100565

Projekt GESTIK
Gestenbasierte Interaktion und Kommunikation mit dem mobilen Robotersystem MILVA

Sachbericht gemäß Nr.3.2 BNBest-P-TMWFK

Autoren: Dr. Hans-Joachim Böhme
Markus Krabbes
Anja Brakensiek
Ulf-Dietrich Braumann
Andrea Corradini
Prof. Dr. Horst-Michael Gross

Stand: 21. Dezember 1998

Zuwendungsempfänger: Prof. Dr.-Ing. H.-M. Gross	Förderkennzeichen: B 601-95037
Vorhabensbezeichnung: Projekt GESTIK	
Laufzeit des Vorhabens: 1. Juli 1996 bis: 30. Juni 1998	

Inhaltsverzeichnis

1 Aufgabenstellung

Die Steuerung autonom agierender, mobiler Robotersysteme stellt eine international und zunehmend auch national stark forcierte Forschungsrichtung mit breiten Anwendungsfeldern in der Industrieautomatisierung und im Dienstleistungssektor dar. Gerade im Dienstleistungssektor wird in jüngster Zeit unter dem Begriff "Service-Roboter" an einer neuen Generation von Systemen gearbeitet, die alltägliche Dienstleistungen und Hilfestellungen in *unmittelbarer Kooperation und Interaktion mit dem Menschen* vollführen sollen. An die Leistungsfähigkeit der hierzu benötigten Steuerungs- und Sensorsysteme sowie an die Intelligenz und Flexibilität der "Nutzer-Roboter-Schnittstellen" werden allerdings extrem hohe Anforderungen gestellt, insbesondere was die Fähigkeit zur Kommunikation und Interaktion zwischen Nutzer und Serviceroboter und damit die Akzeptanz durch den Nutzer betrifft.

Die Ziele dieses Forschungsvorhabens resultierten in zwei parallel bearbeiteten Schwerpunktthemen:

(a) Entwicklung lernfähiger neuronaler Architekturen zur non-verbalen, visuell basierten Kommunikation zwischen menschlichem Nutzer und Roboterplattform MILVA (Abbildung 1).

(b) Entwicklung neuronaler Verfahren zur visuell geführten lokalen Roboternavigation und zur handlungsrelevanten Gesteninterpretation.

Abbildung 1: *MILVA - Multisensorielles Intelligentes Lernfähiges Vehikel in neuronaler Architektur*

Das folgende Szenario soll die im Projekt angestrebte finale Funktionalität der Roboterplattform verdeutlichen: In der aktuellen Position erfolgt eine aktive visuelle Beobachtung des Einsatzfeldes mit dem Ziel, mögliche vorhandene Personen und deren Zuwendung zur Plattform (Interesse an Interaktion) zu detektieren. Ohne weiter spezifizierte gestikulierte Navigationsanweisungen nähert sich MILVA dem detektierten potentiellen Nutzer an. Dabei erfolgt kontinuierlich eine Auswertung der vom Nutzer eingenommenen Körperstellung mit dem Ziel, vorab fest vereinbarte statische Posturen (siehe Abbildung 2) zu erkennen und diese in ein entsprechendes Fahrverhalten umzusetzen.

Abbildung 2: *Posturenalphabet: 'hallo', 'halt', 'fahre nach links', 'fahre nach rechts', 'komm links neben mich', 'komm rechts neben mich'*

1.1 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Vorhaben wurde nach der „Richtlinie für die Förderung der Verbesserung der Drittmittelfähigkeit wissenschaftlicher Forschungseinrichtungen in Thüringen“ (Richtlinie 1) beantragt und durchgeführt.

Als Forschungsplattform für das Vorhaben diente das aus Landesmitteln finanzierte, mobile Robotersystem MILVA mit on-board Neurocomputer und multi-sensorischer Datengewinnung. Neben dem MILVA-System wurden ein stationäres Kamerasystem sowie der Miniaturroboter Khepera zur Methodenentwicklung eingesetzt.

1.2 Planung und Ablauf des Vorhabens

Das Vorhaben wurde entsprechend dem im Antrag zum Projekt „GESTIK“ vorgesehenen Arbeitsprogramm bearbeitet. Naturgemäß ergaben sich während der Bearbeitung eine Reihe von inhaltlichen Präzisierungen der vorgegebenen Zielstellung, auf die in den Abschnitten 2 und 2.5 noch gesondert eingegangen wird. Diese Präzisierungen hatten jedoch keine prinzipielle Änderung der Projektzielstellung zur Folge.

1.3 Anknüpfung an den wissenschaftlichen und technischen Stand

Der zum Zeitpunkt der Antragstellung bekannte Stand projektrelevanter wissenschaftlicher Erkenntnisse wurde im Antragstext dargelegt. Bezüge zu während der Projektbearbeitung bekanntgewordenen projektrelevanten Forschungsergebnissen Dritter enthält Abschnitt 2.

1.4 Zusammenarbeit mit anderen Stellen

Das GESTIK-Projekt war ein ausschließlich am FG Neuroinformatik der TU Ilmenau bearbeitetes Forschungsvorhaben. Während der Projektdauer wurden kontinuierlich die relevanten Ergebnisse der mit dem FG kooperierenden Partner (Uni Bochum, Uni Bielefeld, Uni Dortmund, CMU Pittsburgh) sowie aktuelle publizierte Beiträge aus der internationalen Literatur berücksichtigt. Insbesondere genannt werden soll an dieser Stelle der fruchtbringende Erfahrungsaustausch während der nationalen bzw. internationalen Konferenzen und Workshops, die größtenteils mit eigenen Beiträgen beschickt wurden, sowie der vom FG selbst veranstaltete SOAVE'97-Workshop, der aufgrund der sehr guten Resonanz 1999 wiederholt werden soll.

2 Ausführliche Darstellung

Die ausführliche Darstellung der erreichten methodischen Ergebnisse behandelt in zwei separaten Unterabschnitten die beiden definierten Teilthemen des Projektes. Auf die Integration der beiden Teilprojekte bzw. auf offene bzw. weiterführende Arbeiten und aufgetretene Probleme wird in den entsprechenden inhaltlichen Passagen bzw. im Abschnitt 2.5 eingegangen.

Den letzten Stand der Arbeiten, der teilweise schon über das GESTIK-Projekt hinausgeht, dokumentieren zwei Publikationen, die als Anlage diesem Abschlußbericht beigefügt sind.

Neben der Methodenentwicklung lag ein weiterer Projektschwerpunkt auf der Überarbeitung bzw. Neukonzeption der Hardware-Architektur der Plattform, deren technischer Realisierung sowie der notwendigen softwaretechnischen Umstellung. Folgende Aspekte sollen in diesem Zusammenhang hervorgehoben werden:

- kompletter Neuentwurf der informationsverarbeitenden Hardware, bestehend aus 2 PC-Systemen und CNAPS-Neurocomputer-Integration bei Beibehaltung der VME-basierten Antriebssteuerung
- wesentliche Erweiterung der Sensor-Ausstattung um Doppelkamera-Active-Vision-System mit 5 Freiheitsgraden (2 PTU's plus drehbarer „Hals“), Frontkamera zur Navigation und Doppelmikrofon-Anordnung
- konsistentes LINUX-Softwarekonzept incl. der Entwicklung einer hardware-angepaßten grafischen Entwicklungsumgebung GRASIM (siehe auch Abschnitt 2.2.2)

2.1 Erzielte wissenschaftliche Ergebnisse zum Teilthema 1: Neuronale Architekturen zur gestik-basierten Nutzer-Roboter-Kommunikation

Für eine benutzerfreundliche Mensch-Maschine-Interaktion spielen *natürliche* Kommunikationsformen und Eingabemedien, wie Sprache und Gestik, eine immer größere Rolle (vgl. Ansätze in (Darrell, T. et al., 1997, Kortenkamp, D. et al., 1996)). Das Ziel des GESTIK-Projektes lag deshalb in der Entwicklung einer neuronalen Architektur zur visuellen, gestenbasierten Interaktion eines Nutzers mit unserem mobilen Robotersystem MILVA[1] in Realwelt-Umgebungen (siehe z.B. (Boehme, H.-J. et al., 1998a, Boehme, H.-J. et al., 1998b)). Die dabei notwendigen Teilleistungen und die daraus abgeleitete neuronale Architektur verdeutlicht Abbildung 3.

[1] http://cortex.informatik.tu-ilmenau.de/technik.html

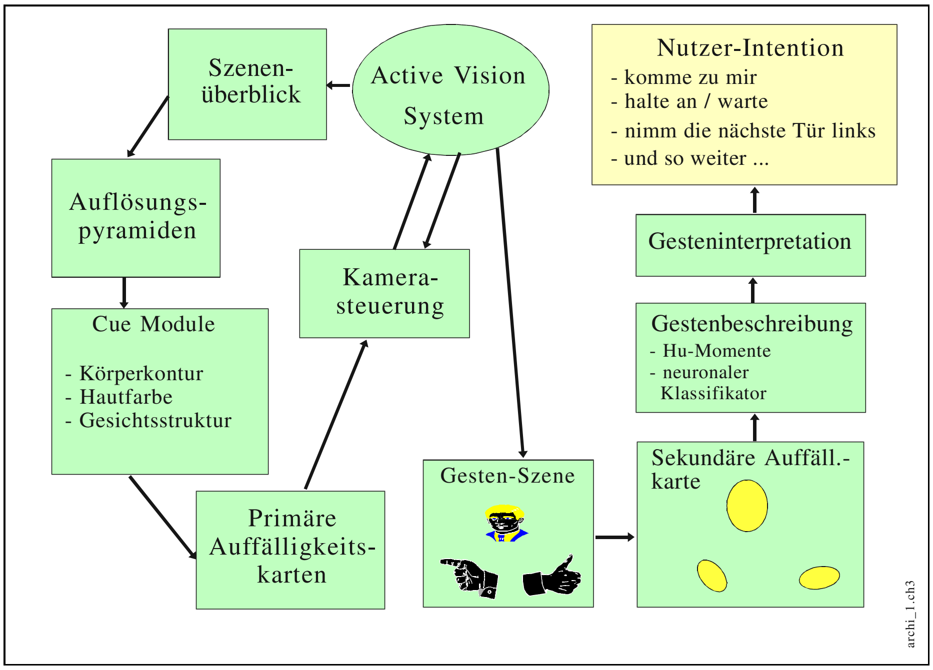

Abbildung 3: *Gesamtarchitektur zum Teilthema 1*

Das Doppelkamera-Active-Vision-System bildet den zentralen Ausgangspunkt. Zunächst erfolgt im Szenenüberblick die Lokalisation von Personen bzw. einer Person. Dazu wird das Bild zunächst in eine Multiskalen-Repräsentation überführt. Auf jeder Auflösungsstufe erfolgt die Suche nach personenspezifischen Merkmalen. Das Ergebnis der Merkmalsdetektion bilden die primären Auffälligkeitskarten, die die wahrscheinlichste Position einer Person bzw. mehrerer Personen im Raum liefern. Auf die der Plattform naheste Person richtet sich jetzt die zweite Kamera und erfaßt dabei einen Bildbereich, der diese Person etwa in Normgröße enthält und die Auswertung der gestikulierten Navigationsanweisungen ermöglicht. Daran schließt sich die Beschreibung der Relation von Kopf- und Handregionen als gestenrelevante Bildbereiche und deren Klassifikation in die 6 zu erkennenden Kommando-Posturen an. Unter der Gesteninterpretation wird dann die Zuordnung der zu detektierenden Kommando-Posturen zu festen Navigationsanweisungen verstanden. Dies stellt gleichzeitig die Schnittstelle zum Teilthema 2, der visuell geführten lokalen Roboternavigation, dar.

2.1.1 Personenlokalisation

Die Lokalisation eines potentiellen Nutzers unter realen Umgebungsbedingungen muß auf mehreren, parallel detektierten pesonenspezifischen Merkmalen (visuelle Cues) beruhen, deren Fusion zu einem robusten personenspezifischen Auffälligkeitssystem führt. Eine Person, die mit dem Roboter in Interaktion treten will, wird aller Wahrscheinlichkeit nach näherungsweise frontal zu ihm ausgerichtet sein, so daß unter günstigen Aufnahmebedingungen alle Merkmale (Hautfarbe, Gesichtsstruktur, Kopf-Schulter-Kontur) gut ausgeprägt sind und robust detektiert werden können. Um neben den xy-Koordinaten im Bild auch die grobe Entfernung des Nutzers zur Kamera schätzen zu können, werden die strukturbasierten Merkmale in einer Auflösungspyramide berechnet, in der das Bild jeweils um den Faktor $1/\sqrt{2}$ unterabgetastet wird. Aus der am besten zutreffenden Ebene und der bekannten Größe eines Gesichts bzw. einer Kopf-Schulter-Silhouette kann dann eine Entfernungsabschätzung vorgenommen werden.

Im folgenden werden zunächst die zur Lokalisation verwendeten Cue-Module, die personenspezifische Merkmale detektieren, etwas detaillierter beschrieben. Daran schließt sich ein Abschnitt zur Fusion der personenspezifischen Merkmalskarten und zur Selektion der wahrscheinlichsten Personenregion an.

Hautfarbklassifikation Die Hautfarbe spielt bei vielen bekannten Verfahren zur Mensch-Maschine-Interaktion eine zentrale Rolle (z.B. (Littmann, E. and Ritter, H., 1995), (Schiele, B. and Waibel, A., 1995) oder (Yang, J. and Waibel, A., 1997), (Maggioni, C. and Kämmerer, B., 1996)). Einen umfassenden Überblick über derzeit existierende Systeme liefert KOHLER in (Kohler, M., 1998). Dabei müssen jedoch in allen Fällen starke Einschränkungen bzgl. des Szeneninhaltes und vor allem bzgl. der Umgebungsbedingungen definiert werden, die für den angestrebten Einsatzfall in keiner Weise zu legitimieren sind. Aus diesem Grund handelt es sich bei der Hautfarbe in dem hier vorliegenden Fall um kein eindeutiges Merkmal, obgleich das Vorhandensein dieses Merkmals die Hypothese, daß sich an dieser Stelle ein Gesicht bzw. Hände einer Person befinden, sehrwohl unterstützen kann. Dies rechtfertigt einerseits die Verwendung dieses Merkmals, andererseits ist eine Kombination mit weiteren personenspezifischen Merkmalen aus den oben genannten Gründen zwingend erforderlich.

Hautfarbene Regionen können in der gesamten Szene, d.h. auch im Hintergrund (vgl. Abb.5 rechts: schwarz entspricht der größten Hautwahrscheinlichkeit) auftreten, zum anderen variiert die Hautfarbe verschiedener Personen unter unterschiedlichen und vorab nicht bekannten Beleuchtungsbedingungen stark.

Für die Erstellung einer Hautfarbdatenbank wurden per Hand segmentierte Hautpartien von farbadaptierten Portraitaufnahmen unterschiedlicher Hauttypen verwendet.

Ziel der zunächst vorgenommenen *Farbadaptation* ist es, weitgehend reproduzierbare Farbwerte bei Beleuchtungsänderungen zu erhalten (siehe (Pomierski, T. and Gross, H.-M., 1996)). Dafür werden die RGB-Farbwerte eines Bildes in einen am Fachgebiet entwickelten, pysiologisch motivierten Elementarfarbraum transformiert, in welchem die Korrektur der Farbtripel stattfindet. Die Forschungsarbeiten zur Entwicklung dieses Farbraumes wurden 1997 mit dem Forschungspreis des TMWFK für angewandte Forschung gewürdigt.
Die RGB-Farbwerte werden durch eine Lineartransformation in den Elementarfarbraum transformiert, der durch eine Rot-Grün-, Blau-Gelb- und Weiß-Schwarz-Achse aufgespannt wird. Die im Farbraum dargestellten Pixel eines Farbbildes ergeben eine Punktwolke, deren größte Ausdehnung bei optimaler Beleuchtung entlang der Unbuntachse verläuft. Durch eine Farbadaptation können Bilder, deren Punktwolken aufgrund nichtoptimaler Beleuchtungsverhältnisse aus der Unbuntachse ausgelenkt sind, so adaptiert werden, daß weitgehend reproduzierbare Farbwerte vorliegen (siehe (Pomierski, T. and Gross, H.-M., 1996)). Obwohl das verwendete Adaptionsverfahren nur bedingt für die Indoor-Einsatzumgebung, in der meist kein farblich ausgewogener Bildinhalt vorliegt, geeignet ist, lassen sich dennoch stabilere und reproduzierbarere Farbwerte als ohne Adaption erzielen. Aktuelle Weiterentwicklungen des o.g. Adaptionsverfahrens, die auch bei farblich unausgewogenem Bildinhalt optimale Ergebnisse liefern, laufen derzeit und sollen dann integriert werden.

Die Lage der Hautfarbpunktwolke des von uns verwendeten Datensatzes nach erfolgter Farbadaptation ist in Abbildung 4 dargestellt.

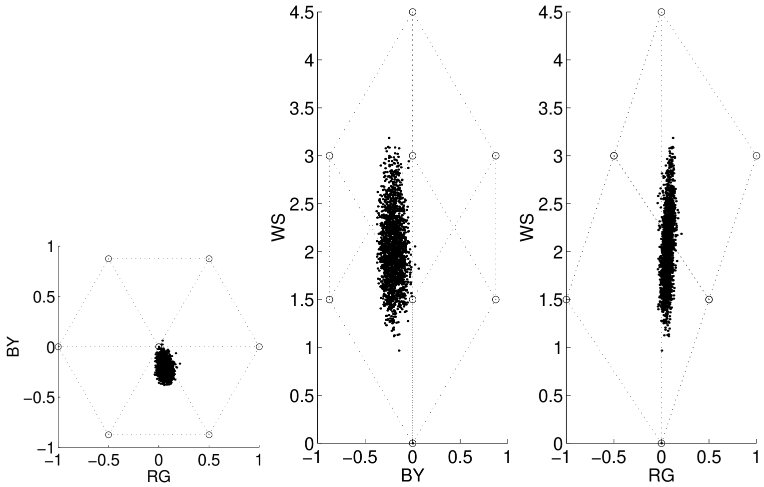

Abbildung 4: *Hautfarbwolke des Trainingsdatensatzes im neurophysiologisch motivierten Elementarfarbraum: Rot-Grün/Blau-Gelb-, Blau-Gelb/Weiß-Schwarz-, Rot-Grün/Weiß-Schwarz-Farbebene;*
linke Abb.: Farben der Eckpunkte im Uhrzeigersinn (von rechts): Magenta-Rot, Orange-Rot, Gelb, Grün, Cyan-Blau, Violett-Blau

Somit wird eine relativ stabile Farbempfindung realisiert, die allerdings immer noch szenenabhänging ist. Die bei Beleuchtungsänderung (Tageslicht, Neonlicht) auftretenden sprunghaften Farbänderungen können mit diesem Verfahren jedoch hinreichend genau kompensiert werden, so daß das Hautfarbmodell wieder zutrifft.

Als nächster Verarbeitungsschritt erfolgt nun die *Hautfarbklassifikation mit Hilfe eines Gaußmodells*. Aufgrund der großen Varianz in der Hautfarbe wird zur Nutzerlokalisation ein relativ grobes Hautfarbmodell verwendet, welches in erster Linie sicherstellen soll, daß Hautpartien möglichst sicher als hautfarben detektiert werden. Ein weiteres spezifischeres Hautfarbmodell, welches auf den aktuellen Nutzer adaptiert wird, wird zur Handdetektion eingesetzt. Um die möglichen Variationen in der Hautfarbe zuzulassen, werden zwei Maßnahmen ergriffen: zum einen wird bei jedem aufgenommenen Bild eine Farbadaptation durchgeführt, um die Beleuchtungseinflüsse zu kompensieren und zum anderen wird zur Klassifikation ein intensitätsnormiertes Hautfarbmodell verwendet.
Farbmodell: Als Farbmodell dienen die intensitätsnormierten Farbanteile r und g (chromatische Farben) aller Farbtripel dieser Hautfarbdatenbank (vgl. auch (Yang, J. and Waibel, A., 1997)), wodurch das Hautfarbmodell unspezifischer wird .

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B} \tag{1}$$

Für das verwendete Farbmodell wird das Farbhistogramm (s. Abb.5 mitte) durch eine 2-dimensionale Gaußfunktion approximiert, indem der Mittelwert m und die Kovarianzmatrix C des Datensatzes gebildet werden (s. Abb.5 links).

$$p(\mathbf{rg}) = \frac{1}{2\pi\sqrt{|\mathbf{C}|}}\, e^{-\frac{1}{2}(\mathbf{rg}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{rg}-\mathbf{m})}$$

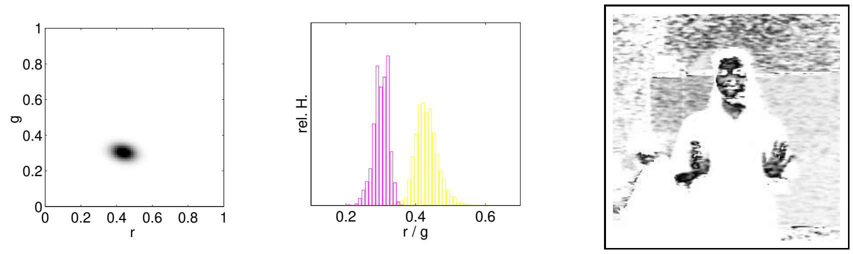

Abbildung 5: *v.l.n.r.: berechnetes Hautfarbmodell (Gaußfunktion); Histogramm der Farbanteile der Datenbank; Hautklassifikationsbeispiel*

Das in den ersten Zwischenberichten beschriebene Verfahren zur Hautfarbklassifikation, welches auf einem unüberwacht trainierten Growing-Neural-Gas-Netzwerk (unüberwachtes GNG) basierte, findet aus folgenden Gründen keine Verwendung mehr: In ausführlichen Tests unter verschiedenen Beleuchtungsbedingungen konnte zwar gezeigt werden, daß durch das Verfahren der Farbadaptation im Elementarfarbraum insgesamt bessere Klassifikationsergebnisse erzielt werden konnten, das Training des GNG mit den Farbpaaren der Rot-Grün/Gelb-Blau-Ebene des Elementarfarbraumes aber keinen bedeutenden Vorteil bringt. Nachteilig wirkt sich ebenfalls aus, daß eine personenspezifische Anpassung des Farbmodells über ein GNG nur umständlich zu realisieren ist, hingegen die Parameter einer Gaußfunktion schnell nachgeführt werden können. Weiterhin kann das verwendete neuronale Modell aufgrund der lokalen rezeptiven Felder das zugrundeliegenden gaußförmige statistische Modell nur mit großem Aufwand in der erforderlichen Güte approximieren.

Wurde eine Person (Gesicht) mit Hilfe der verschiedenen Cues (s. auch nachfolgende Abschnitte) gefunden, wird die aktuelle Gesichtsfarbe zur Bildung eines *spezifischeren Hautfarbmodells* herangezogen, so daß im damit klassifizierten Bild möglichst nur noch Gesicht und Hände des Nutzers detektiert werden. Dabei werden Mittelpunkt, Ausdehnung und Orientierung der Gaußglocke des allgemeinen Farbmodells (5 Parameter) an die aktuelle Hautfarbverteilung der Person angepaßt. Ziel der Anpassung des Hautfarbmodells ist zum einen, die Position der Hände möglichst sicher zu erkennen, zum anderen, das Tracken (Verfolgen) eines Nutzers zu erleichtern. Abbildung 6 stellt exemplarisch die Ergebnisse der Hautfarbklassifikation mit dem unspezifischeren, groben Hautfarbmodell und mit dem personenspezifisch adaptierten Hautfarbmodell gegenüber.

Abbildung 6: *Ergebnis der Hautfarbklassifikation m.H. des groben Modells (links, die dunkelsten Punkte besitzen die größte Hautfarbwahrscheinlichkeit) und Hautfarbklassifikation mit dem auf die vordere Person adaptierten spezifischen Hautfarbmodell (rechts)*

Gesichtsstrukturdetektion Auch hier gelten zunächst sinngemäß die gleichen Aussagen hinsichtlich Szeneninhalt und Umgebungsbedingungen, die bzgl. der Hautfarbe postuliert wurden, so daß die sichere Detektion von Gesichtern unter variablen Realwelt-Bedingungen noch immer eine Herausforderung darstellt.

Für die Detektion von Gesichtern wurden im Verlauf der Projektbearbeitung verschiedene Verfahren entwickelt und untersucht. Der zunächst favorisierte, eigenvektorbasierte Ansatz (siehe auch (Rowley, H. A. et al., 1995, Brakensiek, A. et al., 1998a)), bei dem die gesamte Grauwertmatrix jeder Auflösungsebene pixelweise mit den ersten n $(3 < n < 10)$ Eigengesichtern gefaltet und die resultierenden Fitwerte mit einem neuronalen Klassifikator klassifiziert wurden, lieferte zufriedenstellende Ergebnisse. Die Substitution des neuronalen Klassifikators durch ein abstandsbasiertes Verfahren, welches den originalen mit dem aus den Fitwerten rekonstruierten Bildausschnitt vergleicht, brachte keine entscheidenden Verbesserungen. Beide Verfahren besitzen jedoch den entscheidenden Nachteil, daß die unter Realweltbedingungen erreichbaren Ergebnisse die algorithmische Komplexität nicht rechtfertigen, da diese die notwendige echtzeitfähige Anwendung ohne zusätzliche Spezialhardware verhindert. Aus diesem Grund wurde ein weiteres Verfahren entwickelt, welches

- aus dem eigenvektorbasierten Ansatz abgeleitet werden kann und das
- einen guten Kompromiß zwischen Berechnungsaufwand und Leistungsfähigkeit darstellt.

Das Verfahren der *normierten mittelwertfreien Skalarproduktbildung* (NMS) verwendet analog zum eigenvektorbasierten Ansatz eine Gesichtsdatenbank und extrahiert daraus ein normiertes, mittelwertfreies „mittleres" Gesicht (siehe auch (Boehme, H.-J. et al., 1999)), was näherungsweise dem eigenwertgrößten Eigenvektor entspricht. Mit dem so ermittelten Prototypen wird auch hier die gesamte Grauwertmatrix jeder Auflösungsebene pixelweise gefaltet (jedoch nur noch eine anstelle von vorher n Faltungsoperationen), wobei jeder zu faltende Bildbereich vor der Faltung in eine normierte und mittelwertfreie Darstellung transformiert wird. Das Faltungsergebnis stellt jetzt unmittelbar ein Maß für das Vorhandensein eines Gesichts im betreffenden Bildausschnitt dar. Abbildung 7 stellt die Ergebnisse des eigenvektorbasierten Ansatzes und des NMS-Verfahrens anhand eines Referenzbildes gegenüber. Die enorm reduzierte Berechnungskomplexität rechtfertigt die beim NMS-Verfahren etwas größere Zahl an falsch positiven Detektionsergebnissen.

Detektion der Kopf-Schulter-Kontur Auch bei der Detektion der Kopf-Schulter-Kontur wird davon ausgegangen, daß eine Person, die mit der Plattform in Interaktion treten will, sich dieser frontal zuwendet.

Der im Rahmen des Projekts entwickelte Ansatz, der dem von OREN ((Oren, M. et al., 1997)) ähnelt, beruht auf drei wesentlichen Säulen:

- Zur Bestimmung eines mittleren Konturmodells (Konturverlauf, Abbildung 8) werden spezielle orientierungssensitive Filter, sogenannte steuerbare Filter, verwendet, die sich im Vergleich zu den häufig

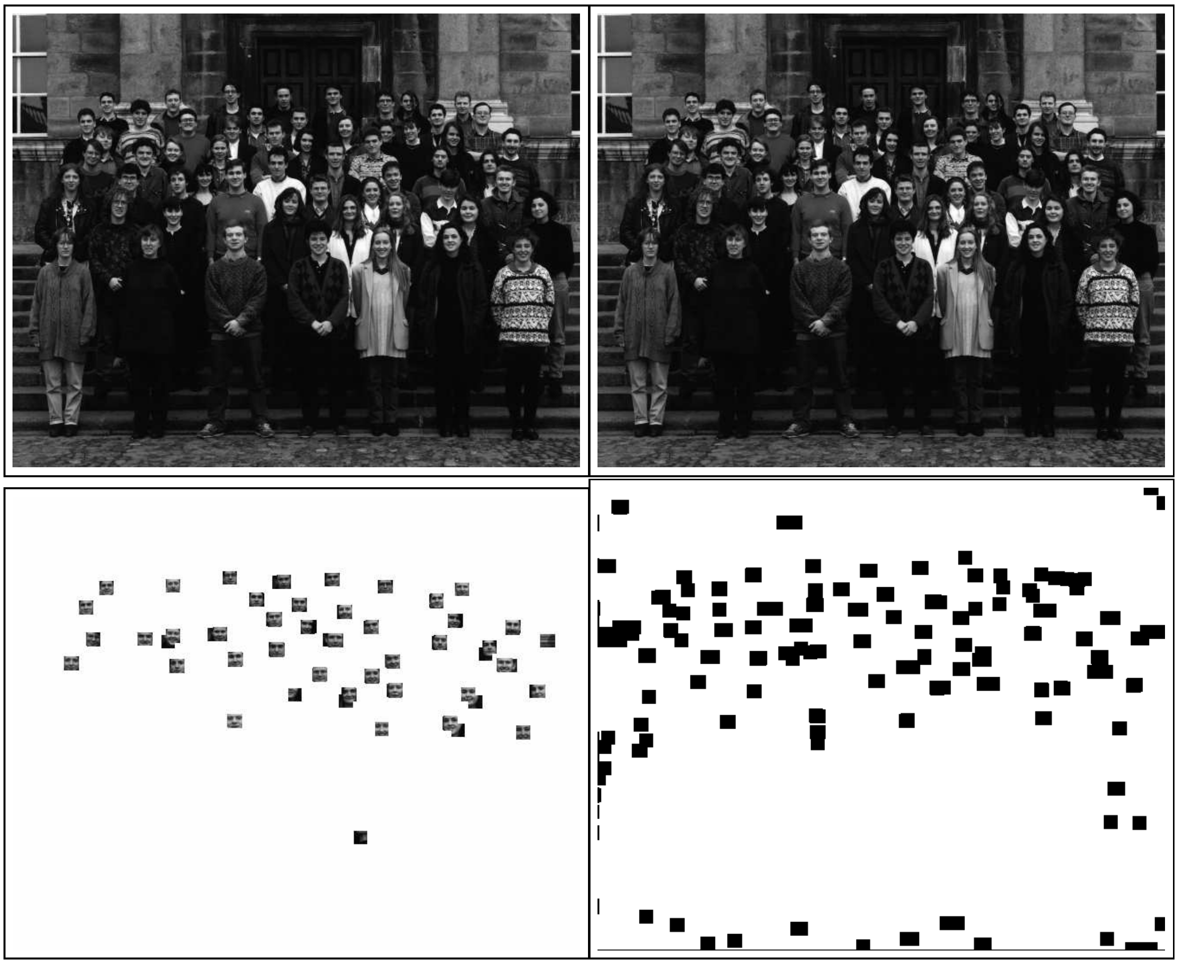

Abbildung 7: *Originalbild (oben) und Ergebnisse der Gesichtsstrukturdetektion anhand des eigenvektorbasierten Verfahrens mit Klassifikation der Fitwerte durch einen neuronalen Klassifikator (links) und anhand des NMS-Verfahrens (rechts).*

genutzten Gabor-Filtern durch eine deutlich reduzierte Berechnungskomplexität bei gleichzeitig erreichbarer höherer Orientierungsgenauigkeit auszeichnen.

- Die Detektion einer Kopf-Schulter-Kontur erfolgt auf allen Ebenen einer Auflösungspyramide, was die Berechnungseffizienz des Gesamtverfahrens erhöht und die Abschätzung der Entfernung der detektierten Kontur von der Kameraebene gestatttet.
- Das Maß für das Vorhandensein der gesuchten Kontur wird mittels einer speziellen Abstandsmetrik ermittelt, die den Abstand zwischen dem Prototyp-Muster (mittlerer Konturverlauf) und dem im Bildausschnitt vorgefundenen Konturverlauf betrachtet.

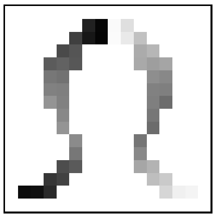

Abbildung 8: *Darstellung des berechneten Konturmodells als mittlerer Konturverlauf, der aus einem Satz von Beispielaufnahmen ermittelt wurde; die Orientierungswinkel sind durch Grauwerte kodiert (*$0°$*: schwarz;* $90°$*: mittelgrau;* $180°$*: weiß).*

Die Abbildung 9 verdeutlicht das Gesamtverfahren zur Kopf-Schulter-Kontur-Detektion.

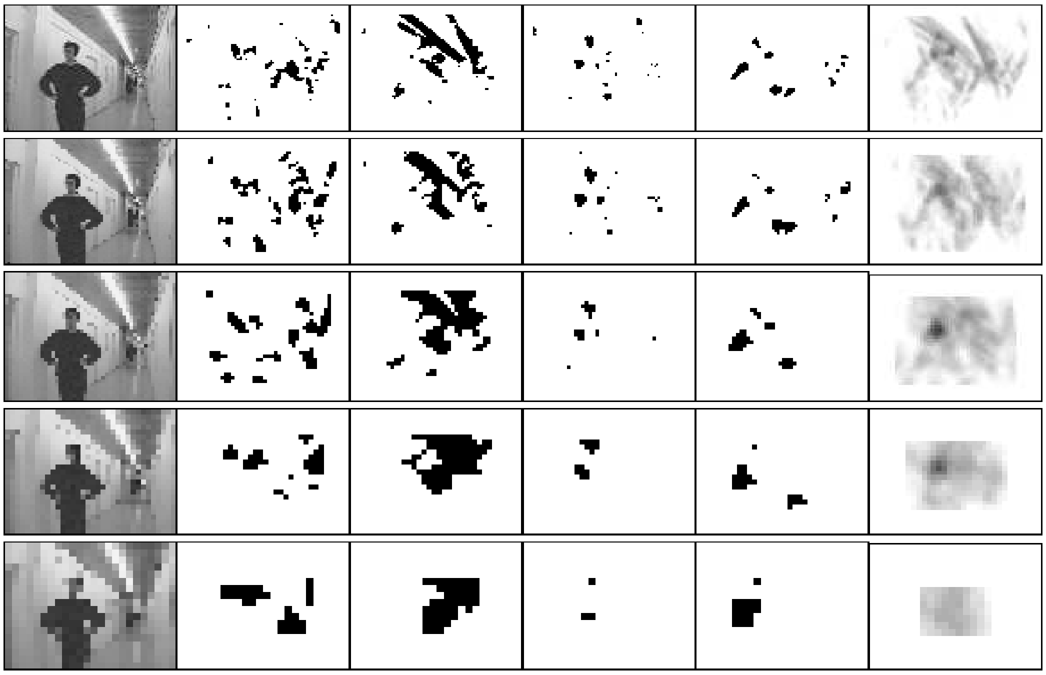

Abbildung 9: *Exemplarische Darstellung des Verfahrens zur Kopf-Schulter-Kontur-Detektion; jede Zeile korrespondiert mit einer Ebene der Auflösungspyramide; v.l.n.r.: Originalbild, Ergebnisse der Orientierungsfilterung für die Orientierungswinkel* 0°, 45°, 90° *und* 135°, *Resultat des Abstandsmaßes zum Prototyp-Muster (dunkle Bereiche markieren hohe Übereinstimmung).*

Selektionsprozeß zur Auswahl des sichersten Personenkandidaten Abbildung 10 gibt einen Überblick über das gesamte personenspezifische Auffälligkeitssystem.

Die Cue-Module bestehen aus Pyramiden, die die detektierten Merkmale auf den verschiedenen Auflösungsebenen repräsentieren und ebenenweise die Eingänge für die *primären Auffälligkeitskarten* bilden. Diese Karten enthalten dynamische Neuronen, die sowohl innerhalb einer Ebene als auch zwischen benachbarten Ebenen in Wechselwirkung treten. Der dabei ablaufende Selektionsprozeß (siehe (Amari, S., 1977)) liefert auf *der* Auffälligkeitsebene *einen* Aktivitätspeak, der die wahrscheinlichste Position einer Person im Raum indiziert. Die Abbildungen 11 bis 13 zeigen jeweils in der rechten Spalte das Resultat der Personenlokalisation sowie die dazugehörigen Ergebnisse der verschiedenen Cue-Module und fassen den aktuellen Entwicklungsstand des personenspezifischen Auffälligkeitssystems zusammen.

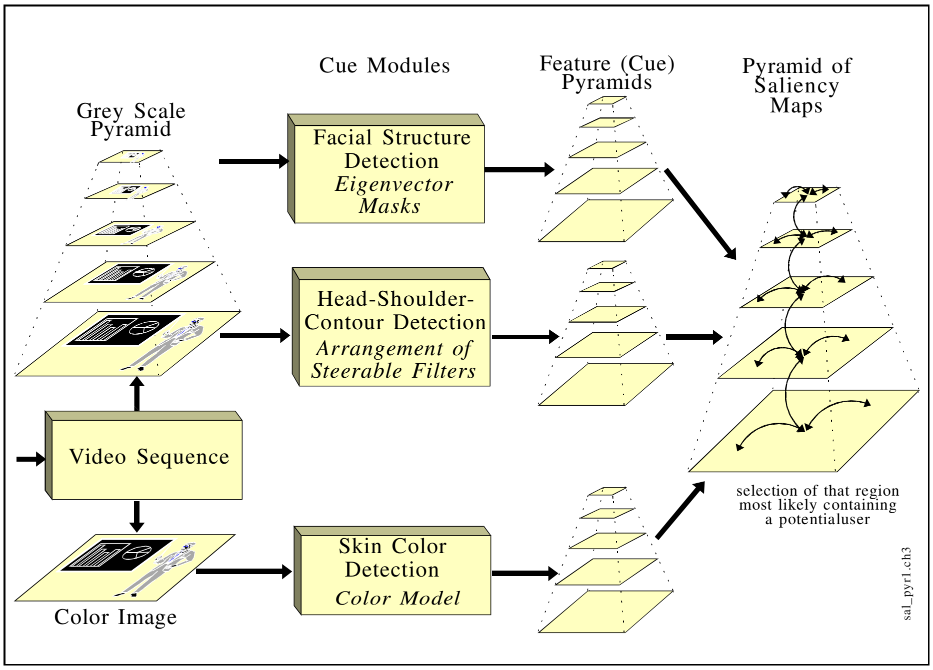

Abbildung 10: *Gesamtarchitektur des personenspezifischen Auffälligkeitssystems.*

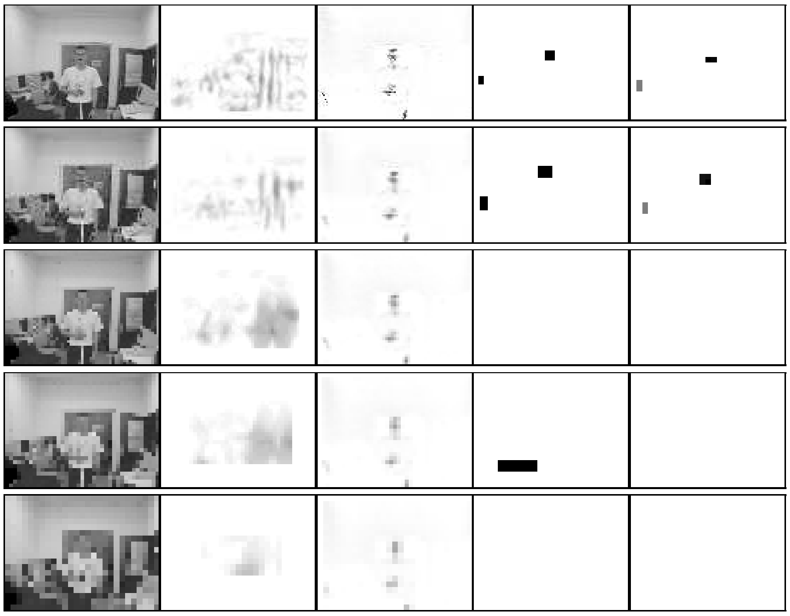

Abbildung 11: *Lokalisationsergebnis im Innenbereich auf 2.0m: In jeder Zeile sind alle Zwischenergebnisse und Ergebnisse für eine der fünf Auflösungsstufen (Abstand* $1/\sqrt{2}$*) dargestellt. Die fünf Spalten enthalten: Input, das Ergebnis der Anwendung des Filterarrangements (Ähnlichkeitsmaß), das Ergebnis der Farbklassifikation, die Gesichtsdetektion und schließlich das Selektionsergebnis innerhalb des dreidimensionalen Feldes dynamischer Neuronen. Die Lokalisation der Person erfolgt sowohl auf der nullten und der ersten Pyramidenebene. Aufgrund des eingegangenen Kompromisses bzgl. des Abbruchkriteriums zur Begrenzung der Iterationsschritte des dynamischen Selektionsprozesses bleiben neben dem Hauptblob (starkste Ausprägung auf der ersten Ebene) z. T. separate Nebenblobs stehen, die aber nur wenig Einfluß auf das Selektionsergebnis nehmen.*

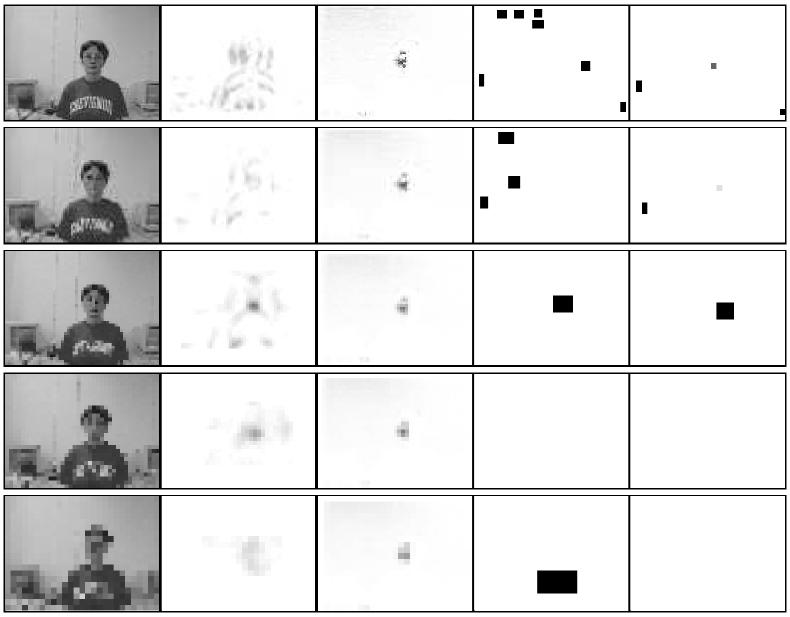

Abbildung 12: *Lokalisationsergebnis im Innenbereich auf ca. 1.4m: Hier wurde gegenüber Bild 11 die Entfernung der Person verringert, so daß die Lokalisation im Unterschied dazu auf der mittleren Pyramidenebene erfolgt.*

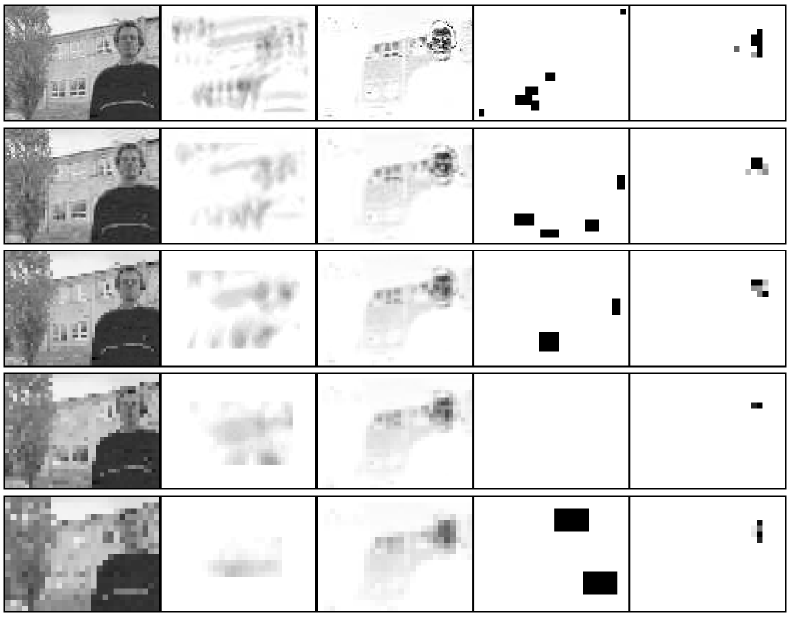

Abbildung 13: *Lokalisationsergebnis im Außenbereich auf ca. 1.4m: Im Unterschied zu den Bildern 11 und 12 liefert hier die gesichtsspezifische Filterung nur fehlerhafte Resultate. Dadurch sinkt die Schärfe und die Genauigkeit der Entfernungsbestimmung der Person, wobei das Blobzentrum wieder auf der ersten Ebene liegt, die nullte Ebene die meisten aktiven Elemente aufweist, der Blob sich jedoch auch auf die höheren Ebenen erstreckt.*

2.1.2 Erkennung statischer Kommandoposturen

Nachdem die aktuell interessanteste Personenregion lokalisiert wurde, erfolgt nun die Beobachtung dieser Person mit Ziel, die bereits dargestellten Kommando-Posturen zu detektieren, um diese dann in ein adäquates Fahrverhalten umsetzen zu können. In diesem Abschnitt werden die Verarbeitungsschritte beschrieben, die zur Erkennung der Kommando-Posturen notwendig sind. Der erste Schritt umfaßt die Kamerasteuerung, die die betrachtete Person erfaßt, diese in ungefähr gleichbleibender Größe im Bild hält und die sogenannten *Gestenszenen* liefert (siehe auch Abbildung 3). Die Kamerasteuerung bildet dabei ein eher technisches Problem und wird hier nicht ausführlich behandelt.

Farbbasierte Segmentierung der Gestenszenen Die Segmentierung von Gesichtern und Händen als die gestenrelevanten Teile des Bildes basiert ausschließlich auf dem Merkmal Hautfarbe, da sich die Position der Hände mit strukturbasierten Verfahren aufgrund der großen Erscheinungsvariabilität nicht in Echtzeit bestimmen läßt.

Um ein auf die aktuell erfaßte Person optimal angepaßtes Hautfarbmodell zu erhalten, werden aus der Gesichtsregion der Person Farbpixel extrahiert, aus denen ein wie im Abschnitt 2.1.1 beschriebenes Gauß-Modell gebildet wird. Die letztendliche Segmentierung erfolgt über den Abstand zu diesem Modell und eine anschließende Schwellwertoperation, was zu einem binären Segmentierungsergebnis führt (siehe Abbildung 14).

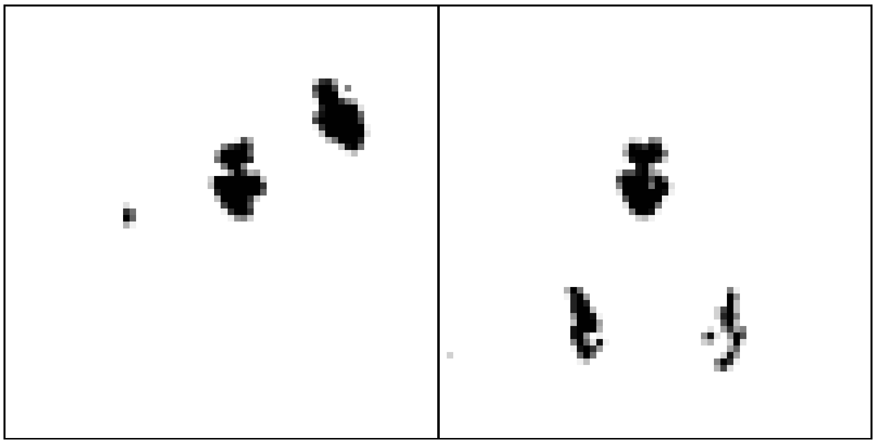

Abbildung 14: *Hautfarbbasierte Segmentierung der Gestenszene für zwei Kommandoposturen (links „Hallo“, rechts „Halt“)*

Beschreibung der segmentierten Gestenszenen durch statistische Momente Die Beschreibung natürlicher Gesten mittels statistischer Momente wurde bereits in (Rigoll, G. and Kosmala, A., 1997) vorgeschlagen. Von der hautfarbsegmentierten und anschließend auf eine Dimension von 64×64 Pixel unterabgetasteten Gestenszene wird ein 9-dimensionaler Merkmalsvektor berechnet, der die 9 zentralen Momente des Bildes umfaßt. In der Abbildung 15 sind die verwendeten Momente graphisch illustriert.

Erkennung der Kommando-Posturen mit neuronalen Klassifikatoren Die letztendliche Zuordnung der in der Gestenszene vorliegenden Anordnung von Gesicht und Händen in die Posturenklassen erfolgt mit Hilfe von neuronalen Klassifikatoren, denen Standard-Netzwerkparadigmen zugrundeliegen. In diesem Abschnitt werden zwei im Rahmen des Projektes untersuchte neuronale Klassifikatoren hinsichtlich ihrer wesentlichen Eigenschaften gegenübergestellt.

Für das Training der Klassifikatoren wurden jeweils 360 Merkmalsvektoren verwendet, die 4 Klassen zuzuordnen waren, wobei jede Klasse einer Kommando-Postur (Hallo, Stop, Links, Rechts) entspricht.

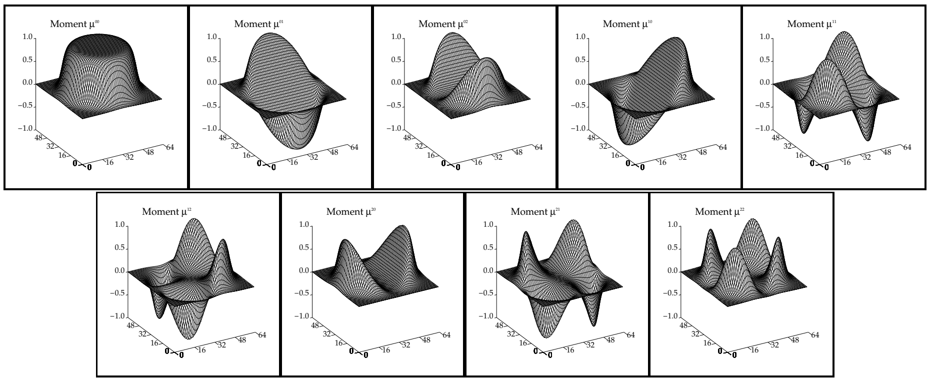

Abbildung 15: *Graphische Illustration der Momente, die zur Beschreibung der Gestenszene angewendet werden; die Momente wurden von einem Bild mit homogenem Grauwert berechnet*

Zunächst wurde ein Radial-Basis-Function-Netzwerk (RBF-Netzwerk), bestehend aus 9 Eingabe-, 20 Hidden- (RBF) und 4 Ausgabeknoten, verwendet. Das Training der RBF-Schicht basierte auf dem Neural-Gas-Algorithmus, die Gewichte zur Ausgabeschicht wurden mit der vom Perceptron bekannten Delta-Lernregel adaptiert.

Beim zweiten Netzwerk handelte es sich um ein modifiziertes Counterpropagation-Netzwerk. Dieses Netzwerk wies eine mit dem RBF-Netzwerk identische Topologie auf. Auch hier wurden die Gewichte von der Eingabe- zur Hiddenschicht mit dem Neural-Gas-Algorithmus trainiert. Die Lernregel zur Adaption der Gewichte von der Hidden- zur Ausgabeschicht basierte auf dem Outstar-Modell nach GROSSBERG ((Grossberg, S., 1969)).

Die folgende Tabelle faßt die mit den beiden neuronalen Klassifikatoren erreichten Ergebnisse zusammen. Das RBF-Netzwerk erzielte ein robuste Performance, und die Anzahl der faslch klassifizierten Muster war relativ klein, wohingegen das Counterpropagation-Netzwerk eine hohe Zahl falsch klassifizierter Muster aufwies und zusätzlich viele Muster gar nicht klassifizieren konnte. Der Vorteil des Counterpropagation-Netzwerks besteht in der im Vergleich zum RBF-Netzwerk sehr viel kürzeren Trainingszeit, was jedoch die schlechten Ergebnisse hinsichtlich der Performance nicht aufwiegt.

Network	Topology	# of Trainingspatterns	# of Testpatterns	# of false classified patterns	# of unclassified patterns	Recognition Rate
RBF	9-20-4	180	180	6	10	91.2 %
CP	9-20-4	180	180	40	48	75.6 %

Ausgehend von den beschriebenen Ergebnissen erfolgt gegenwärtig die Umsetzung des RBF-basierten Netzwerkmodells auf der Roboterplattform MILVA.

2.2 Erzielte wissenschaftliche Ergebnisse zum Teilthema 2: Visuell geführte lokale Roboternavigation mit lernfähigen Neuronalen Architekturen

Das Ziel des GESTIK-Projektes bestand neben der in Abschnitt 2.1 vorgestellten neuronalen Architektur zur Nutzer-Roboter-Kommunikation in der Entwicklung einer neuronal basierten Steuerarchitektur, mit der sich ein mobiler Roboter unter „Blickkontakt“ zum Anwender entsprechend dessen gestenbasierten Anweisungen teilautonom bewegt. Sich auf der Fahrtroute befindende statische und dynamische Hindernisse sollen dabei geeignet umfahren werden. Dem menschlichen Interaktionspartner kommen in diesem Szenario folgende Funktionen zu:

(a) Der Anwender ist der primäre Handlungsantrieb für das System, welches ständig versuchen soll, sich diesem auf eine definierte Entfernung anzunähern, solange er dies nicht durch eine entsprechende (gestikulierte) Instruktion unterbindet.

(b) Situationen, in denen sich dem System mehrere Handlungsalternativen eröffnen, sind vom Anwender durch eine eindeutige Anweisung aufzulösen. Voraussetzung dafür ist, daß entsprechende Konfliktsituationen erkannt werden und sich der Handlungsvorschlag des Nutzers geeignet in den systeminternen Entscheidungsprozeß einbinden läßt.

(c) Wegen der rein visuell basierten Navigation ist nicht auszuschließen, daß das Fahrzeug mit nicht eindeutig interpretierbaren sensorischen Situationen operieren muß. Hier soll der Anwender assistierend eingreifen, um falsche Hypothesen zu korrigieren.

Im Rahmen des GESTIK-Projektes wurde ein neuronales Agentensystem konzipiert, mit dem auf Basis eines lokalen Handlungsrepertoirs in enger Interaktion mit einem Anwender komplexe Verhaltensleistungen erreicht werden können. Grundlage dafür bildet eine geeignete Problemdekomposition in einer heterarchischen Struktur, mit der es gelingt, den gesamten situationsspezifischen Aktionsraum selektierbar zu repräsentieren. Die Realisierung der lokale Handlungsziele verfolgenden Agenten zur Navigation eines mobilen Roboters wird ausführlich vorgestellt. Grundlage dieser intentionalen Agenten bilden überwacht vortrainierte sensomotorische Mappings. Die resultierenden verschiedenen Handlungsvorschläge werden basierend auf ihrem energetischen und zeitlichen Kontext flexibel fusioniert.

Mit dem vorgestellten Agentensystems kann ausschließlich lokales Navigationsverhalten erreicht weden, da sich die reale Umgebung im visuellen Eindruck der Navigationskamera des Systems niemals eindeutig bezüglich der globalen Position im Einsatzfeld darstellt bzw. darstellen soll (keine eindeutigen Landmarken). Dies stellt im Kontext des Projektziels keine Einschränkung dar, da sich im Zusammenwirken von Anwender und Fahrzeug die erwünschte Systemleistung vollständig und widerspruchsfrei umsetzen läßt. Auf diese Weise war es vielmehr möglich, die lokalen Verhaltensleistungen konsequent zu entwickeln und gleichzeitig demonstrabel umzusetzen, weil eine globale Bewegungsplanungsebene zunächst völlig entkoppelt ist, sich aber konzeptionell problemlos ergänzen läßt.

Zielsystem war die Roboterplattform MILVA. Eine zentral montierte Weitwinkel-Kamera liefert den ausschließlich für die Navigation genutzten visuellen Datenstrom. Als Experimentalplattform der hier vorgestellten Untersuchungen diente zunächst der Miniaturroboter KHEPERA (kreisrund, $\o = 5,5cm$; mittig montierte Farbkamera), dessen praktischer Einsatz sich ausgesprochen unproblematisch gestaltet. Aufgrund einer definierten Labyrinth-Umgebung ergaben sich zunächst keine Probleme bzgl. der Untergrund-Hindernis-Trennung in den zur Navigation verwendeten Kamera-Bildern, sodaß keine aufwendige Vorverarbeitung der Rohdaten notwendig war, um stabile und reproduzierbare Eingabeinformationen zu gewährleisten. Bei der anschließenden Übertragung der Architektur auf die MILVA-Plattform war dies jedoch in ungleich stärkerem Maße der Fall, was durch die stark variierende optische Erscheinung von Indoor-Umgebungen und die dort auftretenden Störungen (schwankende Beleuchtungsbedingungen, Reflexionen, Schatten etc.) bedingt ist.

In den folgenden Abschnitten wird zunächst der Entwurf der neuronalen Steuerungsarchitektur und deren Test mit dem KHEPERA-System in einer Labyrinth-Umgebung beschrieben. Der Umsetzung dieser Architektur auf die MILVA-Plattform mit den sich dabei ergebenden Realwelt-Bedingungen und Realwelt-Problemen widmet sich der abschließende Abschnitt zum Teilthema 2.

2.2.1 Neuronale Architektur zur Handlungsorganisation intentionaler Agenten

Problemdekomposition und Repräsentation von Handlungsantrieben Wie bereits angedeutet, sind die folgenden beiden signifikanten Eigenschaften notwendig, um zu einer interaktiven Verhaltensleistung zu gelangen:

- Die sich in einer bestimmten Situation bietenden Handlungsalternativen müssen voneinander separierbar sein. Hierbei ist einerseits die Erkennung von sich gleichzeitig bietenden Handlungsoptionen für das System notwendig, um u.U. eine Anfrage an den Anwender stellen zu können und so Konflikte aufzulösen. Andererseits wird damit erst die Ausschöpfung der gesamten sich bietenden Aktionsvielfalt gewährleistet.
- Jede der einzelnen Handlungsalternativen muß sich in einer Weise unterstützen lassen, in der sowohl die Entscheidung des Anwenders bzw. einer hierarchisch höher liegenden Verarbeitungsebene einbindbar ist, als auch weiterhin das lokale Wissen der Struktur erhalten bleibt und genutzt werden kann.

Zur Umsetzung dieser Spezifika bietet sich eine Dekomposition in prinzipiell identisch arbeitende, aber verschiedene Handlungsziele verfolgende Agenten an, deren Handlungsvorschläge so die gesamte situationsgemäße Aktionsvielfalt repräsentieren.

Implizite Repräsentation von Handlungsantrieben in intentionalen Agenten Es bietet sich an, menschliche Problemlösungsstrategien und deren Teilaspekte auf neuronale Agenten abzubilden, die jeweils versuchen, diese Teilaspekte optimal zu bearbeiten. Daher erhält jeder Agent im Rahmen des Gesamtproblems eine Zweckbestimmung, die über einen agentenspezifischen Handlungsantrieb, eine Intention, beschrieben wird. Während in einigen Lernverfahren die Handlungsantriebe über entsprechende Bewertungsfunktionen ausgedrückt werden (z.B. Reinforcement-Lernen), besteht bei überwachten Lernverfahren die Möglichkeit, durch „Vormachen“ einer charakteristischen Problemlösungsstrategie eine implizite Beschreibung der Systemziele zu erzeugen. Dies wird in der Literatur als Experten-Cloning bezeichnet.

Für die Problematik eines lokalen Navigationsverhaltens mit den Teilaspekten gerichtete schnelle Lokomotion, Vermeidung statischer und dynamischer Hindernisse, eingeschränkte Exploration usw. bietet sich an, separate Agenten zu realisieren, die in ihrem Wechselspiel

- Flurbereiche möglichst mittig durchfahren,
- sich bietende Abbiegemöglichkeiten nutzen (Explorationsverhalten, z.B. Nutzersuche),
- Kreuzungsbereiche möglichst geradlinig überfahren (im Sinne eines Beharrungsvermögens) oder
- in Gefahrenbereichen anhalten bzw. umkehren.

In der Realisierung solcher Agentenausrichtungen durch entsprechende Cloning-Strategien zeigt sich ein grundsätzlicher Unterschied zu den Reinforcement-Methoden, bei denen durch entsprechende Bewertungsfunktionen eine Spezialisierung der Agenten auf oben genannte Verhaltensleistungen erreicht wird. Dies ist beim überwachten Agententraining in dieser Weise nicht möglich, weil den Agenten zu **allen** sensorischen Situationen eine situations- und intentionsadäquate Systemantwort präsentiert werden muß. Deshalb manifestiert sich die Realisierung der oben aufgeführten Teilaspekte nicht in einer Trainingsstrukturierung, die das Gesamtproblem in solche Teilaspekte (Situationen) zerlegt, sondern vielmehr in der Definition von Handlungskonzepten, nach denen der Experte beim Vormachen unter mehreren situationsbezogen optionalen Aktionen in konsistenter Weise eine bestimmte auswählt. Das Spektrum dieser festzulegenden Handlungsregeln muß nun für alle auftretenden Teilprobleme adäquates Verhalten ermöglichen. Für die Roboternavigation besteht eine, wenn auch triviale, Möglichkeit der Definition solcher Handlungskonzepte in der Trennung: „Fahre so weit wie möglich...

(a) und nutze die nächste sich bietende Gelegenheit, nach rechts abzubiegen!“

(b) und nutze die nächste sich bietende Gelegenheit, nach links abzubiegen!“

(c) geradeaus und biege nur ab, wenn keine Alternative besteht!“

Diese Einzeldefinitionen sind zwar in jeder Situation anwendbar, ermöglichen jede für sich aber noch keine Bewältigung der Gesamtproblematik. Im Zusammenwirken dieser Handlungsziele wird dagegen ein komplexes Gesamtverhalten erreicht, mit dem nicht nur die verschiedenen Teilaspekte der Roboternavigation abgedeckt werden, sondern das auch die oben genannten Voraussetzungen für eine interaktive Systemleistung erfüllt. So wird beispielsweise bei einem auftretenden Hindernis der Agent „3“ versuchen, bis kurz vor einer Kollision darauf zuzufahren. Die beiden anderen Handlungsziele repräsentieren dagegen ständig alle Möglichkeiten, von diesem Pfad abzuweichen. Spätestens wenn auch der „Geradeaus-Agent“ zu einer Ausweichbewegung gezwungen ist, wird über den „Gewinner“ der beiden „Abbiege-Agenten“ die günstigere Alternative herausgebildet. Dabei existieren ständig Eingriffsmöglichkeiten durch eine höhere Verarbeitungsebene, indem einzelne der optionalen Aktionen unterstützt werden.

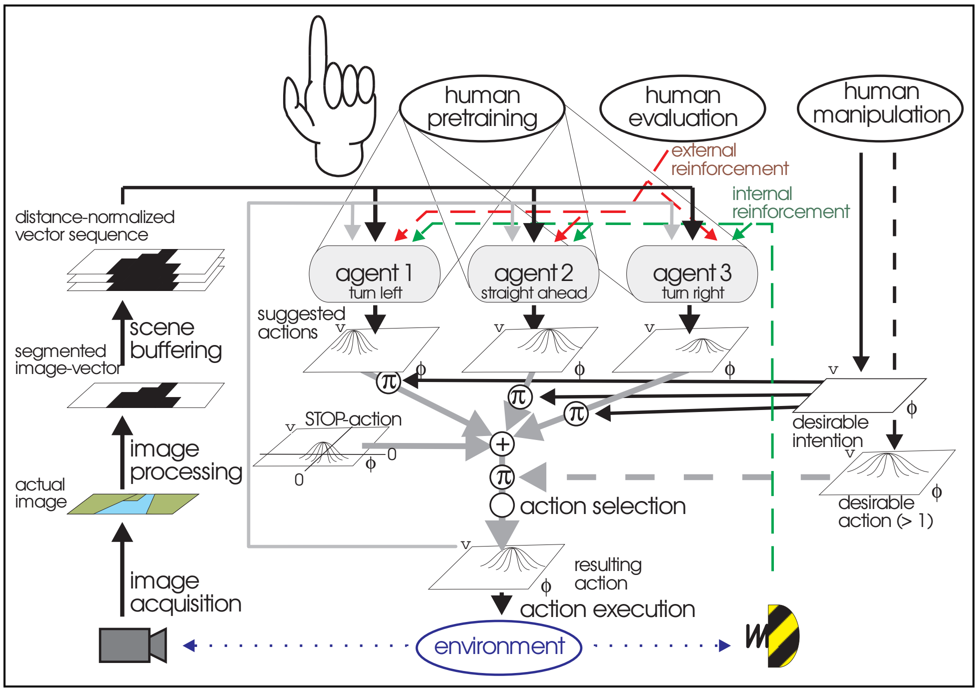

Abbildung 16: *Gesamtübersicht der verteilten Agentenstruktur. (siehe Text für detaillierte Beschreibung.)*

Handlungsorganisation durch Agentenfusion Wie bereits angedeutet werden die einzelnen Agenten nun in vielen Situationen entsprechend ihrer Intention (trainierten Handlungsziele) kontroverse Handlungsvorschläge generieren, die situationsspezifisch in einer Weise zu fusionieren sind, die die Aspekte des globalen Fahrverhaltens (z.B. Fahrt zum Nutzer oder entsprechend Nutzeranweisung) berücksichtigt. Ein Strukturentwurf zur Zusammenfassung der Handlungsvorschläge solcher Einzelagenten ist in Bild 16 dargestellt: Expertenwissen fließt dabei auf 3 Kanälen in das Gesamtsystem ein:

(a) Die verschiedenen Agenten bilden den visuellen Datenstrom (distance-normalized vector sequence) auf situationsadäquate Aktionen (suggested actions) ab, die in zweidimensionalen Karten durch Geschwindigkeit v und Lenkwinkel ϕ topologisch kodiert sind. Sie erhalten, wie bereits erläutert, ihre intentionale Ausrichtung durch Belehrung mit unterschiedlichen, von einem Experten erstellten Trainingsmengen (human pretraining, siehe Abschnitt *Realisierung der intentionalen Agenten*).

(b) Die Aktionsvorschläge der über identischen Datenströmen operierenden Agenten werden überlagert (+) und können sowohl intentionsorientiert, also agentenspezifisch (desirable intention), als auch aktionsorientiert (desirable action) durch Anweisungen des Nutzers verstärkt werden (human manipulation). Weil beispielsweise bei Abbiegemanövern u.U. zunächst in die entgegengesetzte Richtung ausgeholt wird, muß dabei der handlungszielorientierte (agentenspezifische) Nutzereingriff die primäre Anweisungseinbindung darstellen. Dieser Eingriff findet zwar bereits auf einer symbolischen Ebene statt, basiert aber dabei auf den vom Experten während des Agententrainings verwendeten Begriffen. Durch den rein modulierenden Eingriff kann sichergestellt werden, daß das Fahrzeug zwar diese externen Anweisungen befolgt, aber dabei immer nur mit dem eigenen situationsgemäßen Verhaltensrepertoire agiert. Die Einkopplung der vom Anwender ausgehenden Anweisungen wird konzeptionell so realisiert, daß über ein Gesten(Posen)-Alphabet, mit dem eine visuelle Nutzer-Roboter-Kommunikation geführt werden soll, eine Interpretation zu einer entsprechenden Agenten- bzw. Aktionsunterstützung möglich wird (siehe hierzu (Boehme, H.-J. et al., 1998a)).

(c) Zur weiteren Adaption der Agenten muß die auf Basis des energetischen und zeitlichen Kontextes letztendlich ausgewählte Aktion (action selection) auf die Agenten zurückgeführt werden, um diese dem erzielten Handlungserfolg zuzuordnen. Basis für die Bewertung ist die sensorische Erfahrung von Kollisionen, die als Reinforcement-Signal (internal reinforcement) dazu führen soll, daß das Gesamtsystem in einer späteren vergleichbaren Situation eine andere (bessere) Aktion ausführt. Das Modul human evaluation verkörpert die Handlungsbewertung durch einen Experten über ein external reinforcement. Die Identifikation dieser Kollisionen o.ä. vorwegnehmenden Bewertungssignale aus einem multimodalen Datenstrom (visuell, akustisch) mittels Konditionierung ist Gegenstand weiterer Forschungsarbeiten am Fachgebiet.

Als Ergebnis der agenteninternen Adaptionsvorgänge kann es, wie auch in unbekannten Situationen, dazu kommen, daß ein oder mehrere Agenten keinen Aktionsvorschlag liefern. Für diese Fälle wird vergleichbar einer permanenten Besorgnis ständig eine unterschwellige Stop-Absicht (STOP-action) den Aktionsvorschlägen überlagert, um sich dann mit der Anhalteaktion (v=0, ϕ=0) durchzusetzen.

Realisierung der intentionalen Agenten Als Ausgangspunkt für einen überwacht trainierten Einzelagenten einer derartigen verteilten Steuerarchitektur bietet sich der an der CMU Pittsburgh entwickelte ALVINN-Ansatz (Pomerleau, 1993) an, dessen szenariospezifische Umsetzung und Erweiterung in diesem Abschnitt betrachtet wird. Die Grundidee von ALVINN besteht in einer direkten Abbildung eines Kamerabildes auf einen Lenkwinkel, um damit ein Straßenfahrzeug auf gewöhnlichen Straßen zu steuern. Dabei dient ein Multi-Layer-Perceptron als Basisstruktur, die das von einem Trainer vorgegebene Verhaltensmuster approximiert und dabei den analogen Wert des einzustellenden Lenkwinkels in Form einer topologischen Ausgabekodierung repräsentiert (siehe Abb. 18).

Für die Steuerarchitektur des KHEPERA dient ebenfalls ein zweilagiges Multi-Layer-Perceptron in reiner Feed-Forward-Architektur als Netzwerk, zu dessen Belehrung klassisches Backpropagation eingesetzt wird. Die Vorverarbeitung der Input- und Outputdaten ist in den Abbildungen 17 und 18 dargestellt. Die Trainingsmenge bildet eine 2500 Beispielsituationen umfassende Datenbank, die von einem „Experten“ erstellt wurde und in etwa 50 Trainingsepochen präsentiert wird. An dieser Stelle muß betont werden, daß die Navigation durch den Experten bei der Erstellung der Trainingsdaten ausschließlich mittels der Videobilder der on-Board-Kamera des KHEPERA erfolgte, um konsistente Datenquellen in Trainings- und Recallphase zu garantieren.

Probleme / Erweiterungen der ALVINN-Architektur In dieser ersten Implementierungsform meistert der KHEPERA alle gutartigen Situationen auch in einem unbekannten Labyrinth weitgehend kollisionsfrei und zeigt dabei ein erwartungsgemäßes (dem Training entsprechendes) Verhalten (Abb. 19). Wichtig ist dabei, daß immer Elemente des Untergrunds im Blickbereich der Kamera bleiben. Hierbei macht sich das schmale Blickfeld der Kamera stark einschränkend bemerkbar (Abb. 20).

So kann es ab bestimmten Lenkeinschlägen dazu kommen, daß das Fahrzeug mit Objekten kollidiert, die es in der aktuellen Situation visuell gar nicht wahrnimmt. Dies führt bereits bei der Erstellung der Trainingsmuster durch den Experten zu einer angepaßten Verhaltensweise. Typisch dafür ist, daß soweit wie möglich eine Geradeausfahrt durchgeführt wird, und nur dort ein Abbiegen stattfindet, wo der vor dem Fahrzeug liegende Fahrweg

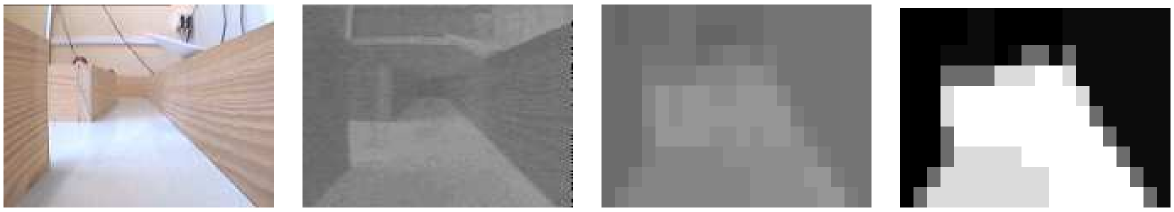

Abbildung 17: *Die Phasen der Bildvorverarbeitung auf dem KHEPERA: das Originalbild (links), die Blau-Gelb-Aktivierungen dieses Bildes nach (Pomierski, T. and Gross, H.-M., 1996) (2. v. links), nach der 3-fachen Unterabtastung des relevanten Bildausschnitts (2. v. rechts) und der Dynamikanpassung (rechts).*

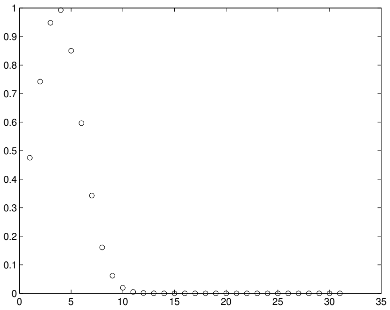

Abbildung 18: *Eindimensionale topologische Kodierung des Lenkwinkels über einem m-dimensionalen Outputvektor: Da es beim KHEPERA mit seinen beiden getriebenen Rädern keine explizite Steuergröße „Lenkwinkel" gibt, muß diese aus den Geschwindigkeiten der beiden Räder ermittelt werden (LS=LeftSpeed,RS=RightSpeed):*

$$y_i = e^{-\frac{\left(i-\frac{m+1}{2}\cdot(1+\frac{LS-RS}{LS+RS})\right)^2}{M}} \qquad i \in [1\ldots m];$$

Darstellung für $m = 31, M = 10, RS = 5 \cdot LS$

versperrt ist. In den folgenden Abschnitten werden deshalb Erweiterungen vorgestellt, die es ermöglichen, den gesamten sich bietenden Aktionsspielraum auszuschöpfen.

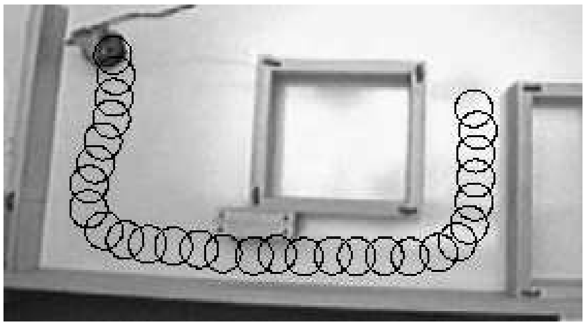

Abbildung 19: *Aufzeichnung einer KHEPERA-Fahrt mittels einer farbigen Markierung (Kreis = KHEPERA-Grundfläche, Bild = Startposition): Das mit der ALVINN-Architektur erreichbare (weil trainierbare) Fahrverhalten: In Gassen richtet sich das Fahrzeug mittig aus, ist nur eine Wand sichtbar, so wird dieser mit einem mittiger Flurfahrt entsprechendem seitlichen Abstand gefolgt.*

Da dieses Problem bei der MILVA-Roboterplattform mit ihrer 3-Rad-Kinematik auch wegen des Schleppkurvenverhaltens in ähnlicher Weise auftritt (Abb. 20), mußten bereits im KHEPERA-Szenario Strukturen entwickelt werden, die auch hier durch eine sensomotorische Abbildung ein adäquates Verhalten erzeugen. Dabei sollten Architekturen gefunden werden, die die zurückliegende Information bis zu einem festen, den Fahrzeugabmessungen entsprechenden **Weg**-Horizont gleichberechtigt berücksichtigen.

Wie in Abbildung 21 deutlich wird, bietet sich gegenüber der Einführung von rekurrenten Verbindungen oder dynamischen Neuronen (z.B. mit PT_1-Verhalten) die Einrichtung eines **Wegfensters** an, in dem zusätzlich die Inputvektoren zurückliegender Situationen mitgeführt werden. Diese Struktur stellt sich als eine vereinfachte Anwendung der „Sliding-Window"-Technik in TDNN-Netzen nach (Waibel, A. and Hampshire, J., 1989) dar. Wichtig ist bei einer solchen Erweiterung der Netzwerk-Input-Schicht, daß sich der Versatz der gleichzeitig dargestellten Situationen auf eine konstante Wegbasis bezieht. Um dies auch bei variabler Geschwindigkeit zu realisieren, werden bis zu einem bestimmten Horizont die Inputvektoren gemeinsam mit den zugehörigen Schrittdistanzen zwischengespeichert, von denen die vom aktuellen Standpunkt aus definiert zurückliegenden Wegpunkte ausgewählt werden.

In der Anwendung zeigte diese Erweiterung die erhofften Erfolge nicht nur im Ausrichtungsverhalten zu den Hindernissen bzw. Wegverbreiterungen im erwünschten Abstand (Abb. 22), sondern aufgrund der in der History zusätzlich repräsentierten Information auch in einer Reduktion von Konfliktsituationen und einer Tolerierung

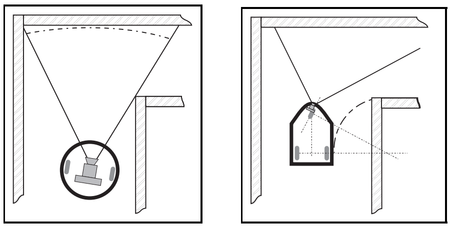

Abbildung 20: *Der schmale Sichtbereich der KHEPERA-Kamera (Öffnungswinkel 60^o) schränkt die erreichbare Verhaltensleistung erheblich ein (links). Vergleichbare Verhältnisse bei MILVA (rechts).*

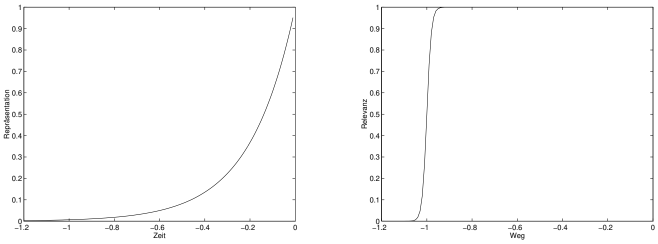

Abbildung 21: *Während rückgekoppelte Netzwerkstrukturen zurückliegende Information vergleichbar mit PT_1-Verhalten repräsentieren (links) bleibt die Relevanz von Bildinformation für ein kollisionsvermeidendes Verhalten über einen den Fahrzeugabmessungen entsprechenden Weg gleichbleibend hoch und bricht erst danach vollständig ein (rechts).*

kurzzeitigen Verlusts des Untergrund-Sichtkontakts. Bewährt hat sich dabei ein Wegfenster der Tiefe 3 mit einer Wegdifferenz von 3 cm (entspricht halbem KHEPERA-Durchmesser). Zusätzliche Vergangenheitsebenen erhöhen zu stark die Anzahl der zu adaptierenden Wichtungen, weiter auseinanderliegende Ansichten variieren zu stark, um sie durch ein Netzwerk in einem sensomotorischen Zusammenhang nachvollziehen zu können.

Grundsätzlich wird mit der Einführung des Wegfensters die sensorielle Basis für die Ausschöpfung aller mög-

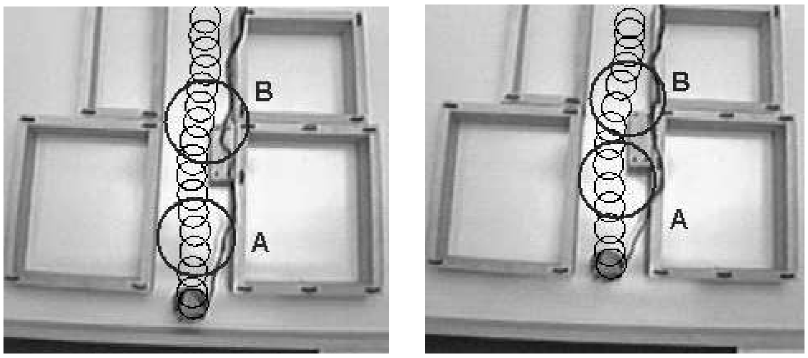

Abbildung 22: *Aufzeichnung der Fahrtrajektorien bei einer Flureinengung: Mit dem Netzwerk ohne History-Input ist ein frühes Einlenken (A) notwendig und die zu frühe Rückkehr zur Flurmitte ist nicht zu vermeiden (B) (links). Mit der zusätzlichen Information des Wegfensters erfolgen sowohl die Ausweich- (A) als auch die Rückkehrbewegung (B) mit dem erwünschten Abstand zum Hindernis (rechts).*

lichen Aktionsalternativen (der Vorwärtsfahrt) geschaffen, da nun auch an Kreuzungen und Abzweigungen abgebogen werden kann, auch wenn die seitlichen Einfahrten im aktuellen Bild nicht mehr vollständig repräsentiert sind. Damit ist es nun möglich, mehrere, auf verschiedenen Verhaltenszielen basierende Lernmengen zu erstellen, mit der die intentionale Auslegung der verschiedenen Agenten erreicht wird.

Geschwindigkeitsdimension und Aktivierungsfokus Parallel zu der oben beschriebenen Input-Erweiterung wurde mit der Einführung der Geschwindigkeit der Aktionsraum um eine Dimension erweitert, um auch das Geschwindigkeitsverhalten des Trainers während der Fahrt nachzubilden. Der bisherige m-dimensionale Output-Vektor wurde deshalb in einen m$\times$n-dimensionalen umgewandelt, auf dem nun in einer zweidimensionalen (ca. M$\times$N breiten) GAUSS-Funktion Lenkwinkel und Geschwindigkeit abgebildet werden (Gl. 2, Abb.

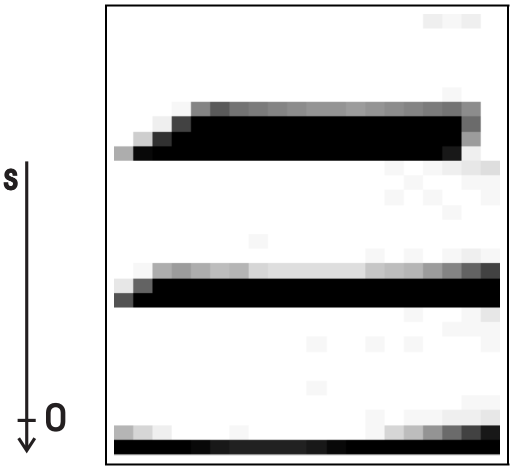

Abbildung 23: *Wegfenster aus 3 Inputvektoren (jeweils 10 Bildzeilen): Während der aktuelle Inputvektor (ganz unten) eine mehrdeutige Situation darstellt, ist durch die zurückliegenden Inputvektoren die Rechtsabbiegung erkennbar.*

24).

$$y_{ij} = e^{-\frac{\left(i-\frac{m+1}{2}\cdot(1+\frac{LS-RS}{LS+RS})\right)^2}{M}} \cdot e^{-\frac{\left(j-\frac{N}{2}-\frac{LS+RS}{LS_{max}+RS_{max}}\cdot(n-N)\right)^2}{N}} \qquad i \in [1\ldots m]; j \in [1\ldots n] \tag{2}$$

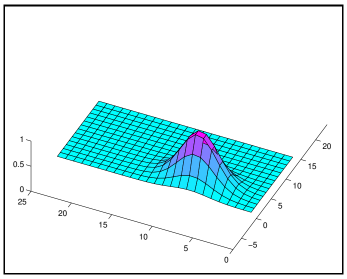

Abbildung 24: *2–dimensionale topologische Aktionskodierung für langsames Rechtsabbiegen als ca. 10×5 breite GAUSS-Glocke auf einem 25×15-dimensionalen Outputvektor (in Flächendarstellung).*

Aus methodischer Sicht bietet die 2-dimensionale Repräsentation gegenüber einer denkbaren 2-fach eindimensionalen Abbildung von Geschwindigkeit und Lenkeinschlag entscheidende Vorteile, weil sich über der Fläche der Output-Neuronen mehrere Handlungsvorschläge **separierbar** darstellen lassen. So läßt sich mit der erhaltenen Ausgabeaktivierung in geeigneter Weise ein dynamisches neuronales Feld nach (Amari, S., 1977) und (Kopecz, 1996) anregen, mit dem genau die Effekte erreicht werden können, die eine „intelligente" Auswertung des Netzoutputs beinhaltet: (I) Durch Mechanismen der Maximumsselektion wird auf die Region fokussiert, die sich sowohl in ihrer räumlichen Ausprägung als auch in ihrer Aktivität von ihrer Umgebung am stärksten abhebt. Alle nicht unterstützten Bereiche werden in ihrer Aktivität inhibiert. (II) Nach der Selektion eines lokalen energetischen Maximums wird dieses auch dann weiter fokussiert, wenn dessen Aktivität in einem begrenzten

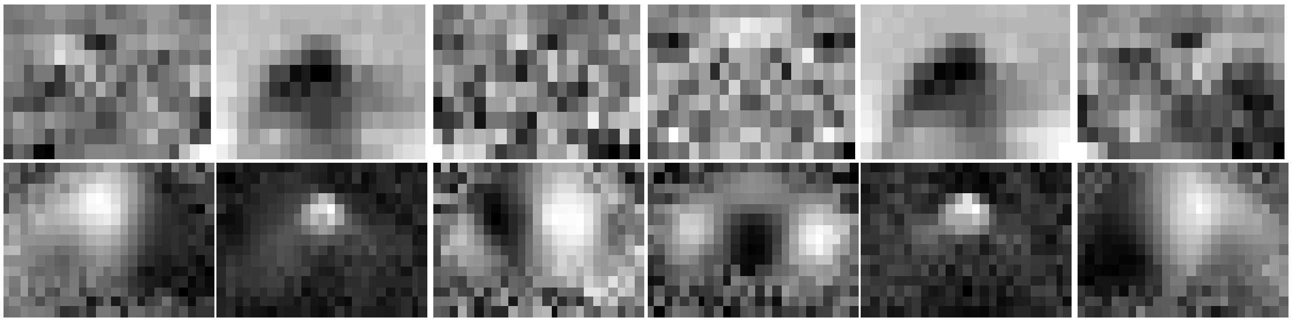

Abbildung 25: *Die afferenten (oben) und efferenten (unten) Wichtungen der 6 Hidden-Neuronen eines trainierten Netzwerks mit 2–dimensionaler Outputkodierung, aber ohne Wegfenster: Während in den Wichtungsmustern der Inputschicht (obere Zeile) sich deutlich die Struktur des Input-Datenstroms wiederspiegelt (vergl. Abb. 17) zeigt die Outputschicht (untere Zeile) starke exzitatorische (hell) und inhibitorische (dunkel) Wichtungen in den typischen Regionen des Aktionsraumes (Stopaktion bei v=0,ϕ=0: in den Aktionskarten mittig unten).*

Maße unter die anderer absinkt. Durch die Hystereseeigenschaft ergibt sich ein gewisses Beharrungsvermögen, so daß auch bei kurzzeitigen Einbrüchen der lokalen Aktivierung der Fokus nicht verloren wird. (III) Auch bei einer sich verschiebenden lokalen Aktivierung wird deren Zentrum weiter abgebildet. Damit können die räumlich nie stationären „Blobs“ in der Outputebene des Netzwerks stabil verfolgt werden (Tracking) (Krabbes, M. et al., 1997a).

Die besten Resultate in der Auswertung der Ausgabeaktivierung der einzelnen Agenten wurden erzielt, wenn in den durch das nachgeschaltete dynamische neuronale Feld selektierten Regionen die Position des maximal aktivierten Neurons zur Erzeugung des Fahrbefehls herangezogen wurde. Die Fusionierung der Aktionsvorschläge mehrerer Agenten kann in ähnlicher Weise erfolgen, indem der Eintrag deren Ausgabeaktivierungen in eine gemeinsame Karte durch die jeweiligen neuronalen Felder regional gebahnt wird. Über der resultierenden „Gesamterregung“ wird durch ein weiteres dynamisches neuronales Feld wiederum die energiereichste Region selektiert, in der die maximale Aktivierung die letztendlich auszuführende Aktion bestimmt.

Ergebnisse Um eine erste Demonstration des Zusammenwirkens des hier vorgestellten Agentenkollektivs zu realisieren, wurde vergleichbar den Braitenberg-Vehikel (Braitenberg, V., 1984) eine „Such“-Funktionalität implementiert, indem ein farbiges Objekt entsprechend seiner horizontalen Position im Kamerabild in eine entsprechende Agentenmodulation umgesetzt wurde: Proportional zum Abstand des „zu suchenden“ Objekts von der Bildmitte wurde der diese Verschiebung kompensierende Agent (wenn rechts, dann „turn right“; wenn links, dann „turn left“) in seinem Aktivierungseintrag in die gemeinsame Aktionskarte bis zu 4-fach verstärkt. Zur Verfolgung des Suchobjektes wurde ebenfalls ein (hier eindimensionales) dynamisches neuronales Feld vorteilhaft angewendet. Durch geeignete Parametrierung wird dabei ein weiteres Verhalten nutzbar: die Erregung in den äußeren Neuronen des neuronalen Felds wird bei einem Herauslaufen des Objektes aus dem Bildrand solange erhalten, bis neue Inputerregung den Fokus auf sich zieht. Durch diesen Speichereffekt sucht der Roboter weiter in die Richtung, in der er das Objekt aus dem Blickfeld verliert.

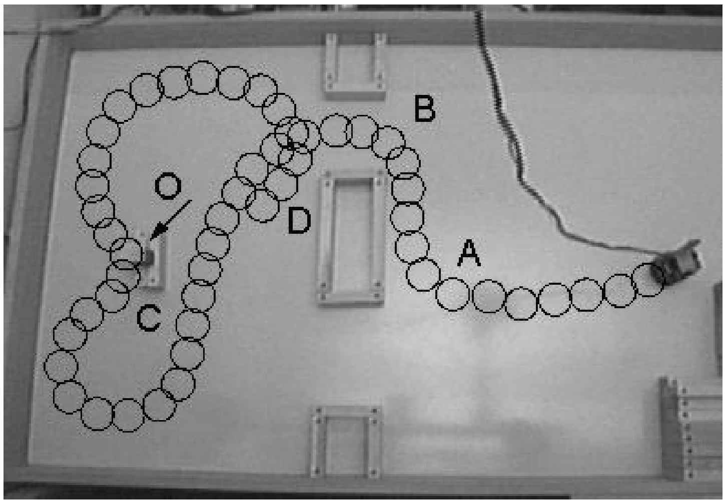

Abbildung 26: *Das „Suchverhalten“ des KHEPERA wird in der Aufzeichnung deutlich sichtbar (Zielobjekt links auf einem Hindernis stehend – O): dem Hindernis (verdeckt nicht das Zielobjekt) wird im letzten Moment ausgewichen, danach wird immer wieder auf das Objekt zugefahren, ohne aber damit zu kollidieren, weil es selbst als Hindernis wahrgenommen wird.*

Abbildung 26 verdeutlicht verschiedene Aspekte der Wechselwirkungen zwischen den Agenten: Die wechselseitige Verstärkung der beiden Abbiegeagenten verursacht ein auf das Objekt gerichtete, aber unstetige Bewegung, wodurch dicht vor dem Hindernis letztendlich der „turn left“-Agent zu einer Ausweichbewegung nach rechts gezwungen ist. Der vermeintlich kürzere Weg wird durch die erzeugte Systemreaktivität nicht zwingend gefunden (A). Da der „turn left“-Agent verstärkt bleibt, wird die Einfahrt nach links genutzt (B). Das Zielobjekt selbst stellt ebenfalls ein Hindernis dar und verursacht so eine Ausweichbewegung, die eine Kollision verhindert (C). Nach der Ausweichbewegung kehrt der Roboter wieder zum Zielobjekt zurück (D). Die großen Kurvenradien auf freien Flächen entsprechen dem Trainingsverhalten in solchen Situationen, weil keine Objekte zur Beobachtung der Eigenbewegung vorhanden sind und bei engeren Kurvenradien mit Objekten im toten Blickwinkel kollidiert werden kann.

2.2.2 Umsetzung der konzipierten Architektur auf die MILVA-Plattform

Neuimplementierung unter GRASIM Zu den wesentlichen Ergebnissen des GESTIK-Projektes muß die Realisierung der grafischen Entwicklungs- und Simulationsumgebung GRASIM gezählt werden, zu der die Projektbearbeiter wichtige Impulse beigetragen haben.

Die bisher beschriebene Architektur wurde zunächst auf dem Miniaturroboter Khepera implementiert. Dabei konnte deren prinzipielle Funktionalität im Khepera-Labyrinth nachgewiesen werden.

Die Umsetzung auf die MILVA-Plattform wurde erst zum Projektende in Angriff genommen und wird derzeit im Rahmen von studentischen Arbeiten fortgesetzt. Dabei ergaben sowohl hinsichtlich der veränderten technisch-geometrischen Gegebenheiten der Plattform als auch hinsichtlich der weitaus komplexeren Umgebungsbedingungen eine Reihe neuer Probleme. Während die Anpassung der Architektur an die MILVA-Gegebenheiten unproblematisch ist, erfordert die robuste Vorverarbeitung der Bilddaten, die notwendig wird, um für die zur Navigation eingesetzten Netzwerke weitgehend reproduzierbare Eingangsinformation zu garantieren, einen enormen Entwicklungsaufwand, der im Rahmen des GESTIK-Projektes nicht mehr umgesetzt werden konnte. Die stabilen Gegebenheiten des Khepera-Labyrinths mit blauem Untergrund, holzfarbenen Wänden und die weitestgehend mögliche Eliminierung störender Einflüsse wie Beleuchtungsschwankungen oder Schattenwürfe erforderten im Gegensatz dazu keine aufwendige Bildvorverarbeitung (siehe auch Abbildung 17), da sich eine Trennung in befahrbaren Untergrund und Hindernisse quasi implizit aus den Bilddaten ergab.

Erste Ergebnisse zur Umsetzung der konzipierten Architektur auf der MILVA-Plattform einschließlich der notwendigen Methoden zur Bildvorverarbeitung können in (Weber, S., 1998) nachgelesen werden. Sie stellen im eigentlichen Sinn keine Resultate des GESTIK-Projektes mehr dar, da sie erst nach Auslaufen der Projektförderung in Angriff genommen wurden.

Integration der gestikulierten Navigationsanweisungen Die Einbeziehung der Nutzer-Intention in das aktuelle Fahrverhalten konnte im Rahmen des GESTIK-Projektes nicht mehr praktisch umgesetzt werden. Die im Teilprojekt 2 entwickelte und in Abbildung 16 dargestellte Gesamtarchitektur zur adaptiven Verhaltensorganisation soll jedoch die Grundlage für die Integration der beiden Teilprojekte bilden. Neben der festen Abbildung der erkannten Kommando-Posturen auf die entsprechenden Navigationsleistungen (zum Nutzer fahren, anhalten, nach rechts bzw. links abbiegen etc.) sind insbesondere die im folgenden benannten Eigenschaften dieser Architektur für die erfolgreiche Integration der beiden Teilprojekte wesentlich:

- Alle Module einschließlich der Nutzer-Intention verwenden zweidimensionale Aktionskarten als einheitliche Repräsentation der Aktionsvorschläge, was deren Fusion und die Auswahl eines resultierenden Aktionsvorschlags wesentlich vereinfacht.
- Die eigentliche Selektion stellt das Resultat eines dynamischen, nichtlinearen Wechselwirkungsprozesses innerhalb von Neuronenfeldern dar. Dieser Wechselwirkungsprozeß wurde theoretisch gut aufgearbeitet und bietet aufgrund der einfachen Handhabbarkeit bei entsprechender Parameterwahl ein breites Spektrum an Möglichkeiten, das resultierende Navigationsverhalten auf die konkreten Gegebenheiten abzustimmen.
- Die konzipierte Architektur stellt implizit sicher, daß die Aktionsvorschläge der einzelnen Navigations-Agenten immer erhalten bleiben und nur solche Aktionen ausgeführt werden können,
  - die in der gegebenen sensorischen Situation überhaupt ausführbar und
  - die zur aktuellen Nutzer-Intention adäquat sind.

Literatur

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87.

Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. (1998a). Neural Architecture for Gesture-Based Human-Machine-Interaction. In *Gesture and Sign-Language in Human-Computer Interaction*, Lecture Notes in Artificial Intelligence, pages 219–232. Springer.

Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Krabbes, M., Corradini, A., and Gross, H.-M. (1998b). User Localisation for Visually-based Human-Machine-Interaction. In *International Conference on Automatic Face- and Gesture Recognition*, pages 486–491. IEEE Computer Society Press.

Boehme, H.-J., Braumann, U.-D., Corradini, A., and Gross, H.-M. (1999). Person Localization & Posture Recognition for Human-Robot Interaction. In *GW'99 - The 3rd Gesture Workshop, Gif-sur-Yvette, France*.

Braitenberg, V. (1984). *Vehikel - Experimente mit kybernetischen Wesen*. Rowohlt.

Brakensiek, A., Braumann, U.-D., Boehme, H.-J., and Gross, H.-M. (1998). Neuronale Verfahren zur Mensch-Roboter-Interaktion: Personenlokalisation und Gestenerkennung. In *43. Internationales Wissenschaftliches Kolloquium der TU Ilmenau*, pages 633–638.

Darrell, T., Basu, S., Wren, C., and Pentland, A. (1997). Perceptually-driven Avatars and Interfaces: active methods for direct control. In *SIGGRAPH'97*. MIT Media Lab Perceptual Computation Section, TR 416.

Grossberg, S. (1969). Some Networks That Can Learn, Remember, and Reproduce Any Number of Complicated Space-Time Patterns. *Journal of Mathematics and Mechanics*, 19(1):53–99.

Kohler, M. (1998). Vision Based Gesture Recognition Systems. `http://ls7-www.informatik.uni-dortmund.de/html/englisch/gesture/vbgr-table.html`.

Kopecz, K. (1996). Neural field dynamics provide robust control for attentional resources. In *Aktives Sehen in technischen und biologischen Systemen*, pages 137–144. Infix-Verlag.

Kortenkamp, D., Huber, E., and Bonasso, P.R. (1996). Recognizing and interpreting gestures on a mobile robot. In *Thirteenth National Conference on Artificial Intelligence (AAAI-96)*.

Krabbes, M., Boehme, H.-J., Stephan, V., and Gross, H.-M. (1997). Extension of the ALVINN-Architecture for Robust Visual Guidance of a Miniature Robot. In *EUROBOT'97 - 2nd Euromicro Workshop on Advanced Mobile Robots*, pages 8–14. IEEE Computer Society Press.

Littmann, E. and Ritter, H. (1995). Neural and Statistical Methods for Adaptive Color Segmentation - A Comparison. In *Mustererkennung 1995*, pages 84–93.

Maggioni, C. and Kämmerer, B. (1996). GestureComputer – History, Design, and Applications. In *Proceedings of the Workshop on Computer Vision in Man-Machine Interfaces*.

Oren, M., Papageorgiou, C., Osuna, E., and Poggio, T. (1997). Pedestrian Detection Using Wavelet Templates. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 193–199.

Pomerleau, D. (1993). *Neural Network Perception for Mobile Robot Guidance*. Kluwer Academic Publishers.

Pomierski, T. and Gross, H.-M. (1996). Biological Neural Architectures for Chromatic Adaptation resulting in Constant Color Sensations. In *ICNN'96, IEEE International Conference on Neural Networks*, pages 734–739. IEEE Press.

Rigoll, G. and Kosmala, A. (1997). New improved feature extraction methods for real-time high performance image sequence recognition. In *ICASSP'97*.

Rowley, H. A., Baluja, S., and Kanade, T. (1995). Human Face Detection in Visual Scenes. Technical report, Carnegie Mellon University.

Schiele, B. and Waibel, A. (1995). Gaze Tracking Based on Face-Color. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 344–349.

Waibel, A. and Hampshire, J. (1989). Building blocks for Speech. *Byte*, (8):235–242.

Weber, S. (1998). Monocular Visual Robot Navigation in Indoor Environments. Technical report, TU Ilmenau, Fachgebiet Neuroinformatik. Report 1/99, ISSN 0945-7518.

Yang, J. and Waibel, A. (1997). Skin-Color Modeling and Adaptation. Technical report, Carnegie Mellon University.

2.3 Verwertbarkeit und voraussichtlicher Nutzen der Ergebnisse

Die aus der Sicht des Fachgebiets Neuroinformatik bzgl. der Förderrichtlinie und bzgl. der Methodenentwicklung wesentlichen Ergebnisse sind im folgenden stichpunktartig zusammengefaßt:

Ergebnisse bzgl. der Förderrichtlinie

- Das Fachgebiet Neuroinformatik konnte gemeinsam mit 5 weiteren Fachgebieten aus 3 Fakultäten einen Antrag auf die Einrichtung eines Graduierten-Kollegs zum 1.10.1999 stellen. Die Thematik -Intelligente und interaktive Mensch-Maschine-Schnittstellen" bildet eine unmittelbare Fortsetzung der im GESTIK-Projekt gewonnenen Erkenntnisse.
- Zum BMBF-Ideenwettbewerb „Mensch-Technik-Interaktion in der Wissensgesellschaft" wurde vom Fachgebiet gemeinsam mit 5 weiteren beteiligten Partnern (Lehrstühle Prof.v.Seelen, Prof.v.d.Malsburg des Institutes für Neuroinformatik der Ruhr-Universität Bochum, Stinnes BauMarkt AG Esslingen, Neurotec/mediaCenter GmbH Friedrichshafen, Optimess GmbH Gera) eine Ideenskizze eingereicht, die jedoch nicht für die finale Antragstellung ausgewählt wurde, da das Konsortium "Intelligente anthropomorphe Assistenzsysteme" unter Führung des Forschungszentrums für Anwendungsorientierte Wissensverarbeitung Ulm inhaltlich ähnlich gelagert und hinsichtlich der industriellen Beteiligung deutlich mächtiger ist.
- In Vorbereitung von Drittmittel-Vorhaben konnte eine bislang sehr konstruktive und fruchtbare Kooperationsbeziehung zur Stinnes BauMarkt AG in Esslingen als potentiellem Endanwender intelligenter Service-Systeme aufgebaut werden. Mit der Filiale in Erfurt konnten bereits erste Voruntersuchungen und Tests der entwickelten Verfahren für solche Systeme durchgeführt werden. Die Erfurter Filiale wird sich auch in ein wiederum beim TMWFK eingereichtes Vorhaben, welches sich vordergründig mit der Realisierung eines robusten Zuwendungsverhaltens für intelligente Service-Systeme beschäftigen soll, als assoziierter Partner einbringen.

Methodenentwicklung

- Für die visuelle Personenlokalisation unter Realwelt-Bedingungen konnte ein personenspezifisches Auffälligkeitssystem (Multi-Cue-Ansatz in Verbindung mit 3-dimensionalen neuronalen Feldern zur Hypothesenselektion) entwickelt und praktisch umgesetzt werden, welches dem aktuellen „state-of-the-art" entspricht. Bzgl. der Realisierung der Teilmodule zur Personenlokalisation sollen insbesondere die folgenden Teilleistungen hervorgehoben werden:

- Für die sichere Detektion der Hände einer Person wurde ein Verfahren entwickelt, welches die Hautfarbe des bereits lokalisierten Gesichts als implizite Referenz für die Erstellung eines personen- und situationsspezifischen Farbmodells heranzieht.
- Die entwickelten Verfahren zur Gesichtsstrukturdetektion und zur Kopf-Schulter-Partie-Detektion bilden einen sehr guten Kompromiß zwischen Berechnungskomplexität und Leistungsfähigkeit.
- Ausgehend von der Theorie zweidimensionaler dynamischer neuronaler Felder wurde ein dreidimensionaler Selektionsprozeß innerhalb einer Multiskalen-Repräsentation umgesetzt, der die Teilergebnisse auf jeweils benachbarten Auflösungsebenen zu einem robusten Selektionsergebnis fusioniert.

• Die visuell-basierte Indoor-Navigation mobiler Systeme stellt noch immer eine große wissenschaftliche Herausforderung dar und wird derzeit noch verhältnismäßig wenig angewendet. Bzgl. der im Rahmen des GESTIK-Projektes durchgeführten Forschungsarbeiten sind die folgenden Punkte aus Sicht des Fachgebiets wesentlich:

  - Ausgehend vom ALVINN-Ansatz konnte eine neuronale Architektur zur Handlungsorganisation intentionaler Agenten konzipiert und auf dem Khepera-Roboter demonstriert werden, die neben der unmittelbaren sensomotorischen Abbildung auch die Einbeziehung der aktuellen Nutzer-Intention in die Generierung des Fahrverhaltens einbeziehen kann.
  - Für den Einsatz im Indoor-Bereich wurde eine Reihe von Erweiterungen zur ALVINN-Architektur erarbeitet. Dies umfaßte insbesondere die Einführung eines geeigneten Wegfensters am Netzwerkeingang, die Art der Aktionskodierung und die letztendliche Aktionsauswahl, die auch hier auf einer Felddynamik miteinander wechselwirkender Neuronen beruht.

• Das am Fachgebiet laufende EU-TMR-Projekts DYGEST konnte nutzbringend in das GESTIK-Projekt mit einbezogen werden. Hervorgehoben werden sollen dabei insbesondere die Arbeiten zur Personenlokalisation und zur Posture-Erkennung, die in enger Kooperation zwischen beiden Projekten entstanden.

Die Ergebnisse zur Methodenentwicklung finden ihren Ausdruck in zahlreichen hochrangigen und durchweg referierten Publikationen, die während der Projektlaufzeit entstanden.

Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. (1997). Visually-Based Human-Machine-Interaction in a Neural Architecture. In *Proceedings SOAVE'97 - Selbstorganisation von adaptivem Verhalten*, volume 8 of *Fortschrittsberichte*, pages 166–175. VDI Verlag.

Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., and Gross, H.-M. (1998a). Neural Architecture for Gesture-Based Human-Machine-Interaction. In *Gesture and Sign-Language in Human-Computer Interaction*, Lecture Notes in Artificial Intelligence, pages 219–232. Springer.

Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Krabbes, M., Corradini, A., and Gross, H.-M. (1998c). Neural Networks for Gesture-based Remote Control of a Mobile Robot. In *International Joint Conference on Neural Networks*. IEEE Computer Society Press.

Boehme, H.-J., Braumann, U.-D., Brakensiek, A., Krabbes, M., Corradini, A., and Gross, H.-M. (1998b). User Localisation for Visually-based Human-Machine-Interaction. In *International Conference on Automatic Face- and Gesture Recognition*, pages 486–491. IEEE Computer Society Press.

Boehme, H.-J., Braumann, U.-D., Corradini, A., and Gross, H.-M. (1999). Person Localization & Posture Recognition for Human-Robot Interaction. In *GW'99 - The 3rd Gesture Workshop, Gif-sur-Yvette, France*.

Boehme, H.-J., Gross, H.-M., Braumann, U.-D., Heinke, D., Pomierski, T., and Brakensiek, A. (1996). Active-vision zur Selbstorganisation von Verhalten in senso-motorischen Systemen. In *Aktives Sehen in technischen und biologischen Systemen*, pages 60–66. infix-Verlag.

Brakensiek, A., Braumann, U.-D., Boehme, H.-J., and Gross, H.-M. (1998a). Neuronale Verfahren zur Mensch-Roboter-Interaktion: Personenlokalisation und Gestenerkennung. In *43. Internationales Wissenschaftliches Kolloquium der TU Ilmenau*, pages 633–638.

Brakensiek, A., Braumann, U.-D., Böhme, H.-J., Rieck, C., and Groß, H.-M. (1997). Farb- und strukturbasierte, neuronale Verfahren zur Lokalisation von Gesichtern in Real-World-Szenen. In *Mustererkennung 1997, 19. DAGM-Symposium Braunschweig*, Informatik aktuell, pages 113–120. Springer Verlag.

Brakensiek, A., Braumann, U.-D., Corradini, A., Böhme, H.-J., and Gross, H.-M. (1998b). Neuronale Verfahren zur Lokalisation und Bewertung von Handgesten in Real-World-Szenen. In *3. Internationaler Workshop Neuronale Netze in der Anwendung*. Technische Universitaet Magdeburg.

Braumann, U.-D., Brakensiek, A., Boehme, H.-J., and Gross, H.-M. (1998a). Gaborfilterbasierte visuelle Personenlokalisation mit dreidimensionalen Feldern dynamischer Neuronen. In *3. Internationaler Workshop Neuronale Netze in der Anwendung*, pages 31–38.

Braumann, U.-D., Brakensiek, A., Böhme, H.-J., and Groß, H.-M. (1997). Ein neuronales Verfahren zur Segmentierung hautfarbener Bildregionen. In *Proceedings 3. Workshop "Farbbildverarbeitung", Erlangen*, pages 67–72. IRB-Verlag Stuttgart.

Braumann, U.-D., Corradini, A., Böhme, H.-J., and Groß, H.-M. (1998b). Konturbasierte Personenlokalisation mittels dreidimensionaler neuronaler Felder und steuerbarer Filter. In *Mustererkennung 1998, 20. DAGM-Symposium*, Informatik aktuell, pages 385–392. Springer Verlag.

Corradini, A., Braumann, U.-D., Boehme, H.-J., and Gross, H.-M. (1998a). 3D Neural Fields and Steerable Filters for Contour-Based Person Localization. In *Proceedings Workshop on Virtual Intelligence - Dynamic Neural Networks, VI-DYNN'98*. in press.

Corradini, A., Braumann, U.-D., Brakensiek, A., Krabbes, M., Boehme, H.-J., and Gross, H.-M. (1998b). Visual Person Localization with Dynamic Neural Fields: Towards a Gesture Recognition System. In *Proceedings X. Italian Workshop on Neural Nets (WIRN98)*. in press.

Gross, H.-M. and Böhme, H.-J. (1997). Verhaltensorganisation in interaktiven und lernfähigen mobilen Systemen – das MILVA Projekt. In *SOAVE'97 - Selbstorganisation von adaptivem Verhalten*, volume 8 of *Fortschrittsberichte*, pages 1–13. VDI-Verlag.

Gross, H.-M., Stephan, V., and Boehme, H.-J. (1996). Sensory-based Robot Navigation using Self-organizing Networks and Q-learning. In *WCNN'96, World Congress on Neural Networks*, pages 94–99. Lawrence Erlbaum Associates.

Krabbes, M., Boehme, H.-J., Stephan, V., and Gross, H.-M. (1997a). Extension of the ALVINN-Architecture for Robust Visual Guidance of a Miniature Robot. In *EUROBOT'97 - 2nd Euromicro Workshop on Advanced Mobile Robots*, pages 8–14. IEEE Computer Society Press.

Krabbes, M., Boehme, H.-J., Stephan, V., and Gross, H.-M. (1997b). Handlungsorganisation intentionaler neuronaler Agenten. In *SOAVE'97 - Selbstorganisation von adaptivem Verhalten*, volume 8 of *Fortschrittsberichte*, pages 35–44. VDI-Verlag.

Krabbes, M., Boehme, H.-J., Stephan, V., and Gross, H.-M. (1997c). Handlungsorganisation mittels intentionaler neuronaler Agenten. In *AMS'97 - 13. Fachgespräch Autonome Mobile Systeme*, pages 154–65. Springer-Verlag.

Krabbes, M., Weber, S., Boehme, H.-J., and Gross, H.-M. (1998). Monokulare visuelle Hindernisdetektion auf Basis merkmalsbasierter Bildsegmentierung. In *AMS'98 - 14. Fachgespräch Autonome Mobile Systeme*, Informatik aktuell, pages 85–92. Springer.

Im Berichtszeitraum wurden keine Erfindungen gemacht.

2.4 Im Projektverlauf bekanntgewordene Ergebnisse Dritter

Soweit projektrelevant, wurden die entsprechenden Ergebnisse Dritter zu Teilsystemen der Gesamtarchitektur im Abschnitt 2 bereits aufgeführt. Was die im GESTIK-Projekt angestrebte und zumindest teilweise realisierte Gesamtarchitektur anbetrifft, so ergeben sich Parallelen zu

- den Arbeiten von KORTENKAMP (NASA, USA), in denen ebenfalls ein vordefiniertes Alphabet gestikulierter Kommandos zur Robotersteuerung eingesetzt wird,

- dem Robotik-Institut an der Carnegie Mellon University Pittsburgh (USA) und der Robotik-Gruppe der Universität Bonn, die sich bspw. mit dem Einsatz intelligenter mobiler Systeme als Museumsführer beschäftigen, sowie
- dem Institut für Neuroinformatik der Ruhr-Universität Bochum und dem FAW Ulm, die sich ebenfalls der Entwicklung intelligenter mobiler Service-Systeme widmen.

Die Parallelen beziehen sich jedoch auch hier mehr auf den Einsatz bestimmter Verfahren und Teilmethoden und weniger auf die eigentliche Architektur. Die NASA-Gruppe ausgenommen, besteht zu allen genannten Einrichtungen seitens des Fachgebiets Kontakt und Erfahrungsaustausch.

2.5 Probleme und Schwierigkeiten bei der Erlangung des Projektziels

Naturgemäß traten Probleme und Schwierigkeiten bei der Erlangung des Projektziels auf, was in erster Linie in den folgenden Punkten begründet war:

- Da zu Projektbeginn seitens des Fachgebiets nur wenig Erfahrung bzgl. der Komplexität der zu bearbeitenden Themenstellung vorlag, erwies sich der im Antrag angestrebte Umfang als teilweise unrealistisch.
- Insbesondere die von Beginn an durchgesetzte starke Orientierung auf die Realwelt-Tauglichkeit der entwickelten Methoden erwies sich
  - als Nachteil, was einen möglichst schnellen Methodenfortschritt anbetrifft,
  - jedoch als ausgesprochener Vorteil, was die Funktionsfähigkeit, Robustheit und Demonstrabilität der Gesamtarchitektur für den realen Einsatz unter variablen Umgebungsbedingungen anbelangt.
- Die für den Realwelt-Einsatz erforderliche Echtzeitfähigkeit der entwickelten Verfahren erforderte einen deutlich höheren Mehraufwand als dies zur Antragstellung erwartet wurde. So ließen sich bspw. die ursprünglich favorisierten Eigenvektor-Verfahren zur Gesichtsstrukturdetektion nicht wie gewünscht umsetzen. Der notwendige Gesamtaufwand zur Personenlokalisation erwies sich insgesamt als viel umfangreicher, als das zum Zeitpunkt der Antragstellung erwartet worden war.
- Zum Ende des GESTIK-Projektes und darüber hinaus ergab sich die Notwendigkeit aufwendiger Verfahren zur Bildvorverarbeitung, um eine robuste Funktionalität des entwickelten erweiterten ALVINN-Ansatzes in der vorliegenden Indoor-Umgebung sicherzustellen. Dies war zum Zeitpunkt der Antragstellung zumindest in diesem Maße nicht vorherzusehen und liegt ebenfalls in der fehlenden diesbezüglichen Erfahrung begründet.
- Weiterhin treten beim Einsatz komplexer technischer Systeme in realen Umwelten zwangsläufig auch technisch bedingte Probleme auf, die vom Ausfall von Kameras und Framegrabbern, fehlenden Treibern bis hin zu Kabelbrüchen und dem Ausfall ganzer Module reichen und teilweise einen erheblichen Prozentsatz der täglichen Arbeitszeit der Bearbeiter gebunden haben. Dieser Aspekt ist bei weiteren Projekten in weitaus höherem Maße zu berücksichtigen, insbesondere was die Planung von Finanzmitteln für Material und Reparaturen als auch technische Mitarbeiter betrifft.

Prof. Dr.-Ing. H.-M. Gross
- Zuwendungsempfänger -