# Jahresbericht 2003
# Lehrstuhl für Mustererkennung

## Inhaltsverzeichnis

# 1 Mitarbeiter am LME

**Leitung:** **Prof. Dr.-Ing. H. Niemann**

**Professor:** **Prof. Dr.-Ing. J. Hornegger**

**Sekretariat:**
Iris Koppe
Kristina Müller

**Sekretariat SFB 603:**
Martina Montel-Kandy

**Bildverarbeitung:**
Dipl.-Inf. Frank Deinzer (bis 31.12.2003)
Dr.-Ing. Joachim Denzler (bis 31.8.2003)
Dipl.-Ing. Christian Derichs
Dipl.-Inf. Benjamin Deutsch
Dipl.-Inf. Rainer Deventer
Dipl.-Inf. Christopher Drexler (bis 28.2.2003)
Dipl.-Inf. Christoph Gräßl
Dipl.-Inf. Marcin Grzegorzek
Dipl.-Inf. Frank Mattern (bis 31.10.2003)
Kailash N. Pasumarthy, M.S.
Dipl.-Inf. Jochen Schmidt
Dipl.-Inf. Ingo Scholz
Dipl.-Inf. Florian Vogt
Dipl.-Math. (FH) Stefan Wenhardt
Dipl.-Inf. Timo Zinßer
Dr.-Ing. Matthias Zobel (bis 31.3.2003)

**Medizinische Bildverarbeitung:**
Dipl.-Inf. Marcus Prümmer

**Sprachverarbeitung:**
Dipl.-Inf. Johann Adelhardt
Dr. phil. Anton Batliner
Dipl.-Inf. Carmen Frank
Dipl.-Inf. Christian Hacker
Dipl.-Inf. Tino Haderlein
Dr.-Ing. Elmar Nöth, Akad. ORat

Shi, Ruiping, M.S.
Dipl.-Inf. Stefan Steidl
Dipl.-Inf. Georg Stemmer
Dipl.-Inf. Viktor Zeißler

**Lehrbeauftragte:**
Dr.-Ing. Ulf Haßler
Dr.-Ing. Thomas Wittenberg

**Nichtwiss. Personal:**
Walter Fentze
Friedrich Popp

# Gäste

| | | |
|---|---|---|
| Novikov, Konstantin | (Russ. Föderation, DAAD) | 01.12.02 – 28.02.03 |
| Osipova, Elena | (Russ. Föderation)<br>(Soroptimist Intern.)<br>(Deutsche Union, Abt.ER) | 15.02.03 – 30.04.03 |
| Levine, Kirill | (Russ. Föderation, LME) | 15.02.03 – 30.04.03 |
| El Bakry, M. Hazem | (Ägypten, Ägyptische Reg.) | 01.03.03 – 16.10.03 |
| Ipalakova, Madina | (Kasachstan, LME-Stipendium) | 01.07.03 – 30.09.03 |
| Parfyonov, Sergey | (Russ. Föderation, IAESTE) | 01.10.03 – 22.12.03 |
| Zhukov, Ilya | (Russ. Föderation, LME-Stip.) | 15.07.03 – 31.08.03 |
| Shidlovsky, Dmitry | (Russ. Föderation, LME-Stip.) | 15.07.03 – 31.08.03 |
| Ober, Jozef | (Polen, Erasmus/Socrates) | 23.06.03 – 26.06.03 |

# 2 Bildanalyse

Leitung: J. Denzler
(F. Deinzer, B. Deutsch, R. Deventer, C. Drexler, C. Gräßl, M. Grzegorzek, J. Schmidt, F. Mattern, I. Scholz, F. Vogt, T. Zinßer )

Schwerpunkt der Forschungstätigkeiten im Bereich der Bildanalyse am Lehrstuhl ist die statistische Objektmodellierung, –erkennung und Verfolgung, grundlagenorientierte Arbeiten zur optimalen Sensordatenauswahl und –fusion im aktiven Rechnersehen sowie Kamerakalibrierung und 3–D Rekonstruktion mit Anwendungen in der erweiterten und virtuellen Realität. Versuchs– und Anwendungsplattform ist projektübergreifend das autonome, mobile System MOBSY, in dem die verschiedenen Verfahren unter Echtzeitbedingungen und in realer, natürlicher Umgebung ihre Leistungsfähigkeit unter Beweis stellen müssen. Bildbasierte Modelle, wie der Lumigraph oder das Lichtfeld, die im Teilprojekt C2 des Sonderforschungsbereichs 603 entwickelt und erweitert werden, fließen in allen Bereichen als eine Alternative zu geometriebasierten Objekt– und Umgebungsmodellen ein.

Als weiterer Forschungsschwerpunkt hat sich der Bereich Rechnersehen für autonome mobile Systeme etabliert. Darunter fallen grundlagenorientierte Arbeiten auf dem Gebiet der probabilistischen Modellierung von Sensordaten– und Aktionsfolgen für das aktive Rechnersehen, optimale Kameraparameterauswahl für die Objekterkennung und –verfolgung sowie Eigenraumverfahren zur 3D–Objektlokalisation und Klassifikation. Bildbasierte Modelle, wie der Lumigraph oder das Lichtfeld, die im Teilprojekt C2 des Sonderforschungsbereichs 603 entwickelt und erweitert werden, fließen in allen Bereichen als eine Alternative zu geometriebasierten Objekt– und Umgebungsmodellen ein. Als Anwendungsszenario dient der Bereich der Service– und Dienstleistungsroboter. Dort wurde sowohl eine Objekterkennungskomponente für Pflegeroboter im Krankenhaus (Projekt DIROKOL) als auch in enger Kooperation mit der Sprachverarbeitung das mobile System MOBSY entwickelt, das während der 25–Jahr–Feier den Gästen als Empfangsdame zur Verfügung stand. Der grundlegende Versuchsaufbau für die Projekte der Bildanalyse besteht aus beweglichen rechnergesteuerten Kameras, die beispielsweise an der Hand eines Roboters montiert sind und dadurch im Arbeitsraum des Roboters frei positioniert werden können, oder rechnergesteuerten Multi–Media Farbkameras, welche die Szene durch gezielte Schwenk/-Neigebewegungen überwachen und Details von Objekten durch Änderung der Brennweite betrachten können. Zur kontrollierten Datenaufnahme, die beispielsweise bei der Erstellung von Stichproben erforderlich ist, existieren zwei rechnersteuerbare Aufbauten, die jeweils aus einem Drehteller und einem Schwenkarm bestehen. An dem Schwenkarm ist eine hochwertige Farb–Kamera befestigt, so dass von einem Objekt auf dem Drehteller Ansichten von einer beliebigen Position auf einer Halbkugel um das Objekt aufgenommen werden können.

Für die laufenden Projekte auf dem Gebiet der optimalen Sensordatenauswahl sowie auf auf dem Gebiet des Rechnersehens für autonome mobile Systeme steht seit Anfang 1998 das auf der Plattform XR4000 der Firma Nomadic basierende System Mobsy zur Verfügung. Die beiden auf der Plattform installierten Rechnersysteme (Pentium Pro und Dual Pentium III 850) ermöglichen eine vollständig Autonomie. Zum Rechnercluster des Lehrstuhls besteht eine Verbindung über ein Funkethernet. Die Plattform verfügt neben Infrarot–, Ultraschall– und mechanischen Senso-

ren über einen Stereo–Kopf mit Schwenk–Neige–Vergenz–Steuerung und Farbkameras zur visuellen Wahrnehmung der Umwelt sowie einem Greifer. Der Aufbau wurde um ein Touchscreen LCD–Display ergänzt, damit Interaktion (Starten von Demoprogrammen, Auswahl von Objekten, Debugging) direkt an der ansonsten vollständig autonom agierenden Plattform möglich wird. Desweiteren wurde die Konstruktion des Aufbaus so verändert, dass ohne größere technische Eingriffe, die Position des Stereokopfes, d.h. der Kameras, verändert werden kann. Diese gesteigerte Flexibilität ist in einem Projekt zum sichtbasierten Greifen eines Objekts notwendig, da sichergestellt werden muss, dass die Kamera zu jedem Zeitpunkt den Greifer der Plattform einsehen kann.

Das anlässlich des 25–jährigen Bestehens des Lehrstuhls für Mustererkennung enwickelte System Mobsy wurde weiterhin gewartet und bei zahlreichen Anlässen (Tag der Informatik, Erstsemestereinführung, Mädchenpraktkum) vorgeführt: Mobsy wartete im 9. Stock vor den Aufzügen, erkannte ankommende Gäste und nahme diese in Empfang. Danach gab er einen kurzen Überblick über angebotene Demos. Außerdem gab Mobsy bei Fragen Auskünfte über laufende Arbeiten am Lehrstuhl. Das System läuft ohne Eingriff von außen robust und fehlertolerant und zeigt die erfolgreiche Integration von Sprach– und Bildverarbeitung in einem Serviceroboter Szenario. Die Akzeptanz bei den Benutzer macht deutlich, dass natürliche Sprache und Dialog als Schnittstelle zum System sowie aktive Kamerasteuerung zur Gesichtsverfolgung wichtige Aspekte in einem solchen Anwendungsgebiet darstellen. Regelmäßige, automatische Rekalibrierung mittels visueller Information sowie Hinderniserkennung mittels Infrarotsensorik stellt den robusten Betrieb auch bei zahlreichen Besuchern im 9. Stock sicher.

## 2.1 Statistische Objekterkennung und Ansichtenplanung

Die Arbeit zur statistischen, erscheinungsbasierten Objekterkennung im Rahmen des von der DFG geförderten Graduiertenkollegs "Dreidimensionale Bildanalyse und -synthese" wurde fortgesetzt. Man hat sich auf den Erkennungsprozess in Szenen, in denen mehrere Objekte vorkommen, konzentriert.

Um Objekte in einer Szene zu erkennen, werden zunächst statistische Modelle für jede mögliche Objektklasse erstellt. Man nimmt Objekte in verschiedenen, bekannten Lagen auf. Dann wird ein Gitter über jedes Grauwerttrainingsbild gelegt, und an jedem Gitterpunkt ein zweidimensionaler Merkmalsvektor mit Hilfe der Wavelet-Multiskalen-Analyse bestimmt. Anschließend werden die Komponenten der Merkmalsvektoren statistisch modelliert, wobei die Normalverteilung angenommen wird. Da ein Objekt einen begrenzten Teil eines Bildes belegt, wird für jede Objektklasse ein so genanntes variables Objektfenster trainiert. Für jedes Trainingsbild bestimmt man die Menge der zum Objekt gehörenden Merkmalsvektoren, die dann in der Erkennungsphase in Betracht gezogen werden. Die übrigen Merkmalsvektoren fasst man zu einem Hintergrundmodell zusammen, das mit Annahme der Gleichverteilung trainiert wird. Beim Lernen der Objektmodelle kann man immer nur eine endliche Anzahl von Ansichten verwenden. Bei der Erkennung ist die Verteilung aller möglichen Objektlagen kontinuierlich. Um diese Kontinuität handhaben zu können, werden Approximationen mit trigonometrischen Basisfunktionen für die Merkmalsvektoren sowie für die Objektfenster durchgeführt.

In der Erkennungsphase wurden Szenen mit einem und mit mehreren Objekten im Bild untersucht, wobei die Arbeiten an Mehrobjektszenen im Vordergrund standen.

Bei Einobjektszenen wird immer nur nach exakt einem Objekt in einem Bild gesucht. Man benutzt in diesem Fall den so genannten Bayes-Klassifikator, indem man für jede mögliche Lagehypothese jeder möglichen Objektklasse eine Wahrscheinlichkeit ermittelt. Als endgültiges Klassifikations- und Lokalisationsergebnis gilt dann die Objektklasse mit der höchsten Wahrscheinlichkeit in der höchst wahrscheinlichen Lage.

Die Mehrobjektszenen bringen außer der Lokalisation und der Klassifikation von Objekten auch ein neues Problem mit sich. Die Anzahl der im Bild auftretenden Objekten aus der trainierten Stichprobe ist nicht bekannt und muss bestimmt werden. Eine globale Zuweisungsfunktion wird definiert, die die einzelnen Merkmalsvektoren im Bild den einzelnen Objekten zuweist. Gesucht wird bei Mehrobjektszenen seriell. Es wird nach dem ersten Objekt im Bild gesucht, die Zuweisungen durchgeführt, dann nach dem zweiten gesucht usw.; so lange bis ein Abbruchkriterium erfüllt ist. Die Suche nach mehreren Objekten lässt sich beschleunigen, indem man die Anzahl der möglichen Objektklassen im Verlauf der Suche reduziert, und die Lageschätzungen wiederverwendet.

Eine wesentliche Erweiterung, die im letzten Jahr durchgeführt wurde, ist die Zusammenführung des statistischen Klassifikators mit dem System für die Fusion von mehreren Ansichten, das auf dem Condensation Algorithmus basiert. Bei manchen Stichproben sind die Objektklassen aus bestimmten Ansichten nicht zu unterscheiden. Dank des neuen Algorithmus kann die Entscheidung über Klasse und Lage eines Objektes aufgrund einer Bildfolge getroffen werden. Die Wahrscheinlichkeitsverteilung über alle möglichen Klassen und Lagen im $n$-ten Bild in der Erkennungsbildfolge hängt von der Wahrscheinlichkeitsverteilung im $(n-1)$-ten Bild in der Folge ab. Je länger eine Bildfolge ist, desto höher ist die Wahrscheinlichkeit für die richtige Klasse und umso besser das Klassifikationsergebnis (siehe Abbildung 1).

Um die Probleme bei realen Anwendungen zu erkennen, wurden neue Experimente durchgeführt. Für die Erkennung wurden Testbilder mit heterogenem Hintergrund aufgenommen. Die Beleuchtungsbedingungen waren anders als im Training. Trotz der hohen Allgemeinheit des Problems konnte man befriedigende Ergebnisse erzielen. Bei Einobjektszenen mit heterogenem Hintergrund liegt die Klassifikationsrate bei 70% und die Lokalisationsrate bei 67.2%. Bei Szenen mit mehreren Objekten konnte die Anzahl der im Bild gefundenen Objekte in 94% der Fälle korrekt geschätzt werden. Die Erkennungsrate bei der Fusion von mehreren Ansichten steigt mit der Anzahl der fusionierten Bilder, was Abbildung 1 entnommen werden kann. Die Klassifikationsrate nach dem fünften Bild in der Erkennungsfolge liegt bei 72,5%, nach dem zehnten beträgt sie 77,5% und nach dem fünfzehnten 85%.

"Szenenerkennung mit Bayes- und Neuronalen Netzen" ist ein Projekt, welches als Teil des vom Graduiertenkolleg behandelten Themas "Dreidimensionale Bildanalyse und -synthese" durchgeführt und von der Deutschen Forschungs Gemeinschaft unterstützt wird. Die gegenwärtige Aufgabenstellung definiert sich innerhalb eines der fortgeschrittenen Gebiete der Bildverarbeitung und findet ihre Anwendungen dort, wo Herausforderungen an das Maschinensehen in komplexen Szenen und Arbeitsumgebungen gelöst werden müssen.

Die Szenenerkennung mit Hilfe von Bayes-Netzen basiert auf der Idee, eine aktive, wissensbasierte Suche auf Bildern durchzuführen, die sich von den konventionellen Algorithmen zur

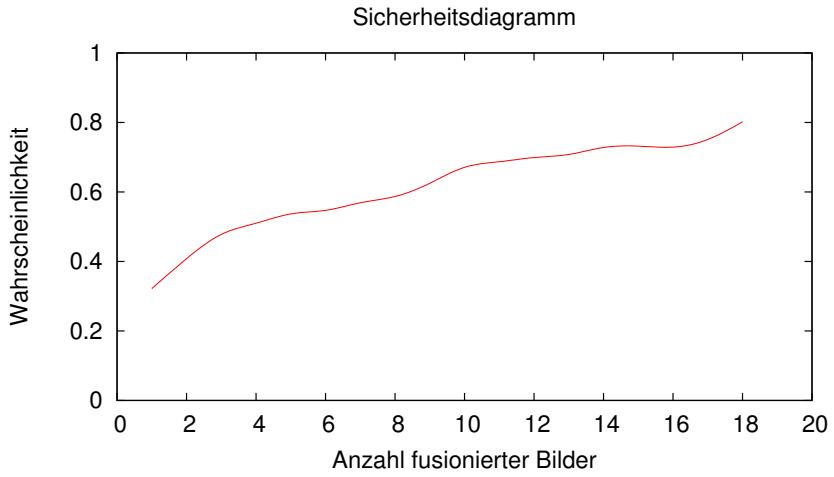


Abbildung 1: Wahrscheinlichkeit für die erwartete Objektklasse in Abhängigkeit von der Anzahl fusionierter Bilder. Je länger eine Bildfolge ist, desto höher ist die Wahrscheinlichkeit für die richtige Klasse.

visuellen Erkennung unterscheidet. Während der indirekten Bildsuche steht genau zu Beginn eines jeden Experimentes ein Stichprobensatz von Trainingsbilddaten verschiedener Klassen zur Verfügung, wobei die Art der zu suchenden Klasse unbekannt ist. Normalerweise wird dann eine rekursive Suche nach Objekten aller Klassen innerhalb eines Bildes durch den Einsatz eines herkömmlichen Objekterkennungssystems unternommen, was aber durch den Bayes-Ansatz, das Ziel der aktuellen Forschung, umgangen werden kann. Die Objektsuche in einem Bild durch Bayes-Netze kann auf eine spezielle Klasse, beziehungsweise auf einen Satz von Klassen, beschränkt werden, sobald die Beziehungen zwischen einzelnen Objekten genau definiert sind. Unsere anfänglichen Ergebnisse haben dabei bewiesen, dass, wenn die strukturellen Beziehungen zwischen den Einzelobjekten eines Bildes richtig festgelegt werden, die Szenensuche mit Bayes-Netzen durchaus effektiv ist, was auch die dargestellten Ergebnisse aussagen.

Allgemeine Büroszenen werden durch Bayes-Netze (BNs) modelliert, wobei die diskreten Knoten jeden der drei Werte $\{$True(T), Unknown(U), False(F)$\}$ annehmen können. Das Lernen der Bayes-Netz-Struktur wird daraufhin manuell durchgeführt, indem angenommen wird, dass die drei diskreten Knoten eine Maus $\Omega_{mos}$, eine Tastatur $\Omega_{keb}$ und eine Tasse $\Omega_{cup}$ repräsentieren, was in Abb. 2 dargestellt ist. Zum parametrischen Lernen der Bayes-Netze werden zwei Ansätze verfolgt, nämlich der (1) Statistische Ansatz (SA) und (2) die Geometrie der Ansicht (GV) zum Berechnen der a-priori Werte. Unter Verwendung der a-priori Werte beider Ansätze werden Verbundwahrscheinlichkeitsverteilungen berechnet und es wird der *J-Tree Inference Algorithm* verwendet, um die a-posteriori Wahrscheinlichkeiten zu ermitteln, mit denen ein bestimmtes Objekt in der Szene gefunden wurde, wenn Belege über die anderen beiden gegeben sind. Für Experi-

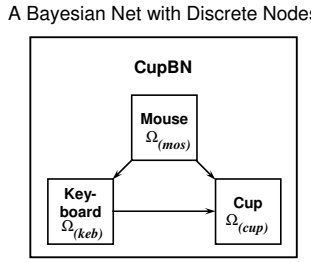


Abbildung 2: Bayes-Netz - Diskrete Knoten

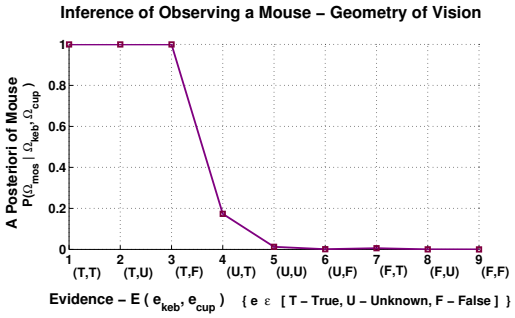


Abbildung 3: Inferenz bei $\Omega_{mos}$

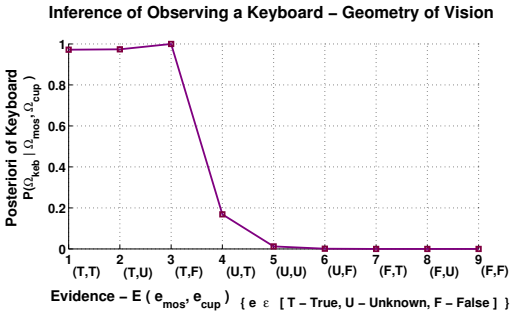


Abbildung 4: Inferenz bei $\Omega_{keb}$

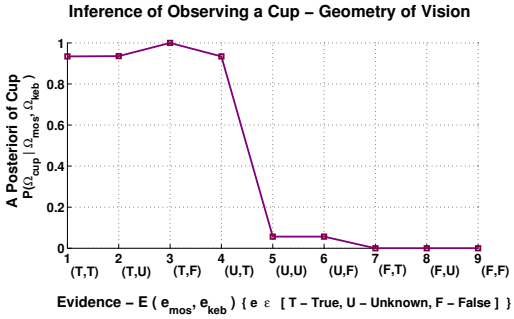


Abbildung 5: Inferenz bei $\Omega_{cup}$

mente, welche die BN-Software benutzen, wird durch die BN-Funktionen automatisch die Inferenz mit allen möglichen Kombinationen der Zustände der diskreten Knoten bestimmt. Für das Vorhandensein der drei Büroszenenobjekte Maus $\Omega_{mos}$, Tastatur $\Omega_{keb}$ und Tasse $\Omega_{cup}$ ergeben sich je nach Ansatz a-priori Werte von 0.85, 0.90 und 0.75 für GV und 0.91, 0.88 und 0.86 entsprechend für SA. Für den Fall, dass 'eins oder beide übrigen Objekte' mit Sicherheit vorhanden sind, resultiert für das Auffinden des dritten Objektes eine a-posteriori Wahrscheinlichkeit von über 0.50, falls der GV-Ansatz verfolgt wird. Die entsprechenden a-posteriori Werte aus SA dagegen liegen unterhalb von 0.25, während a-posteriori Wahrscheinlichkeiten, die mit Hilfe von GV geschätzt wurden (siehe Abb. 3, Abb. 4 und Abb. 5), nachweislich höhere Werte zu liefern versprechen als jene des Statistischen Ansatzes. Die schwachen Ergebnisse von SA liegen darin begründet, dass Abhängigkeiten nicht korrekt zu Grunde gelegt werden konnten - im Gegensatz zu GV. Derzeit ist es allerdings noch zu früh, die Klassifikationsraten der Bayes-Netze mit denen herkömmlicher Erkennungs-Software zu vergleichen.

Anhand der Graphen in Abb. 3, Abb. 4 und Abb. 5 kann man erkennen, dass die BN-Struktur aus Abb. 2 sinnvolle Wahrscheinlichkeiten für alle Knoten geliefert hat, was von der Logik her einleuchtend ist, da falsche oder unbekannte Zustandswerte geringere Wahrscheinlichkeiten liefern, während richtige Zustände natürlich höhere a-posteriori Wahrscheinlichkeiten bereitstellen.

Im kommenden Zeitraumn wird das Augenmerk darauf gerichtet sein, die konventionellen Algorithmen - speziell die Wavelet-Transformation und statistische Objekterkennungssysteme - mit dem Baustein der Bayes-Netze zur aktiven Suche zu verknüpfen. Bestärkt durch die anfänglichen Resultate werden die diskreten Knoten in nächster Zukunft durch kontinuierliche ersetzt werden.

Nachdem in den letzten Jahren im Teilprojekt B2 des Sonderforschungsbereichs 603 bereits

sehr effiziente Methoden zur ansichtenbasierten Objekterkennung, d.h. zur kombinierten Klassifikation und Lokalisation, untersucht und implementiert worden sind, konnte man sich zuletzt wieder vermehrt einem Kernpunkt der eigentlichen Aufgabenstellung, der Ansichtenplanung widmen, welche die zuverlässigen Ergebnisse der Objekterkennung benötigt.

Hierbei existieren aufgrund der vorhergehenden Forschung schon sehr ausgereifte Techniken zur Fusion von Kameraaufnahmen sowie zur optimalen Auswahl eben dieser durch Methoden des Reinforcement Learnings (RL).

Dennoch war es bisher problematisch, eine planmäßige Kamerabewegung in der realen Anwendung auszuführen, da hier bezüglich der Ziele des Teilprojektes kontinuierliche Parameter bereitgestellt werden müssen, was in der Praxis einer Bewegung der Kameras auf einer Kreisbahn mit uneingeschränkten Winkelwerten entspricht. Hinsichtlich der Lokalisation eines Objektes konnte diese Problematik in der Vergangenheit mit Hilfe von kontinuierlichen Parametrisierungen statistischer Ansätze gelöst werden, doch auch für das Reinforcement Learning selber ist es nicht ausreichend, nur einen Satz von diskreten Zuständen und Aktionen zuzulassen. Insbesondere ist dies offensichtlich, da es – neben der allgemeinen Forderung noch Kontinuität – immer eine Rauschüberlagerung in der Bewegung der kameratragenden Plattform MOBSY geben wird, die es per se verhindert, sich ausschließlich auf diskreten Punkten zu justieren.

Im vergangenen Jahr wurde deshalb zur Approximation der RL-Zustands-Aktionswert-Funktion das Verfahren der erinnerungsbasierten Funktionsapproximation so modifiziert, dass es die o.g. Problematik beseitigt und unter Beibehaltung der größtmöglichen Allgemeinheit an die Anwendung angepasst werden konnte. Auf Basis aller durch das Training bekannter Zustands-Aktionswerte wurden dazu beliebige, unbekannte Werte approximiert, die daraufhin ihrerseits wiederum zur Stabilisierung der Wissensbasis herangezogen werden konnten. Zu diesem Zweck war es notwendig, sowohl eine deterministische Transformationsfunktion, welche ausschließlich von der jeweiligen Anwendung abhängt (hier also im Wesentlichen die Transformation durch die Winkeländerung), und eine Abstandsfunktion von RL-Zuständen heranzuziehen. Letztere musste dabei so gestaltet werden, dass sie gerade den Abstand von Verteilungsdichten zueinander berechnen kann, da nur solche im Teilprojekt zum Einsatz kommen werden.

Weiterhin war es notwendig, eine Kernfunktion zur Gewichtung der Zustände der Wissensbasis einzuführen, welche eine gleichmäßige Berechnung über den Zustandsraum gewährleistet. Das dabei auftretende Problem der Bestimmung eines sinnvollen Gewichtungsparameters $D_\Phi$ konnte hervorragend durch die Integration eines bereits früher entwickelten Verfahrens zur automatischen Parameterberechnung für allgemeine Anwendungen behoben werden. Hierzu gilt es, über die Kurvenlänge der Zustands-Aktionswerte-Funktion den Einfluss dieses Parameters so zu regulieren, dass möglichst gut ausgeprägte, aber nicht zu viele lokale Maxima auftreten. Neben der somit erzielten Bereitstellung der geforderten Funktionalität gelang es durch geschickte Ansätze in der Verarbeitung bekannter Zustandswerte, eine Laufzeitoptimierung und Effizienzerhöhung zu erreichen, ohne dabei Einbußen in der Aktionsplanung der realen Anwendung hinnehmen zu müssen.

Ein zusätzlicher wichtiger Bearbeitungspunkt war die Ansichtenplanung mittels statistischer Klassifikatoren, die ausschließlich die Entropie als Gütemaß für die Belohnungen des Reinforcement Learnings vorsehen. Zwar werden damit die einsetzbaren Klassifikatoren stark eingeschränkt, aber die Ergebnisse und Anwendungsmöglichkeiten der statistischen Klassifikatoren haben sich

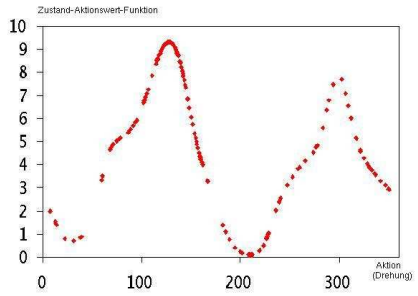


Abbildung 6: ohne Kostenfunktion

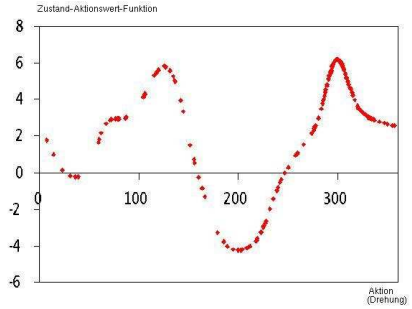


Abbildung 7: mit Kostenfunktion

als sehr effizient herausgestellt. Auch hier wurde der Einsatz der bereits ausgiebig untersuchten Partikelmengen diskutiert und ebenfalls durch geeignete Verfahren für die Verwendung im kontinuierlichen Wertebereich verfügbar gemacht.

Zudem wurden schon einige initiale bzw. fortgeschrittene Ideen bezüglich der Zoom- und Greifplanung akquiriert. In erster Linie wurden die Berechnungen der Greifplanung mit einem Kostenfaktor versehen, der es ermöglichen soll, diejenige von mehreren optimalen Greifpositionen einzunehmen, welche den geringsten Aufwand erfordert, also im Falle des Teilprojektes den minimalen Bewegungsweg. Hierzu verdeutlichen Abb. 6 und Abb. 7 die Änderung der optimalen Aktion unter Verwendung der Kostenfunktion. Somit wurden bereits die Grundlagen für die Verfolgung und Greifplanung bewegter Objekte durch die mobile Plattform MOBSY gelegt, was wiederum das erklärte Ziel der näheren Zukunft sein wird.

## 2.2 Bildanalyse für autonome Systeme

Zentraler Forschungsgegenstand des Teilprojektes B2 im Sonderforschungsbereich 603 ist die Ermittlung der optimalen Kameraparameter für die Aufgaben der Objektverfolgung und der Objekterkennung. Im Rahmen der Objektverfolgung, die als Zustandsschätzproblem eines dynamischen Systems aufgefasst wird, ist das verwendete Optimierungskriterium die bedingte Entropie zwischen Zustand und Beobachtung. Als Zustandsschätzer wurden bislang erweiterte Kalman-Filter und Partikelfilter verwendet. Daneben wurde jetzt die Klasse der so genannten *Unscented Kalman Filter* evaluiert, welche die Wahrscheinlichkeitsdichteverteilung während des Zustandsschätzproblems als eine deterministische Punktemenge darstellen. Die Leistung dieser Filter war vergleichbar mit der des erweiterten Kalman Filters.

Die Objektverfolgung mittels Lichtfeldern, entstanden in Kooperation mit Teilprojekt B6 und C2 des Sonderforschungsbereiches 603, wurde mit dem regionenbasierten Objektverfolger für eine Kamera kombiniert. Lichtfelder gehören zu den ansichtenbasierten Bildgenerierungsverfahren. Sie sind in der Lage, aus trainierten Ansichten neue Ansichten aus beliebigen Blickwinkeln erzeugen zu können, und das in fotorealistischer Qualität. Die regionenbasierte Objektverfolgung liefert dabei die 2-D Position des Objektes im Kamerabild. Die Verfolgung mittels Lichtfeldern benutzt diese Information als Ausgangspunkt zur 3-D Positions- und Lageschätzung mittels einer geeignet gewählten Gütefunktion zum Bildvergleich. Dadurch stehen sowohl die schnelle Reak-

tionszeit der Template-basierten Methode als auch die allgemeine Lageschätzung des Lichtfeld-Verfolgers zur Verfügung.

Als ein Anwendungsgebiet für die Evaluation der im Teilprojekt B2 entwickelten Ansätze und Verfahren dient ein Szenario aus dem Bereich der Dienstleistungsrobotik. Mittels der mobilen Plattform MOBSY wird dabei ein Objekt in einer Büroumgebung erkannt, angefahren und gegriffen. Dies bedarf des koordinierten Zusammenspiels von Objekterkennung, Objektverfolgung, Ansichten- und Greifplanung in einem gemeinsamen System. Die Objektverfolgung behält das Objekt bei optimal eingestellten Brennweiten während der Anfahrt der mobilen Plattform im Blickfeld der Kameras. Mit Hilfe der Ansichtenplanung wird sowohl eine robuste Erkennung des zu greifenden Gegenstandes sowie eine exakte Ermittlung der Objektlage ermöglicht. Dabei bestimmt eine Greifplanung die optimale Greifposition bezogen auf das Objekt und steuert daraufhin die Plattform und den Greifer entsprechend.

Dabei muss jedoch beachtet werden, dass die Objektverfolgung und die Objekterkennung gleichzeitig auf die Kameras zugreifen, jedoch unterschiedliche Anforderungen haben. So kann es für die Objektverfolgung vorteilhafter sein, eine hohe Brennweite zu wählen, während die Objekterkennung eine kleine Brennweite bevorzugt, zum Beispiel um das ganze Objekt im Blickfeld zu haben, oder um das Objekt möglichst in der Größe zu sehen, die auch beim Training des Klassifikators verwendet wurde. Dabei ist es erstrebenswert, dass die Steuerung der Brennweiten durch ein einziges, zentrales System geschieht, welches die Einzelbedürfnisse von Objektverfolgung und Objekterkennung geeignet fusioniert. Dies wurde dadurch erreicht, daß das entropiebasierte Optimierungskriterium der Objektverfolgung, welches die optimale Brennweite bestimmt, um eine Gewichtsfunktion erweitert wurde. Diese Gewichtsfunktion gewichtet die Brennweiten stärker, die die Objekterkennung vorzieht, sofern eine Klassifikation durchgeführt wird. Ist dies nicht der Fall, so ist sie konstant und hat auf die Optimierung keinen Einfluss. Abbildung 8 zeigt diesen Kompromiss in Aktion.

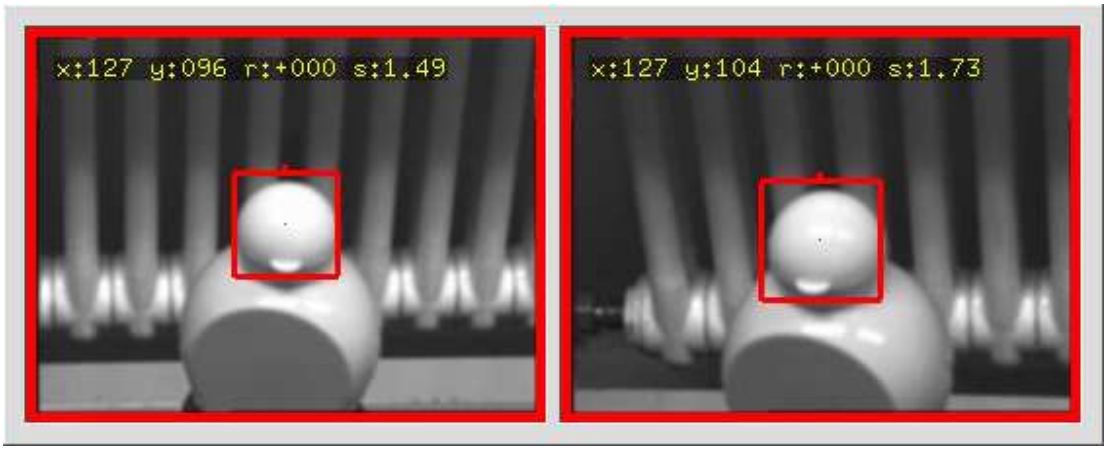


Abbildung 8: Fusion der Anforderung der Objekterkennung und der Objektverfolgung an die Zoomplanung. Nur die linke Kamera wird zur Erkennung eingesetzt, daher überwiegt die Anforderung des Klassifikators an eine kleine Brennweite. Im Bild der rechten Kamera hat der Klassifikator keinen Einfluss, die hohe Brennweite wird allein von der Objektverfolgung bestimmt.

Weiterhin wurden erste Arbeiten an der Aufgabe durchgeführt, die beiden internen Kameras auf dem Roboter durch zwei externe Kameras zu ergänzen. Diese externen Kameras überblicken generell die gesammte Szene des Greifvorgangs, also den Roboter und das zu greifende Objekt. Sie sollen die Objektverfolgung unterstützen, wenn die beiden internen Kameras das Objekt nicht gut erblicken können. Dies kann der Fall sein, wenn das Objekt neben oder hinter dem Roboter steht, oder wenn sich ein verdeckendes Hindernis zwischen dem Roboter und dem Objekt befindet. Letzteres tritt dann auf, wenn das Objekt auf dem Boden und damit zu niedrig ist, um von den Kameras zu jedem Zeitpunkt des Greifvorgangs gesehen zu werden. In diesem Fall ist der Roboter selbst das verdeckende Hindernis. Dieses Thema wurde im Rahmen einer Masterarbeit untersucht. Dabei bestimmen die externen Kameras rein sichtbasiert die Position des Zielobjektes und die Position und Orientierung des Roboters im Raum. Aus diesen Informationen kann die relative Position des Objektes zum Roboter errechnet werden, und ein Greifen wird auch ohne direkten Sichtkontakt möglich.

## 2.3 Bildbasierte Modellierung und erweiterte Realität

Das Teilprojekt C2 des Sonderforschungsbereichs 603 bearbeitet zusammen mit dem Lehrstuhl für Graphische Datenverarbeitung (LGDV) das Thema der „Analyse, Codierung und Verarbeitung von Lichtfeldern zur Gewinnung realistischer Modelldaten". Die in diesem Teilprojekt zentrale Datenstruktur des Lichtfelds erlaubt es, durch sog. bildbasierte Modellierung beliebige Ansichten einer realen Szene zu generieren, deren Aussehen durch eine Sammlung an Bilddaten bekannt ist. Die benötigten Informationen über Parameter und Positionen der verwendeten Kamera werden über Verfahren der „Struktur aus Bewegung" (Structure from Motion) direkt aus den Bilddaten ermittelt, bei denen es sich meistens um einen Bildstrom von einer handgeführten Kamera handelt. Das Teilprojekt C2 wird bereits seit 1998 von der Deutschen Forschungsgemeinschaft (DFG) gefördert und wurde 2003 ein zweites Mal, bis Ende 2006, verlängert.

Ein wichtiges Werkzeug zur Berechnung von Struktur aus Bewegung sind Faktorisierungsmethoden. Sie ermöglichen es, aus Punktkorrespondenzen in einer Reihe von Bildern gleichzeitig Kamerapositionen und 3-D-Struktur der Szene zu berechnen, und bilden so einen Bestandteil zur Rekonstruktion von Lichtfeldern. Hier wurden eine Reihe von Verbesserungsmöglichkeiten untersucht, so etwa ein Algorithmus, der auch mit Merkmalen umgehen kann, die nicht in allen Bildern sichtbar sind. Dieser wurde von einem einfachen, schwach-perspektivischen Modell auf ein perspektivisches erweitert.

Weitere Verbesserungen allgemeiner Faktorisierungsmethoden gehen davon aus, dass die Eingangsdaten der Punktkorrespondenzen verrauscht oder fehlerhaft sind, und modellieren daher die Unsicherheit dieser Punkte. In einer Studienarbeit wurden zwei dieser Verfahren untersucht und verglichen und können nun auch für die Lichtfeldberechnung verwendet werden. Die Betreuung erfolgte gemeinsam mit dem Projekt VAMPIRE (s. Abschnitt 2.4), da die Ergebnisse hier ebenfalls von Interesse sind.

Die bereits 2002 begonnenen Untersuchungen von dynamischen Lichtfeldern wurden 2003 unter einem anderen Blickwinkel fortgeführt. Die ersten dynamischen Lichtfelder wurden durch das Registrieren mehrerer statischer Lichtfelder generiert. Im Gegensatz dazu können sie nun aus einem einzigen Bildstrom erzeugt werden, der kontinuiertliche Bewegung enthält. Eine Ein-

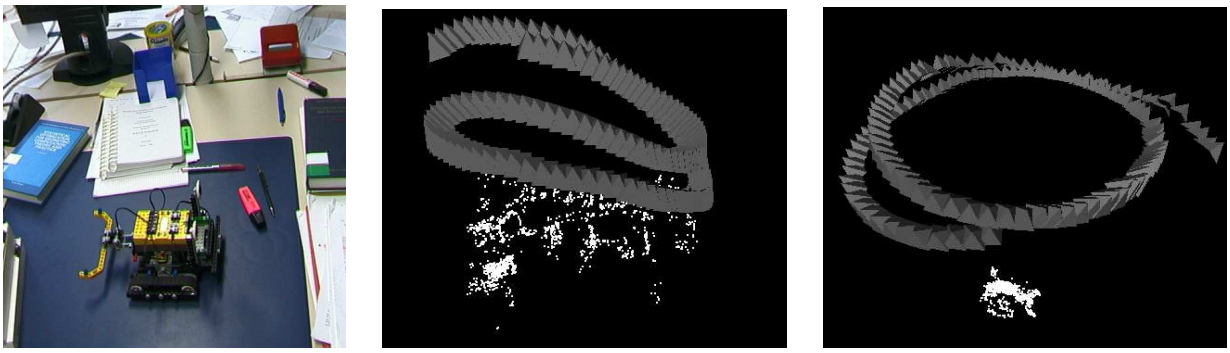

Abbildung 9: Beispielsequenz einer rotierenden Spielzeugraupe (links) mit getrennter Rekonstruktion von Hintergrund (mitte) und Objekt (rechts)

schränkung besteht darin, dass sich nur ein starres, bewegtes Objekt in der Szene befindet. Durch eine Trennung von Hintergrund und Objekt vor der Kalibrierung und Geometrierekonstruktion kann eine Analyse der Objektbewegung durchgeführt werden. Dadurch können Bilder mit gleichen oder ähnlichen Objektpositionen identifiziert werden, was eine Rückführung auf die ursprünglichen, schrittweise statischen Lichtfelder ermöglicht. Abbildung 9 zeigt ein Beispiel für eine dynamische Szene, deren Hintergrund und Vordergrund getrennt rekonstruiert wurden. In der Darstellung des Vordergrunds (rechts) ist deutlich das Objekt als Punktwolke zu erkennen sowie die Bewegung der Kamera relativ zu dem Objekt, das sich um seine Achse drehte.

Im Teilprojekt B6, „Rechnergestützte Endoskopie des Bauchraums“, des Sonderforschungsbereichs 603 (SFB 603) wurden in Zusammenarbeit mit der Chirurgischen Universitätsklinik die grundlegenden Arbeiten zur Einführung einer Rechnerunterstützung in der minimal-invasiven Chirurgie abgeschlossen.

Das System zur Bildverbesserung [34] und Erzeugung einer 3-D-Visualisierung des Operationsgebietes mittels Lichtfeldern wurde während einer realen Operation eingesetzt (siehe Abbildung 10). Dabei wurden drei Lichtfelder erzeugt und die Bilder während der Operation in Echtzeit verbessert. Das System bietet dem Chirurgen folgende Bildverbesserungsmöglichkeiten: Farbnormierung, Entzerrung und zeitliche Filterung. Die Methoden können nach Belieben kombiniert werden. Eine Evaluation der Bildverbesserungsmethoden mit 14 Ärzten und 120 Bildvergleichen pro Arzt [22, 34] zeigte sehr deutlich, dass Entzerrung und zeitliche Filterung subjektiv als Verbesserung der Bildqualität empfunden werden (Rangsummentest mit Signifikanzniveau $\alpha = 0,01$). Bei der Farbnormierung ergab sich ein großer Unterschied zwischen erfahrenen und unerfahrenen Ärzten. Während die unerfahrenen Ärzte die farbnormierten Bilder bevorzugten, wurden diese von den erfahrenen Ärzten als schlechter beurteilt. Vermutlich hängt dies damit zusammen, dass erfahrene Ärzte andere Farben als ungewohnt und daher schlechter beurteilen.

Die Transformation von der Endoskophalterung zur Endoskopspitze (Hand-Auge-Transformation) wurde bisher mit einem semi-automatischen Verfahren berechnet. Es ist auch möglich diese Transformation über ein automatisches Hand-Auge-Kalibrierverfahren zu bestimmen [28].

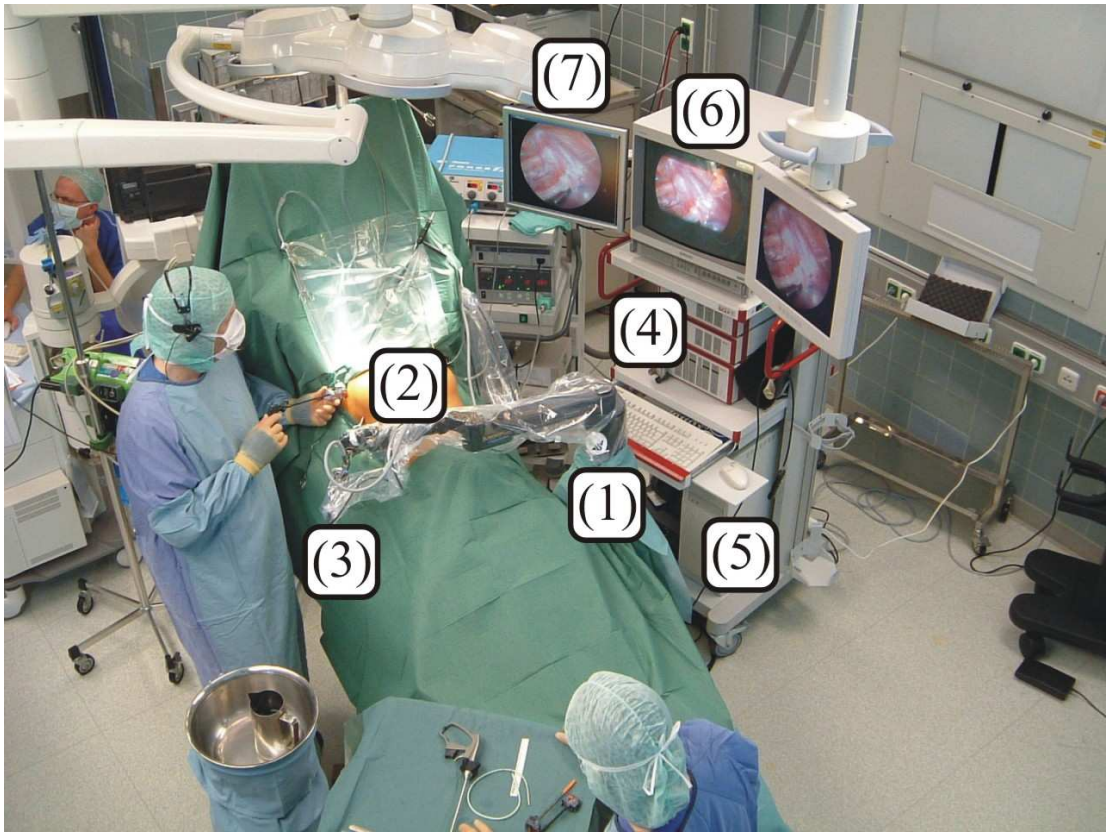


Abbildung 10: Das System zur computergestützten Endoskopie im OP der Chirurgischen Universitätsklinik. (1) Roboterarm AESOP 3000 zur sprachgesteuerten Bewegung und Lagebestimmung des Endoskops, (2) Patient, (3) Kamerakopf und Endoskop, (4) Lichtquelle (5) PC, (6) Video-Endoskopiesystem (Originalbild), (7) Zweiter Monitor (verarbeitetes Bild bzw. 3-D-Visualisierung).

Auf Grund der Ungenauigkeit des Roboterarms muss hierbei eine Datenselektion durchgeführt werden. Die resultierende Genauigkeit ist vergleichbar mit der semi-automatischen Methode. Beide Methoden basieren auf einer Kamerakalibrierung.

Es wurde untersucht, welche Möglichkeiten am besten geeignet sind, um mit Hilfe des Roboterarms Szenengeometrie zu berechnen. Die Szenengeometrie besteht aus 3-D-Punkten der Oberfläche. Es hat sich herausgestellt, dass die Triangulation eines Punktes aus allen Ansichten der Szene, in der der Punkt sichtbar ist, das beste Ergebnis liefert. Hierbei kann eine Ausreißerelimination dadurch erreicht werden, dass lediglich Bildpaare zur Triangulation verwendet werden und der endgültige 3-D-Punkt als Median der berechneten Punkte definiert wird. Ist Szenengeometrie gegeben, so kann die Registrierung der 3-D-Lichtfeld-Visualisierung mit CT-Daten durchgeführt werden, indem zunächst Dreiecksnetze der Oberflächen erzeugt werden und diese dann anhand korrespondierender Punkte grobregistriert und anschließend mit einem Iterative-Closest-Point Algorithmus feinregistriert werden. Der Vorteil für den Arzt liegt in der zusätzli-

Abbildung 11: Der SCORBOT ER VII

chen Information. Bisher wurde ein Datensatz einer Gallenblase registriert. Für die tatsächliche Anwendung ist eine Registrierung mit segmentierten Gefäßdaten notwendig. Diese Problematik wird 2004 angegangen.

Bei Verfahren zur bildbasierten Modellierung und zur erweiterten Realität ist stets eine Rekonstruktion der Szenengeometrie nötig. Es gibt dabei eine Vielzahl von Algorithmen, die für unterschiedliche Voraussetzungen unterschiedlich gute Ergebnisse liefern. Zur Zeit werden Experimente durchgeführt, welche Voraussetzungen in Kombination mit welchem Algorithmus zu einem möglichst guten, d. h. genauen, Ergebnis führen.

Die Voraussetzungen bei einer Rekonstruktion aus einer Bildfolge sind Pfad der Kamera und Brennweite. Um den Pfad möglichst gut bestimmen zu können, wird die Kamera dabei von einem Roboterarm geführt. Hierzu wird der SCORBOT ER VII der Firma Intelitek verwendet. Der Roboter ist in Abbildung 11 dargestellt. Er besitzt 5 Winkelachsen und eine Längsachse. Diese Werte sind über die serielle Schnittstelle auslesbar. Die Software für die Berechnung der Orientierung und Position wurde dabei selbst entwickelt.

## 2.4 Das Projekt VAMPIRE

Im Rahmen des EU-Projekts VAMPIRE (Visual Active Memory Processes and Interactive Retrieval) wird seit Juni 2002 Forschung im Bereich der automatischen Analyse von Videosequenzen durchgeführt. Im Vordergrund stehen dabei die Erkennung von Objekten und Bewegungsabläufen, sowie das Lernen neuer Objekt- und Bewegungsmodelle. Neben dem Lehrstuhl für Mustererkennung sind auch die Universität Bielefeld, die Technische Universität Graz und die University of Surrey am Projekt beteiligt.

Ein Schwerpunkt der Arbeit des Lehrstuhls für Mustererkennung innerhalb des Projekts VAMPIRE liegt im Bereich 3-D Rekonstruktion und bildbasierte Objektmodelle. Dieser Forschungsbereich wird im Rahmen des SFB 603 von einer weiteren Gruppe am Lehrstuhl untersucht (siehe Abschnitt 2.3), so dass auf vorhandenen Ergebnissen aufgebaut werden konnte. Der Einsatz einer

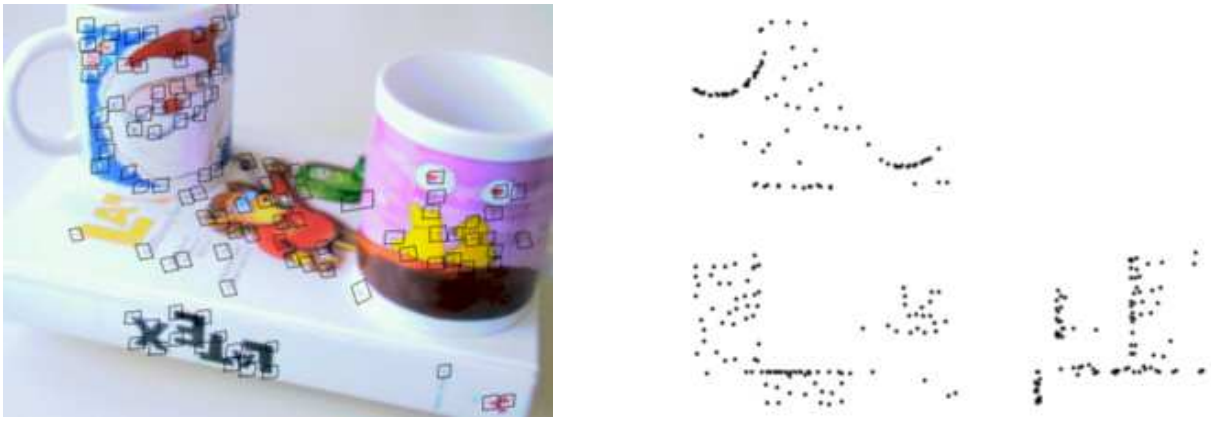

Abbildung 12: Verfolgte Punktmerkmale (links) und drei Ansichten der 3-D Rekonstruktion (rechts) einer Beispielsequenz.

Augmented-Reality-Ausrüstung für die Interaktion des Benutzers mit dem System macht allerdings oftmals die Anpassung, Beschleunigung und Verbesserung der vorhandenen Algorithmen notwendig.

Ein grundlegender Verarbeitungsschritt bei der 3-D Rekonstruktion aus Videosequenzen ist die Verfolgung von Punktmerkmalen. Durch die Anpassung an eine hochoptimierte Bildverarbeitungsbibliothek und die Implementierung einer verbesserten Merkmalsauswahl mit einer hierarchischen Datenstruktur konnte der existierende Punktverfolger auf die bis zu 30fache Geschwindigkeit beschleunigt werden, was einen Echtzeiteinsatz des Verfahrens zulässt. Weiterhin wurde die affine Bewegungsschätzung des Punktverfolgers verbessert, und mit Hilfe eines linearen Beleuchtungsmodells die Robustheit gegenüber Beleuchtungsschwankungen erhöht. Die letztgenannte Erweiterung ermöglicht es, die Translationsparameter von der affinen Schätzung zu übernehmen, welche das aktuelle Merkmal mit dem ersten Auftreten des Merkmals vergleicht. Mit dieser Methode wird ein langsames Abweichen der Merkmalsposition von der tatsächlichen Position, wie es bei einer Translationsschätzung von Bild zu Bild auftreten kann, wirkungsvoll verhindert.

Bei der 3-D Rekonstruktion werden aus den Koordinaten der verfolgten Punktmerkmale die Kameraparameter und die 3-D Geometrie der aufgenommenen Szene berechnet. Auch auf diesem Gebiet wurden erste Untersuchungen durchgeführt. Es hat sich dabei gezeigt, dass die Robustheit der Rekonstruktionverfahren für kalibrierte Kameras deutlich höher ist als für unkalibrierte Kameras, bei denen zusätzlich die intrinsischen Kameraparameter geschätzt werden müssen. Als Ausgangspunkt für weitere Untersuchungen wurde deshalb die Faktorisierungsmethode nach Christy-Horaud implementiert, welche für kalibrierte Kameras mit Hilfe eines affinen Kameramodells eine 3-D Rekonstruktion für ein perspektivisches Kameramodell berechnet. Dieses Verfahren diente auch als Grundlage einer Studienarbeit über robuste Faktorisierungverfahren, die Ausreißer und Schätzfehler besser tolerieren als herkömmliche Verfahren. Bild 12 zeigt das letzte Bild einer Videosequenz, sowie drei Beispielansichten der rekonstruierten 3-D Struktur der aufgenommenen Szene.

Neben der 3-D Rekonstruktion ist die Objektverfolgung ein weiterer Schwerpunkt im Projekt

VAMPIRE, der vom Lehrstuhl für Mustererkennung untersucht wird. Hierbei muss generell zwischen datengetriebenen und modellbasierten Verfahren unterschieden werden. Datengetriebene Verfahren besitzen die Eigenschaft, dass ohne a priori Wissen (beispielsweise Geometrieinformationen) über das zu verfolgende Objekt, eine Verfolgung durchgeführt werden kann. Dies hat den Vorteil, dass sofort nach der Detektion einer Bewegung in einer Szene eine Verfolgung eines unbekannten Objektes stattfinden kann. Nachteilig ist allerdings, dass bei einer Drehung des Objekts entlang einer Achse, die parallel zur Bildebene liegt, das Objekt verloren wird, weil kein Wissen vorhanden ist, wie das Objekt von einer anderen Ansicht aussieht. Diesen Nachteil kompensieren modellbasierte Verfahren, die während eines Trainings signifikante Informationen über das Objekt sammeln. Dennoch sind datengetriebene Verfahren sehr wichtig, da nur diese verwendet werden können, so lange unbekannt ist, welches Objekt sich in der Szene bewegt. Nachdem das Objekt nach einem Klassifikationsschritt erkannt wurde (siehe Abschnitt 2.1), kann von einem datengetriebenen auf ein modellbasiertes Verfahren umgeschaltet werden.

Ein datengetriebenes Verfahren ist der Hyperebenen Ansatz von F. Jurie, der am Lehrstuhl implementiert wurde. Unter der Annahme, dass die zu verfolgenden Objekte eine planare Oberfläche besitzen, lassen sich viele Bewegungsarten in der Bildebene parametrisieren, wie zum Beispiel Translation, Rotation, Skalierung und perspektivische Verzerrung. Ein typisches Problem, das man generell in der Bildverarbeitung antrifft, ist dass durch Veränderung der Beleuchtung, beispielsweise durch Änderung des Einstrahlwinkels der Beleuchtungsquelle (illustriert in Abbildung 13) oder auch Schatteneffekte, sich die Erscheinung des Objekts derartig ändert, dass eine Verfolgung fehlschlägt. Unter Annahme der Planarität des zu verfolgenden Objektes konnte gezeigt werden, dass durch ein lineares Beleuchtungsmodell Änderungen der Beleuchtung kompensiert werden können. Dieses Beleuchtungsmodell wird durch zwei Parameter festgelegt, die eine Änderung der Beleuchtungsstärke und des Kontrastes widerspiegeln, was auf zwei Arten geschehen kann. Zum einen durch eine Schätzung des minimalen quadratischen Fehlers zwischen einem Referenzbild und dem Bild des Objektes, das aktuell verfolgt wird, und zum anderen durch eine Normierung anhand der Verteilung der Intensitäten. In Experimenten konnte bestätigt werden, dass das zweite Verfahren die besten Ergebnisse liefert, was in [19] veröffentlicht wurde.

Ein weiteres typisches Problem, das in einer realen Szene häufig auftritt, sind Verdeckungen des zu verfolgenden Objektes, beispielsweise durch andere Objekte. Um diesem Problem zu begegnen wird versucht, Punkte, die verdeckt werden, zu erkennen und von der Bewegungsschätzung auszuschließen. Ein solches Verfahren, das am Lehrstuhl untersucht wurde, ist die X84 Rückweisungsregel, was ein klassisches statistisches Mittel zur Detektion von Ausreißern ist.

Da die Objektverfolgung im Projekt VAMPIRE in Echtzeit stattfinden soll, empfiehlt es sich, nicht alle Punkte eines Objektes zu verwenden, sondern sich auf eine Teilmenge von 100 bis 200 Punkten zu beschränken. In der ursprünglichen Implementation des Hyperebenen-Verfahrens wurden diese Punkte zufällig aus der Menge der Objektpunkte ausgewählt. Da Objektpunkte in schlecht texturierten Bereichen sich als ungeeignet erwiesen, wurden verschiedene Kriterien untersucht, um die Punktauswahl zu verbessern. In Experimenten konnte gezeigt werden, dass sich durch die Auswahl der Punkte in Bereichen starker Varianz und hoher Gradienten die Bewegungsschätzung verbessern lässt.

Im Jahr 2004 wird der Schwerpunkt von der datengetriebenen Objektverfolgung auf modellbasierte Verfahren verlagert. Ein modellbasierter Hyperebenenansatz scheint für dieses Vorgehen

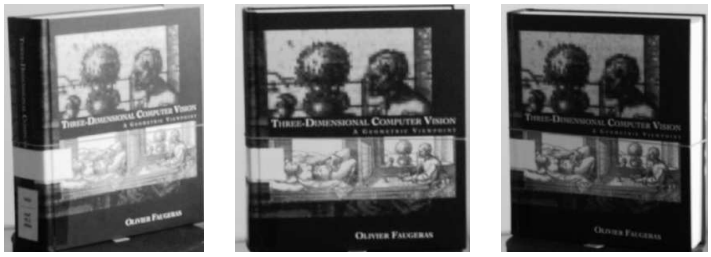


Abbildung 13: Typisches Beispiel für die Veränderung der Erscheinung eines Objektes bei unterschiedlichen Blickrichtungen. Man erkennt, dass auf Grund der Betrachtungsrichtung das Buch im linken Bild deutlich heller erscheint als im mittleren oder rechten.

sehr geeignet. Desweiteren sollen auch Objektmodelle, die aus Lichtfelder generiert wurden, untersucht werden. Wichtige Vorarbeiten wurden bereits im Projekt SFB 603 TP C2 (Siehe Abschnitt 2.3) geliefert.

## 2.5 Die virtuelle Hochschule Bayern

Im Jahr 2003 wurde die Entwicklung einer virtuellen Vorlesung mit dem Titel „Rechnersehen mit Anwendungen in der Augmented Reality sowie beim bildbasierten Rendering“ für die virtuelle Hochschule Bayern (vhb – http://www.vhb.org) weitestgehend abgeschlossen. Projektbeginn war Juni 2002. Das Projekt wurde vom Freistaat Bayern mit Mitteln aus der High-Tech-Offensive Bayern gefördert.

Fachlich soll der Kurs eine Einführung in die Themen 3D-Rekonstruktion, Augmented Reality und Lichtfelder geben. Dabei handelt es sich um aktive Forschungsschwerpunkte des Lehrstuhls für Mustererkennung (vgl. Abschnitte 2.3 und 2.4).

Das Erlernen der Inhalte soll ausschließlich über das Internet möglich sein. Zusätzliche Präsenzveranstaltungen sollen somit nicht nötig sein.

Um den Stoff möglichst auf unterschiedlichen Arten zu vermitteln wurden, auf der Grundlage des didaktischen Konzepts, verschiedene Komponenten entwickelt.

Die beiden wichtigsten Komponenten für die Vermittlung des theoretischen Wissens sind Skript und Folien. Diese beiden Teile enthalten inhaltlich im Prinzip das Gleiche, sind jedoch von der Bedienung und Verwendung her unterschiedlich. Die Folien dienen zum Lesen auf den Bildschirm. Dort sind die zahlreichen Links auch gut nutzbar und durchaus hilfreich zum Nachschlagen, z. B. auf vorherige oder nachfolgende Kapitel. Die Seitengröße ist auch auf das Bildschirmformat angepasst, so dass möglichst wenig geblättert werden muss. Die Folien enthalten auch eine Option zum Vorlesen. So können zwei Sinneskanäle gleichzeitig genutzt werden und der Behaltensgrad erhöht sich. Im Gegensatz dazu ist das Skript zum Ausdrucken gedacht und befindet sich deshalb im DIN A4 Format. Das Skript enthält den Inhalt als Fließtext, was die Notwendigkeit zum Ausdrucken unterstreicht, da von vielen ein Lesen von langen Fließtexten am Bildschirm als nicht komfortabel angesehen wird.

Ein weiterer Baustein sind die Verständnisfragen. Diese werden am Ende eines jeden Ab-

schnitts gestellt und sollen nochmal das Nachdenken über die einzelnen Aspekte anregen. Die Fragen stellen eine wichtige Komponente für die Selbstkontrolle der Studenten dar. Dabei wurden meist offene Fragen gestellt, auf die Verwendung von Multiple Choice Fragen wurde weitestgehend verzichtet. Multiple Choice Fragen sind in ihrer Anwendung zu weit eingeschränkt, als dass man damit komplexe Sachverhalte abfragen könnte. Der Lernende muss sich zunächst selbst eine Lösung überlegen und kann diese dann mit der Musterlösung vergleichen.

Als weitere Komponente wurde eine Experimentierumgebung eingerichtet. Dort können zu verschiedenen Abschnitten der Inhalte Experimente durchgeführt werden. Es wurden prinzipiell zwei Arten von Experimenten realisiert. Einige Versuche brauchen relativ lange Berechnungszeiten und wurden deswegen so umgesetzt, dass die Daten hochgeladen und dort verarbeitet werden. Das Ergebnis kann dann nach spätestens einigen Minuten Bearbeitungszeit wieder heruntergeladen werden. Lange Bearbeitungszeiten benötigen z. B. Videos auf denen Punktverfolgung oder eine 3D-Rekonstruktion durchgeführt wird. Die zweite Art von Experimenten wurde durch Javaapplets realisiert. Dies ist dann möglich, wenn die Berechnung in Echtzeit durchgeführt werdenkann.

Die Kommunikationsumgebung zählt ebenfalls zu den Komponenten. Nur durch eine umfassende Kommunikation kann eine optimale Betreuung gewährleistet werden. Dabei gilt es zu vermeiden, dass sich der Lernende alleine gelassen fühlt, denn sonst kann die Motivation schnell nachlassen, was sich negativ auf die Lernleistungen auswirkt. Für die Kommunikation wurden zwei Systeme installiert: Zwei E-Mailverteiler, einer nur für Studenten, einer für Studenten mit Betreuer und ein Webforum. Das Webforum kann wahlweise als moderiertes oder nicht moderiertes Forum betrieben werden.

Der Kurs wurde in zwei Teile aufgespalten und in jeweils zwei Testbetrieben von Studenten getestet. Das Feedback der Studenten war durchweg positiv, mit der Ausnahme das die Kommunikationsumgebung nicht so gute Noten erhielt. Dies liegt aber vor allem daran, dass die Zahl der Studenten in den Testläufen relativ gering war, die Studenten sich persönlich kannten und deshalb lieber persönlich Kontakt aufnahmen, als über die bereitgestellten Mittel zu kommunizieren.

Die Entwicklung des Kurses wird Anfang 2004 abgeschlossen werden. Nach erfolgreicher Begutachtung durch die virtuelle Hochschule Bayern wird die virtuelle Vorlesung dort in das Kursprogramm aufgenommen.

## 2.6 Statistische Modellierung von Daten

Aufgrund einer steigenden Komplexität von Produktionsprozessen wird es immer schwieriger die Einflußfaktoren auf die Qualität eines Produktes zu erkennen bzw. die Produktionsprozesse zu regeln. Das Teilprojekt C1 des Sonderforschungsbereichs 396 trägt durch Entwicklung von stochastischen Modellen, genau genommen Bayesnetzen, zur Lösung dieser Probleme bei. Ein Bayesnetz wird dabei durch seine Struktur, d.h. die Verbindung der einzelnen Knoten, und durch die Verteilungsfunktion beschrieben, die jedem Knoten zugeordnet ist.

In den zurückliegenden Jahren wurde die Verwendung von Bayesnetzen als selbstadaptive Regler für lineare Systeme erforscht. Ein weiteres Forschungsziel war die Modellierung nichtlinearer Systeme. Die Kernidee ist es, das nichtlineare System durch mehrere Taylorreihen zu

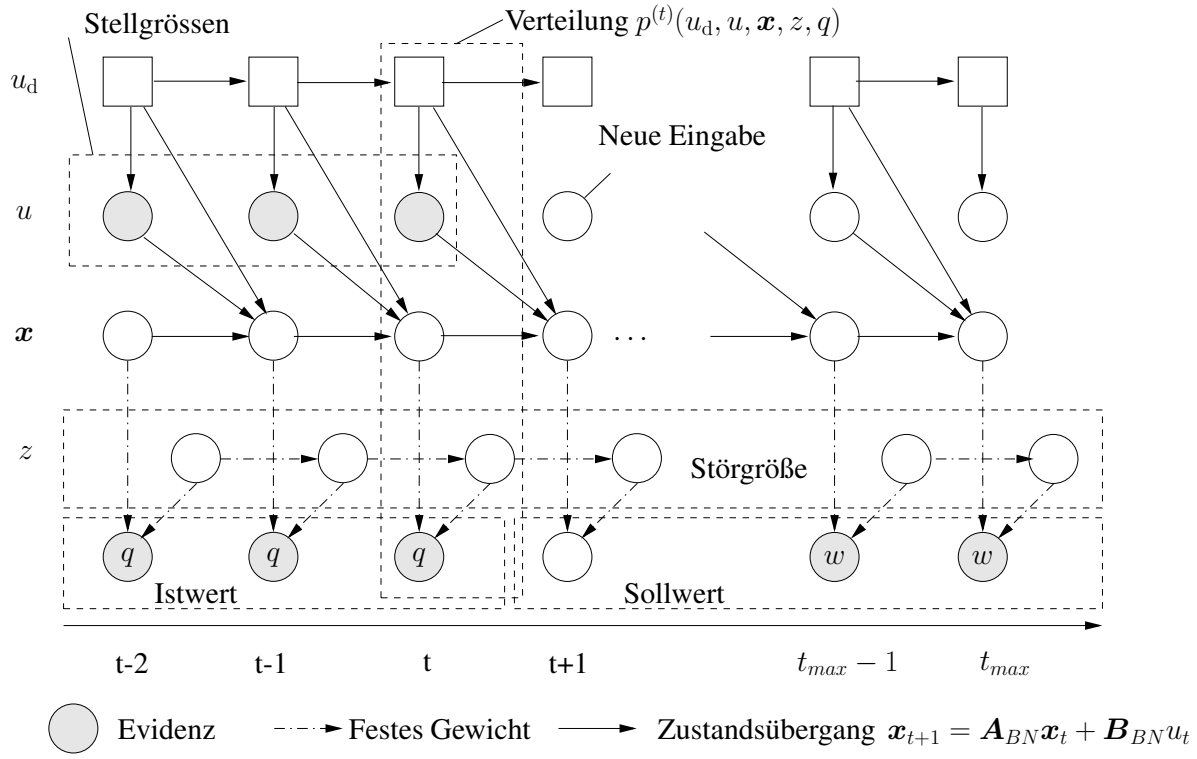


Abbildung 14: Bayesnetz zur Modellierung eines nichtlinearen, dynamischen Systems

approximieren. Ein diskreter Knoten dient als Schalter zwischen den verschiedenen Taylorreihen. Hierbei ist natürlich zu beachten, dass dies kein abruptes Umschalten ist, sondern es werden die Gewichte in einer Mischungsverteilung verschoben.

In dem zurückliegenden Jahr wurden die beiden Ideen kombiniert, so dass auch nichtlineare, dynamische Systeme durch ein hybrides, dynamisches Bayesnetz modelliert werden können. Es hat sich als günstig erwiesen, die Modellierung der Nichtlinearität mit dem Zustandsraummodell zu kombinieren. Als Resultat entsteht das Modell in Abbildung 14. Dieses Modell dient dazu eine eingangsseitige Nichtlinearität, z.B. eine Sättigung, zu modellieren. Der diskrete Knoten $u_{\mathrm{d}}$, der im Gegensatz zu kontinuierlichen Knoten als Quadrat dargestellt ist, schaltet dabei zwischen verschieden Arbeitspunkten um.

Dieses Modell wird mit dem EM-Algorithmus trainiert und nach dem Training als Regler eingesetzt. Dabei werden in den ersten Zeitscheiben ehemalige Ein- und Ausgaben als Evidenz eingegeben. Dadurch ist eine Schätzung der Störgröße $z$ und des Zustandes $\mathbf{x}$ möglich. In den Zeitscheiben, die die Zukunft repräsentieren wird der Sollwert $w$ als Evidenz in die Ausgabeknoten eingetragen. Knoten, für die eine Evidenz vorliegt sind in Abbildung 14 schattiert gezeichnet.

Anschließend wird durch Marginalisierung die Verteilungsfunktion für die Knoten $u_{\mathrm{t+1}}, \cdots,$ $u_{\mathrm{t+i}}$ berechnet. Damit kann auf ein geeignetes Eingabesignal geschlossen werden. Das resultierende Modell ist sehr gut zu Regelungszwecken geeignet. Ein neuer Sollwert wird nach kurzer Zeit erreicht, auch eine Störung wird schnell ausgeregelt. Es ergibt sich eine kleine bleibende Regelabweichung von ca. 0,5% des Sollwertes.

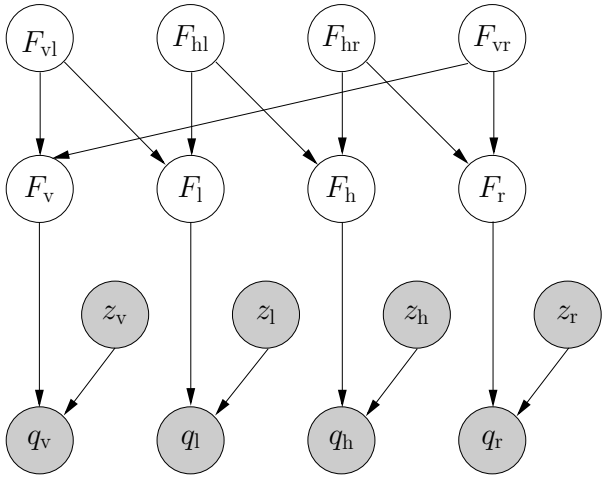


Abbildung 15: Modellierung der IHU-Presse

Ein Problem ist die Inferenzdauer, die bei einer exakten Inferenz exponentiell mit der Anzahl der Zeitscheiben ansteigt. Zur Verkürzung der Inferenzdauer wird momentan ein approximativer Algorithmus erforscht, der eine Inferenz in linearer Zeitdauer ermöglicht.

Die Kernidee einer Regelung, die auf einem Bayesnetz basiert, wurde zur Regelung der Presse beim Innenhochdruckumformen (IHU) herangezogen. Bei der Modellierung der Presse wurde erkannt, dass ein linearer Zusammenhang zwischen den Eingabekräften $F_{\mathrm{vl}}, \cdots, F_{\mathrm{vr}}$ und den Ausgabekräften $F_{\mathrm{v}}, \cdots, F_{\mathrm{r}}$ besteht. Dieser Zusammenhang wird durch die ersten beiden Schichten in Abbildung 15 repräsentiert. Die dritte Schicht repräsentiert die Störgröße, die sich aus der Differenz zwischen Soll- und Istwert ergibt. Die unterste Schicht repräsentiert die beobachtete Ausgabe und dient zur Eingabe des Sollwertes.

Wie schon bei der Regelung nichtlinearer dynamischer Systeme wird nach der Eingabe der Evidenz per Marginalisierung auf mögliche Eingabewerte geschlossen. Als Resultat ergibt sich eine gleichmäßigere Kraftverteilung und damit auch ein gleichmäßigerer Blecheinzug.

Im nächsten Jahr werden sich die Aktivitäten auf das Erforschen eines approximativen Algorithmus konzentrieren. Weitere Themen sind das Erlernen der Struktur und das Anwenden von Bayesnetzen auf die Regelung des Spritzgußprozesses.

# 3 Sprachverstehen

Leitung: E. Nöth
(J. Adelhardt, A. Batliner, C. Frank, C. Hacker, M. Levit, R. Shi, S. Steidl, G. Stemmer, V. Zeißler)

Die inhaltlichen Schwerpunkte der Forschungsaktivitäten zur Sprachverarbeitung bilden das maschinelle Erkennen und Verstehen gesprochener Äußerungen sowie Fragestellungen des multimodalen Mensch-Maschine-Dialogs. Die Arbeiten im Berichtsjahr können zwei anwendungsorientierten Projekten zugeordnet werden: Erkennung von spontaner Sprache und Emotionen im Rahmen des *PF-STAR*–Projekts sowie die Entwicklung des multimodalen Dialogsystems *SmartKom*.

Im von der Europäischen Union geförderten Projekt *PF-STAR* sollen Grundlagen für zukünftige Forschungsaktivitäten im Bereich der Mensch-Technik-Interaktion gelegt werden. Dazu werden vor allem drei Bereiche untersucht: Übersetzung gesprochener Sprache, automatische Erkennung und Ausdruck von Emotionen sowie die Entwicklung von Algorithmen zur Verarbeitung von Kindersprache. Am *PF-STAR*–Projekt sind insgesamt sieben Arbeitsgruppen aus mehreren Universitäten und Forschungseinrichtungen beteiligt. Der Lehrstuhl bearbeitet die Bereiche Emotionserkennung und Erkennung von Kindersprache. Weitere Forschungstätigkeiten, die über den Rahmen des *PF-STAR*–Projekts hinaus gehen, beschäftigen sich mit der Verbesserung der automatischen Erkennung spontaner Sprache auch von nicht-nativen Sprechern.

In dem vom BMBF geförderten Projekt *SmartKom* werden Konzepte für neuartige Formen der Mensch-Technik-Interaktion durch die Entwicklung eines Demonstrationssystems bereits praktisch erprobt. Diese Konzepte sollen die bestehenden Hemmschwellen von Computerlaien bei der Nutzung der Informationstechnologie abbauen und so einen Beitrag zur Benutzerfreundlichkeit und Benutzerzentrierung der Technik in der Wissensgesellschaft liefern. Das Ziel von *SmartKom* ist die Erforschung und Entwicklung einer selbsterklärenden, benutzeradaptiven Schnittstelle für die Interaktion von Mensch und Technik im Dialog. Am *SmartKom*–Projekt sind insgesamt 12 Arbeitsgruppen aus mehreren Universitäten, Großforschungseinrichtungen und Firmen beteiligt. Der Lehrstuhl bearbeitet die Bereiche Prosodie–, Mimik– und Gestik–Interpretation.

## 3.1 Das PF-STAR-Projekt

Das EU-Projekt PF-STAR (**P**reparing **f**uture multi**s**ensorial in**te**r**a**ction **r**esearch) begann am 1. 10. 2002 und ist auf zwei Jahre angelegt. Es werden Grundlagen und Referenzsysteme für zukünftige Forschungsprojekte untersucht und entwickelt. Schwerpunkte der Untersuchungen sind die automatische Übersetzung von Sprache, Erkennung und Synthese von Emotionen in der Sprache, ferner Synthese von Emotionen in Gesichtern und Erkennung von Kindersprache. Die Partner sind innerhalb Deutschlands die Universität Aachen und die Universität Karlsruhe, sowie europaweit die Universitäten Birmingham, Padua und Stockholm, sowie das ITC-Irst in Trient, Italien. Der Lehrstuhl für Mustererkennung der Universität Erlangen ist in den Teilgebieten

- Detektion von emotionaler Sprache und
- Erkennung von gesprochener Sprache von Kindern

beteiligt und leitet zudem das erstere. Die Erkennung von Emotionen findet beispielsweise Anwendung in automatischen Telefondialogsystemen. Man interessiert sich dafür, ob der Anrufer verärgert ist, um rechtzeitig eingreifen zu können und den Benutzer an einen menschlichen Agenten weiterzuleiten. Besondere Schwierigkeit bereitet dabei die sprecherunabhängige Klassifikation von nicht-geschauspielerten natürlichen Emotionen.

Bei der automatischen Erkennung von Sprache bereitet bisher beispielsweise die Erkennung von Kindern besondere Schwierigkeiten. Zu Beginn des Projektes wurden deshalb emotionale Daten von Kindern gesammelt, um eine Brücke zwischen beiden Teilprojekten zu schlagen.

In der ersten Hälfte des Projekts konzentrierte sich die Arbeit also auf das Sammeln und Bearbeiten von Sprachkorpora, auf die Adaption und Berechnung von geeigneten akustischen Merkmalen sowie auf erste Auswertungen und Klassifikationen.

**Sprachkorpora**

An zwei Erlanger Schulen, dem Ohm-Gymnasium und der Montessori-Schule, wurden Kinder im Alter von 10-12 Jahren aufgenommen. Sie hatten zum einen deutsche und altersgemäße englische Texte zu lesen, zum anderen nahmen sie an einem Experiment teil, mit dem emotionale Kindersprachdaten erhoben wurde. Dabei mussten sie dem Roboterhund AIBO der Firma Sony verschiedene Aufgaben stellen, wie etwa zu einem von mehreren Näpfen mit Futter zu gehen, oder ihn durch einen Parcours zu steuern, der auf einem Teppich aufgemalt ist. Da der AIBO nicht perfekt funktionierte und manchmal recht unwillig reagierte, konnten unterschiedliche emotionale Reaktionen ausgelöst werden, von lobend („gut gemacht, Aibolein“) bis verärgert („Steh auf, du blöde Blechbüchse!“). Die Äußerungen der insgesamt 51 Kinder (21 männlich, 30 weiblich), die an diesem AIBO-Experiment teilnahmen, wurden mit einem hochqualitativen Funkmikrophon auf einem DAT-Recorder aufgenommen und später auf den Rechner übertragen. Die Lesedaten wurden direkt am Rechner aufgenommen. Eine besondere Zusammenarbeit fand hier auch mit der Universität Birmingham statt. Dort wurden für britische Kinder Aufnahmen mit gleichem Versuchsaufbau durchgeführt.

Alle Sprachdaten wurden orthographisch transliteriert; die Daten des AIBO-Experiments werden zusätzlich noch nach bestimmten Gesichtspunkten gelabelt, d.h., bei jedem Wort wird notiert, ob es prosodisch auffällig ist und ob es von einen neutralen Benutzerzustand abweicht, z.B. ob es hilflos, verärgert, lobend, oder gelangweilt klingt.

Abb. 16 zeigt den Parcours. Eine weitere Aufgabe war die so genannte Objektlokalisierung (Abb. 17). AIBO sollte dabei zu einem vorgegebenen Futternapf gesteuert werden. Unter allen Umständen musste AIBO davon abgehalten werden, zu einem Napf mit vergifteten Futter zu gehen.

Im Berichtsjahr wurden weitere Sprachkorpora bearbeitet. Das SympaFly-Korpus beinhaltet Dialoge verschiedener Benutzer eines Telefon-Flugbuchungssystems der Firma Sympalog. Im Korpus ist emotionale Sprache von Erwachsenen enthalten. Der YOUTH-Korpus der Carnegie Mellon University beinhaltet amerikanische Kindersprache.

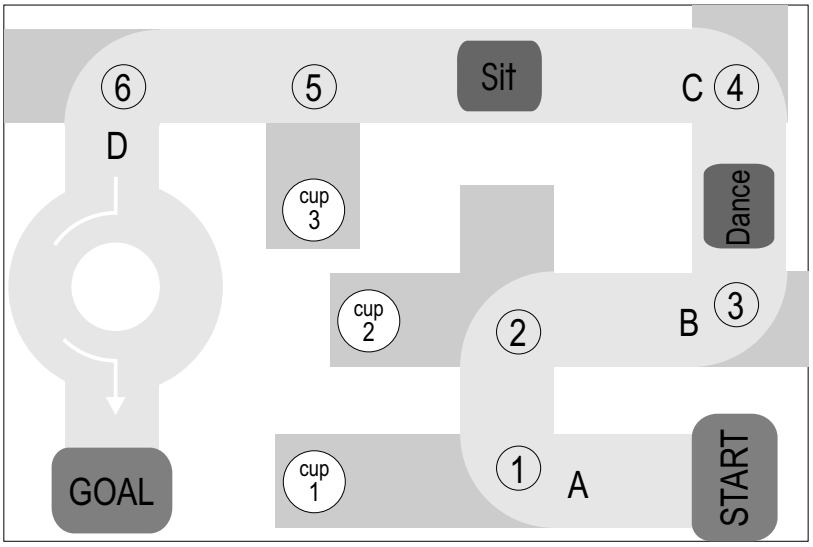


Abbildung 16: Parcours durch den die Schüler und Schülerinnen den Roboterhund AIBO steuern sollten

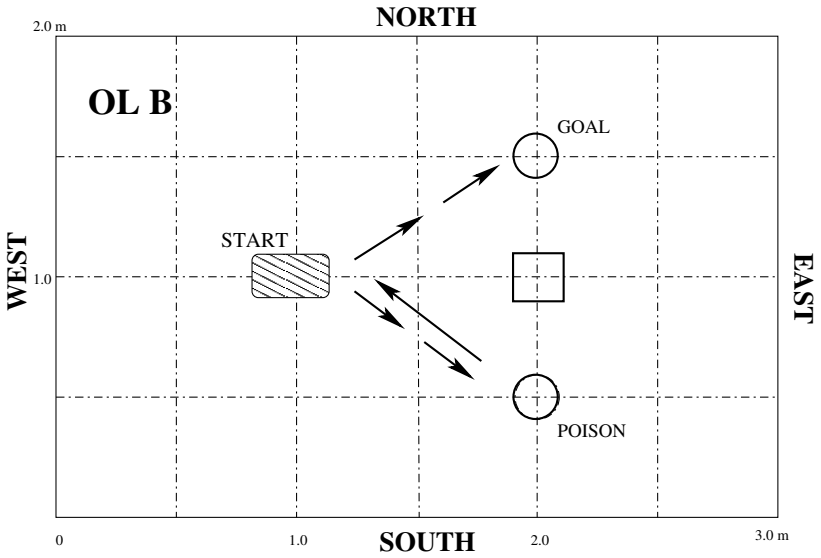


Abbildung 17: Objektlokalisation: Die Schüler und Schülerinnen hatten die Aufgabe, AIBO zu einem vorgegebenen Objekt zu steuern

### Akustische und linguistische Merkmale

Am Institut wird seit längerem ein umfangreicher prosodischer Merkmalvektor verwendet, der Tonhöhe, Dauerphänomene und Energie modelliert und der sich bei unterschiedlichen Fragestellungen bewährt hat, u.a. auch bei der automatischen Klassifikation von emotionalen Benutzerzuständen in den Projekten Verbmobil und SmartKom. Zusätzlich zu diesen prosodischen Merkmalen wurden nun verschiedene spektrale Merkmale berechnet. Diese basieren auf dem

Harmonicity-to-Noise-Ratio (HNR), den Frequenzen von vier Formanten und auf der Energie in vier Bändern, die um die vier Formanten zentriert sind.

Potentielle Anwendungen für die Erkennung emotionaler Zustände bei Kindern liegen im Bereich Edutainment/Entertainment. Automatische Dialogsysteme stellen eine andere für PF-STAR relevante Anwendung dar. Das SympaFly-Korpus wurde unter unterschiedlichen Gesichtspunkten annotiert: prosodische und konversationelle Auffälligkeiten, (emotionaler) Benutzerzustand, sowie Dialogschritt- und Dialogerfolgsrate. Für den Benutzerzustand wurden mit leave-one-out folgende klassenweise gemittelten Erkennungsraten erzielt: 29,7 % für die 9 annotierten Klassen, 49,8 % für 4 Klassen (neutral, emphatisch, hilflos, markiert), und 72,3 % für 2 Klassen (neutral/positiv vs. negativ). Auch Merkmalextraktionsverfahren unseres Partners ITC-irst wurden auf dem SympaFly-Korpus getestet und evaluiert.

In einer Studienarbeit konnte gezeigt werden, dass mit linguistischen Merkmalen wie der Abfolge von Wortklassen oder beispielsweise dem Vorhandensein des Wortes „ja“ in der nächsten Benutzeräußerung bestimmen lässt, ob ein einzelner Dialogschritt erfolgreich ist. Im SympaFly-Korpus wurde mit diesen Merkmalen für den Dialogschritterfolg eine klassenweise gemittelte Erkennungsrate von 82,5 % erzielt. A-posteriori lassen sich auch Aussagen über den Erfolg eines gesamten Dialogs treffen, um beispielsweise in der Entwicklungsphase eines Dialogsystems fehlgeschlagene Dialoge selektieren zu können. Der Dialogerfolg konnte zu 85,4 % richtig klassifiziert werden.

**Kindersprache und nicht-native Sprache**

Bei der Erkennung von Kindersprache ist in der europaweiten Zusammenarbeit in PF-STAR insbesondere auch die Erkennung von nicht-nativer Kindersprache interessant. Alle Partner haben Kinder aufgenommen, die englische Texte lesen. Fernziel ist die Entwicklung von automatischen Aussprache-Tutoren für den Fremdsprachenerwerb. Schwierigkeit hierbei ist die Erkennung von Kindersprache, die Erkennung von nicht-nativer Sprache und die Bewertung der Aussprache.

Ursache dafür, dass Kindersprache schlechter erkannt wird, sind u.a. Verschiebungen im Frequenzbereich. Abbildung 18 zeigt die Verschiebung der Formanten der Laute /e:/ und /O/ von Kindersprache aus den Aibo-Aufnahmen im Vergleich zur Erwachsenensprache im Verbmobil-Korpus. Für die Erkennung von Kindersprache wurden verschiedene akustische Normalisierungsverfahren untersucht und weiterentwickelt. Die bisher besten Ergebnisse konnten mit einer nicht-linearen Verzerrung der Frequenz-Achse erzielt werden. Dabei wird die Frequenz-Verzerrung durch eine Spline-Kurve modelliert, die iterativ verbessert wird. Vergleichende Messungen der Sprechgeschwindigkeit von Kindern und Erwachsenen zeigten, dass hier z. T. erhebliche Unterschiede zwischen den beiden Gruppen vorhanden sind. Eine erste Untersuchung, bei der die Sprechgeschwindigkeit in den Sprachsignalen der Kinder mit dem PSOLA-Verfahren skaliert wurde, zeigten, dass eine Beschleunigung der Kindersprache die Erkennung deutlich verbessert kann. Dies ist in Abbildung 19 dargestellt (Aufnahmen von gelesenen Texten).

Bereits im letzten Jahr wurde in einer Diplomarbeit ein Verfahren entwickelt, bei dem die Erkennung nicht-nativer Sprache durch die Interpolation von akustischen Modellen verbessert wird. Dazu wurden bisher sowohl Trainingsdaten der Muttersprache eines nicht-nativen Sprechers als auch Daten der Zielsprache benötigt. Im Berichtsjahr wurde das Verfahren so weiterentwickelt,

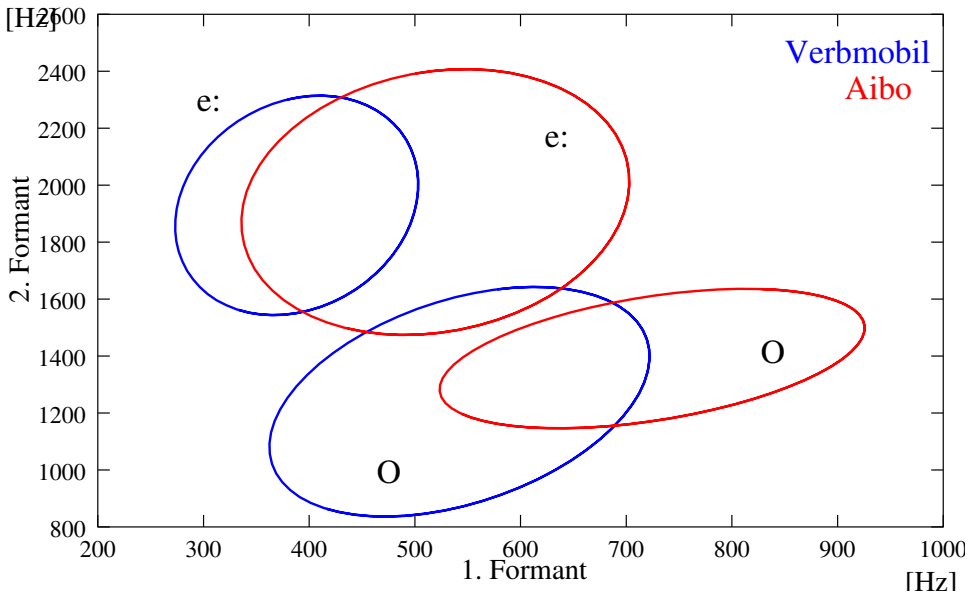


Abbildung 18: Verschiebung des 1. und 2. Formanten bei Kindersprache (Aibo) im Vergleich zu Erwachsenensprache (Verbmobil)

dass keine Daten der Muttersprache eines nicht-nativen Sprechers mehr benötigt werden. So ist die robuste Erkennung nicht-nativer Sprache auch dann möglich, wenn der Akzent des Sprechers dem System nicht bekannt ist.

## 3.2 Das HUMAINE-Projekt

Am 1. 1. 2004 läuft das auf vier Jahre angelegte EU-Projekt HUMAINE „Human-Machine Interaction Network on Emotion“ im 6. Rahmenprogramm der EU unter der Projektnummer 507422 an. In diesem Projekt sollen die Grundlagen gelegt werden für „emotions-orientierte Systeme“, die menschliche emotionale und emotionsähnliche Zustände und Prozesse erkennen, modellieren und beeinflussen. Der Lehrstuhl für Mustererkennung ist in diesem Projekt an den Arbeitspaketen 4 (Signale), 5 (Korpora) und 6 (Interaktion) beteiligt.

## 3.3 Sonstige Arbeiten

Die Forschung am Spracherkennungssystem des Lehrstuhls konzentrierte sich im Berichtsjahr auf die Verbesserung der Erkennung für Sprechergruppen, die meist nur schlecht verstanden werden: nicht-native Sprecher und Kinder.

Im Rahmen einer Studienarbeit wurde untersucht, inwiefern verschiedene Akzente automatisch unterschieden werden können. Als Daten standen Aufnahmen von gelesenem, nicht-nativem Englisch zur Verfügung; die Sprecher stammten aus Deutschland und Italien. Es zeigte sich, dass bereits mit einem sehr einfachen Mischverteilungsklassifikator Erkennungsraten von ca.

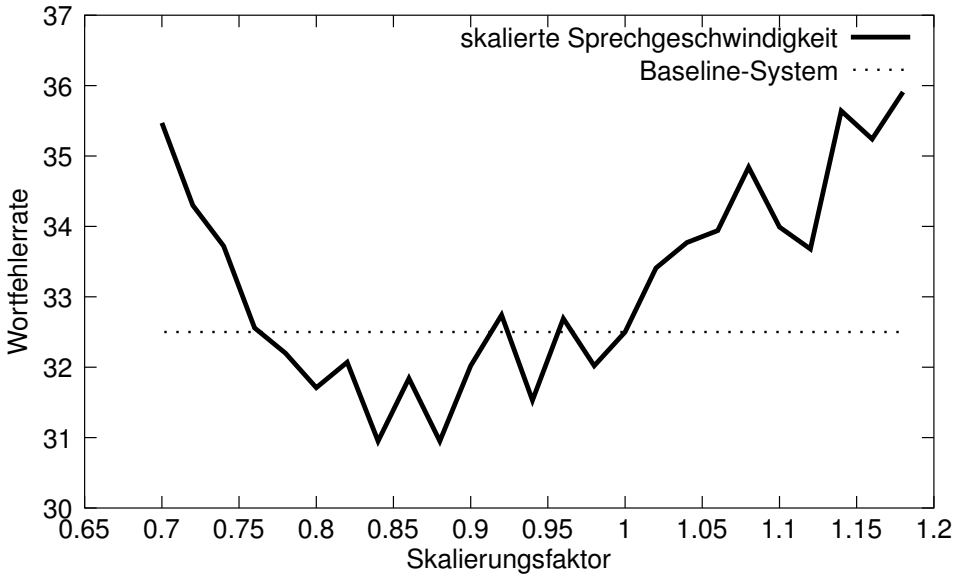


Abbildung 19: Wortfehlerrate eines mit Erwachsenen-Sprache trainierten Erkenners auf Kindersprache für unterschiedliche Skalierungungen der Sprechgeschwindigkeit. Eine Beschleunigung der Kindersprache (Skalierungsfaktor $< 1$) verbessert die Erkennung.

70 % erreicht werden. Die am besten geeigneten Merkmale für dieses Problem scheinen die Mel-Cepstrum Koeffizienten und ihre Ableitungen zu sein.

Eine weitere Studienarbeit beschäftigte sich mit dem Problem, Merkmale für die Spracherkennung zu finden, die besser zwischen den Sprachlauten diskriminieren als die verbreiteten Standardmerkmale, die Mel-Cepstrum Koeffizienten. Es wurde ein datengetriebener Ansatz verfolgt, der auf einer nicht-linearen Transformation, der Kernel-PCA oder der Kernel-LDA basiert. Die Schwierigkeit bestand darin, den benötigten Speicher- und Rechenaufwand der Kernel-Methoden möglichst stark zu reduzieren, ohne die Ergebnisse negativ zu beeinträchtigen. Für zahlreiche aus der Literatur bekannte Stichproben und Klassifikationsprobleme, wie z. B. die Erkennung handgeschriebener Ziffern, wurden hervorragende Resultate erzielt. Leider konnten diese nicht auf das Spracherkennungsproblem übertragen werden; die Erkennungsraten wurden nur geringfügig verbessert.

## 3.4 Das SmartKom-Projekt

Ziel des Projektes *SmartKom* (*http://www.smartkom.org/*), das als eines von vier Leitprojekten durch das BMB+F ins Leben gerufen wurde, war die Entwicklung eines multimodalen multimedialen Dialogsystems. Der Lehrstuhl für Mustererkennung bearbeitete die vier Teilprojekte *Mimikerkennung*, *Erkennung prosodischer Phänomene*, *Emotionserkennung in der Stimme* und *Gestenanalyse & Stifteingabe*. Die Entwicklung von allen entsprechenden Modulen wurde bis zum Ende des Projekts im September des vergangenen Jahres zum Abschluss gebracht. Die Module sind in das *SmartKom*-Demonstrationssystem integriert worden.

Das Prosodiemodul, in dem sowohl die Erkennung prosodischer Phänomene als auch die Emotionserkennung integriert sind, wurde im vergangenen Jahr fertiggestellt. Seine Architektur ist in Abbildung 20 skizziert. Um die maximale Erkennungsleistung zu erzielen, wurden alle Klassifikatoren mit den aktuellen Daten trainiert. Das Training der prosodischen Klassifikatoren erfolgte auf der *SmartKom WoZ*-Stichprobe. Für die Klassifikation der emotionalen Benutzerzustände wurde die auf dem Lehrstuhl aufgenommene *MMEG*-Stichprobe benutzt. Eine Optimierung des Klassifikators hinsichtlich des zum Training ausgewählten Datenmaterials, der benutzten Merkmale und der Klassifikationsparameter brachte eine weitere Verbesserung der Ergebnisse. Ein Teil dieser Experimente, die sich mit der Auswertung von unterschiedlichen Merkmalsgruppen befassen, ist ausfürlich in [1] beschrieben.

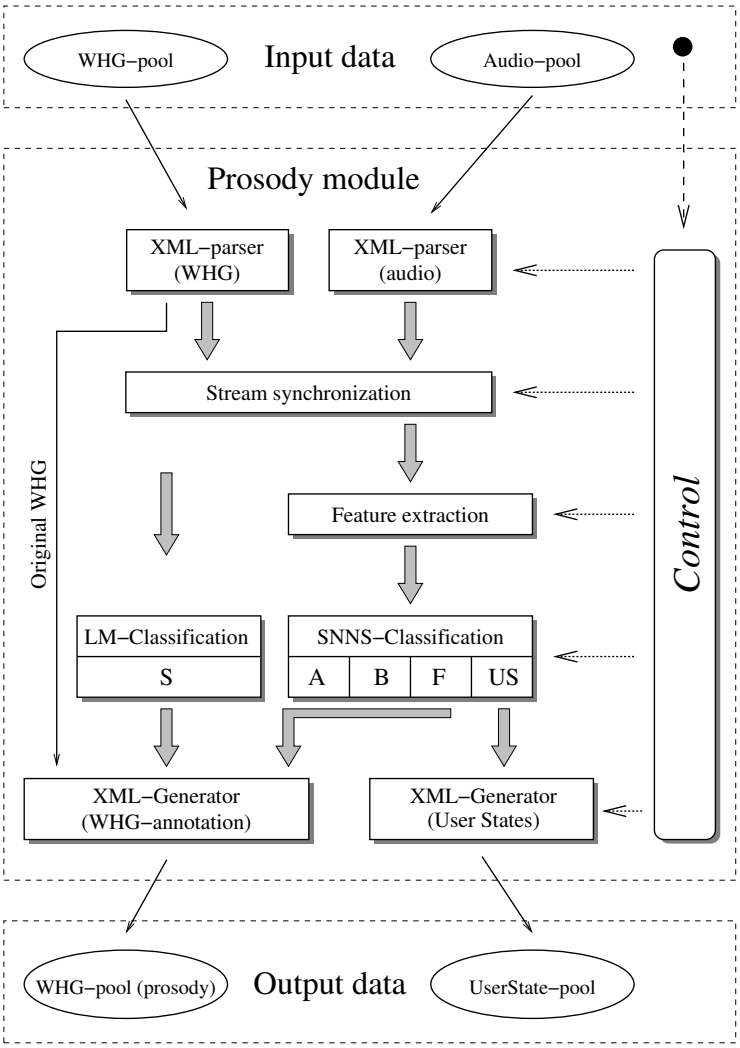


Abbildung 20: Die Struktur des Prosodiemoduls

Eine offene Frage im Bereich der Emotionserkennung stellt die „Obergrenze“ der Erkennungs-

leistung eines automatischen Emotionserkenners dar. Diese wird üblicherweise durch die Erkennungsleistung eines Mensches festgelegt. Daher ist eine vergleichende Analyse zwischen Mensch und Maschine von Interesse, wobei der Mensch sich, genau wie die Maschine, nur anhand der isolierten Sprachdaten, also ohne jegliches Kontext- bzw. Situationswissen, entscheiden muss (siehe auch 3.4). Diese Analyse wird in [38] durchgeführt. Darüber hinaus befasst sich diese Arbeit mit unterschiedlichen Ansätzen zur Selektion des Trainingsmaterials und zeigt, dass bei der Anwendung einer der vorgestellten Selektionsmethoden die Erkennungsleistung des Klassifikators verbessert weerden kann.

Für eine robuste prosodische Analyse ist eine robuste Schätzung von prosodischen Basismerkmalen, vor allem der Grundfrequenz (F0) ausschlaggebend. Ein wesentlicher Teil dieser Verfahren ist die Detektion von Anregungsarten der Sprache. Es wurde festgestellt, dass die Leistung des am Lehrstuhl eingesetzten Algorithmus stark einbricht, wenn die Sprachdaten verrauscht oder verhallt sind, was unweigerlich bei den Aufnahmen mit einem Raummikrofon auftritt (wie z.B. im *SmartKom Public*-Szenario: *http://www.smartkom.org/public_de.html*). Um die Untersuchungen zur robusten Detektion von Anregungsarten der Sprache durchzuführen, wurde eine gemischte heterogene Stichprobe zusammengestellt. Diese Stichprobe enthält Sprachdaten in unterschiedlicher Qualität (von qualitativ guten Aufnahmen mit Nahbesprechungsmikrofon bis hin zu stark verrauschten oder verhallten Aufnahmen und Aufnahmen mit Lombard-Effekt), die mit einer framesynchronen Annotation von Anregungsarten versehen waren. Im Rahmen einer Diplomarbeit wurde nach einer Kombination aus einer Merkmalsmenge und einem Klassifikator gesucht, die die gewünschte Robustheit und Generalisierungsfähigkeit zu bieten hatte. Als optimal hat sich die Kombination aus einem Gauss-Klassifikator und der MFCC-Spracherkennungsmerkmalen erwiesen, mit der z.B. auf der *SmartKom*-Stichprobe eine Steigerung der Erkennungsrate von 85 % auf 95 % erzielt wurde. In das Prosodiemodul ist neben den Klassifikatoren zur Erkennung der Benutzerzustände mittels prosodischer Merkmale auch die Erkennung von syntaktisch-prosodischen Grenzen (sog. M-Grenzen) integriert. Für die Abgabeversion des Moduls wurden die der Klassifikation zugrunde liegenden Sprachmodelle um den aktuellen Wortschatz und weitere *SmartKom WoZ*-Dialoge (Wizard-of-Oz) erweitert. Auf einer Stichprobe aus den SmartKom WoZ-Dialogen wurde damit eine Gesamterkennungsrate von 92 % bei der Klassifikation von Grenze vs. Nicht-Grenze erreicht.

Für die Erkennung von Benutzerzuständen liegt am Lehrstuhl die MMEG-Stichprobe vor, bei der Probanden gleichverteilt verschiedene Benutzerzustände aus dem http://www.smartkom.org-Szenario spielten. Die Daten beinhalten Aufnahmen in den drei SmartKom-Modalitäten Sprache, Mimik und Gestik für die Zustände *neutral*, *ärgerlich*, *freudig* und *zögerlich*. Zur besseren Beurteilung von automatischen, auf Sprache basierenden Klassifikatoren wurden die Daten auch manuell, das heisst von Menschen annotiert und die Ergebnisse ausgewertet und der automatischen Analyse gegenüber gestellt.

Am Lehrstuhl liegt eine Stichprobe von Sprache unter dem Einfluss von Müdigkeit vor, die im Rahmen eines entsprechenden Experiments zusammen mit Begleitdaten aufgenommen wurde. Die Stichprobe beinhaltet im Wesentlichen gelesene Texte des Verbmobil- und des SmartKom-Szenarios. Die Stichprobe wurde nun für die Klassifikation komplett aufbereitet und somit z.B. das zugehörige Lexikon angepasst und die Zeitzuordnung berechnet. Für die Klassifikationsexperimente wurde eine variierbare Abbildung der sieben bzw. acht verschiedenen annotierten

Müdigkeitsgrade auf zwei Klassen vorgenommen.

In Experimenten mit der Müdigkeitsstichprobe wurden bei der wortweisen Klassifikation mit Neuronalen Netzen unter Verwendung prosodischer Merkmale beim Zwei-Klassen-Problem *müde* vs. *nicht-müde* erfolgsversprechende Ergebnisse erzielt. Bei dem Verfahren wurden wortweise berechnete prosodische Merkmale verwendet, die auf einer erzwungenen Zeitzuordnung beruhen. Bei den Experimenten konnte gezeigt werden, dass unterschiedliche Merkmalgruppen der verwendeten Merkmale wie F0-, Energie- oder auch Dauer-basierte Merkmale das Klassifikationsergebnis unterschiedlich stark beeinflussen und damit vom Grad der annotierten Müdigkeit beeinflusst werden. Ausserdem wurde festgestellt, dass gemäß der verwendeten Abbildung auf zwei Klassen die Müdigkeit bei verschiedenen Sprechern unterschiedlich gut erkannt wird.

Das Mimik-Modul hat die Aufgabe, den Benutzerzustand eines Anwenders anhand seines Gesichtsausdrucks zu detektieren und diesen an das Dialogsystem zu melden. Für den Demonstrator, der auf der internationalen Statuskonferenz *Human Computer Interaction* in Berlin vorgestellt wurde, musste eine robuster Version des Mimik-Moduls ins Gesamtsystem integriert werden. Wichtig war in diesem Zusammenhang

- eine Unabhängigkeit von der Umgebung
- und eine stabile Erkennung des Benutzerzustands.

Die Unabhängigkeit von der Umgebung beinhaltet die Beleuchtung und den Hintergrund. Das Umgebungslicht setzt sich zusammen aus Tageszeit-abhängigem natürlichem Licht und Kunstlicht, das nicht vorhersehbar ist. Unterstützt wird die Beleuchtung deshalb durch indirekte Halogenlämpchen am Demonstrator. Um die Anzahl von Fehlklassifikationen bei der Gesichtslokalisation zu verringern und somit das Verfahren unabhängiger vom Hintergrund zu machen, wurde das *Circle-Frequency-Filter* (CFF) auf das Problem adaptiert und im Modul integriert.

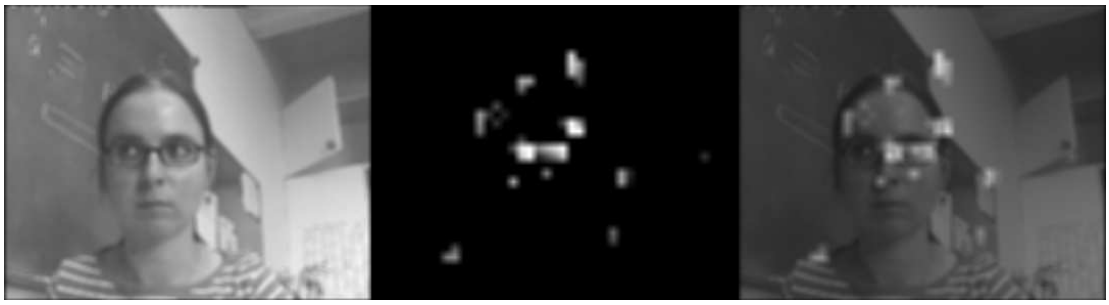

Abbildung 21: Für das linken Originalbild wurden die Bewertungen des CFF berechnet und im mittleren Bild dargestellt. Helle Bereiche bezeichnen gute Bewertungen. Zur besseren Anschaulichkeit wurden diese beiden Bilder addiert und es entsteht das rechte Bild.

Das CFF detektiert Hell-Dunkel-Wechsel auf einer Kreislinie, wie sie im Bereich der Nasenwurzel auftreten. Die Resultate, die ein CFF liefert, sind in Abbildung 21 illustriert. Diese zusätzliche Wissensquelle, kombiniert mit den bisher eingesetzten holistischen Verfahren der SVM und Eigenräumen, erhöht die Lokalisationsgenauigkeit für das Gesicht.

Um die bestehende Funktionalität des Mimik-Moduls für ein breites Publikum im Demonstrator „sichtbar“ zu machen, wurden zwei Fallbeispiele entwickelt. Ein positiver/freundiger bzw. negativer/ärgerlicher Gesichtsausdruck während der Anzeige einer Filmauswahl zeigt dem Dialogsystem die Vorliebe oder Abneigung bezüglich eines bestimmten Film-Genres. In der Fortführung des Dialogs wird dieser Genretyp bevorzugt dargestellt bzw. an das Ende einer Liste gesetzt. Die Benutzerzustände *Freude* und *Ärger* wurden ausgewählt, da sie sowohl vom beobachtenden Publikum, als auch vom System gut erkannt werden.

Ein wichtiger Aspekt für die Gesichtsausdruckserkennung in einem automatischen Sprachdialogsystem ist die Verfälschung des Gesichtsausdrucks durch die Lippenbewegungen aufgrund des Sprechens eines Bedieners. Um dieses Problem zu lösen, können vor der Gesichtsausdrucksklassifikation geeignete Gesichtsregionen ausgewählt werden, die einen Benutzerzustand repräsentieren. Die Auswahl dieser Gesichtsregionen wird von der Mimikerkennung automatisch durchgeführt. Die Ergebnisse der Benutzerzustandsklassifikation einzelner Regionen werden fusioniert und können in Dialogsituationen, in denen der Anwender spricht, als Erkennungsergebnis an das Dialogsystem weiter gegeben werden.

# 4 Professur für medizinische Bildverarbeitung

Leitung: J. Hornegger
(M. Prümmer)

## 4.1 Sprachgesteuerte Gefäßanalyse für die interventionelle Anwendung

Eine Intervention ist ein medizinischer Eingriff zur Behandlung pathologischer Gefäße. Zur Behandlung einer Stenose wird ein Stent mittels eines Katheters beispielsweise an der Hüfte eingeführt und durch das Gefäßsystem zum krankhaften Gefäß durchgeschoben. Um den Stent möglichst präzise auszuwählen und schnell zu plazieren, ist es erforderlich eine Gefäßkarte zu erstellen und eine Quantifizierung des pathologischen Gefäßabschnittes vorzunehmen. Für die richtige Wahl des Stents wird der Gefäßdurchmesserverlauf und die Länge des stenotisierten Gefäßabschnittes benötigt. Die Entwicklung von Algorithmen zur quantitativen Auswertung einer Gefäßaufnahme sowie zur computerunterstützten Gefäßnavigation konzentriert sich zunehmend auf deren Einsatz in 3-D-Angiographie Systemen. Da im Verlauf einer interventionellen Behandlung der Arzt den OP-Tisch und somit den sterilen Bereich verlassen muss um an einer Workstation die Gefäßanalyse durchzuführen, stellt eine sprachgesteuerte Gefäßanalyse im interventionellen Umfeld eine wertvolle Bereicherung dar. Aufbauend auf eine Arbeit zur *quantitativen Analyse von Volumendaten* wurde ein bestehendes SW-Packet (Syngo-Applikation, Siemens Medical Solutions, Forchheim) mit einem Spracherkenner der Firma Sympalog (www.sympalog.de) ausgestattet. Für eine klinische Evaluierung wurde ein Prototyp realisiert, der es ermöglicht eine Stenosen-Selektion und Quantifizierung mittels Sprachkommandos durchzuführen. Ein mit einem Funkmikrofon ausgestatteter Arzt ist damit in der Lage eine Gefäßanalyse durchzuführen, ohne den OP-Tisch verlassen zu müssen. Funktionell kann eine semi-automatische Schwellwertsegmentierung (Sprachbefehl: *Erhöhe bitte den Schwellwert um 32*), die Orientierung und Größe eines 3D-Volumens (Sprachbefehl: *Vergrößere/verkleinere das Volumen* oder *Rotiere das Volumen nach unten*) sowie die Selektion einer Stenose mittels eines computerunterstützten 3D-Zeigers via Sprachbefehl durchgeführt werden. Eine Sprachsteuerung bietet den Vorteil, dass alle Steuerbefehle in einer Kommandoebene angeordnet und somit umständliche hierarchische Menüs zu vermeiden werden können. Der entwickelte Prototyp wurde in eine Siemens Leonardo Workstation integriert und erlaubt eine Sprachsteuerung eines typischen Workflows einer Gefäßanalyse. Der Prototyp wird derzeit am Institut für Röntgendiagnostik der Universität Würzburg klinisch evaluiert und weiterhin untersucht in welchem Umfang die sprachgesteuerte Gefäßanalyse verbessert werden kann.

## 4.2 Koronarangiografie - Rekonstruktion von Herzkranzgefäßen aus Aufnahmen herkömmlicher C-Bogen Systeme

Das Ziel ist die datengetriebene 4D-Rekonstruktion von Koronarien aus herkömmlichen angiografischen Projektionen. Koronarangiografie (Koronarien = Herzkranzarterien, Angiografie

= Gefäßdarstellung) ist ein bildgebendes Verfahren, das mit Hilfe von Kontrastmittel den Innenraum der Herzkranzgefäße sichtbar macht. Mittels einer C-Bogen Anlage werden auf einer Kreisbahn um 180 Grad um den Patienten eine Folge von ca. 200 Röntgenbildern aufgenommen. Aus diesen Projektionen können heute starre Objekte, wie beispielsweise das zerebrale Gefäßsystem, durch Anwendung der inversen Radontransformation dreidimensional rekonstruiert werden. Da in der Kardiologie eine Aufnahme der Bildfolge über mehrere Herzzyklen erfolgt, versagt diese Methode bei Herzkranzarterien. Die Herzbewegung verursacht im rekonstruierten Volumen starke Bewegungsartefakte, die derzeit eine klinische Anwendung verhindern. Mit diesem schwierigen Problem beschäftigt sich die AG Medizinische Bildverarbeitung. Bisher wurde die Infrastruktur zur Rekonstruktion starrer Gefäße geschaffen. Dazu wurde ein Algorithmus (in C++) zur gefilterten Rückprojektion angepasst und eine Schnittstelle zu Matlab geschaffen, um Rekonstruktionsergebnisse zu visualisieren. Weiterhin wurde ein C-Arm Modell, das eine virtuelle Angiographie von Modellen wie zum Beispiel eine pulsierende Kugel oder ein Zylinder ermöglicht, programmiert. Derzeit werden Algorithmen zur Schätzung und Kompensation der Bewegung in der Projektion realisiert.

# 5 Bachelor-Arbeiten

1. Schmidt, F.: Robuste Tiefenkarten mit dem Roboterarm AESOP, November 2003

# 6 Master-Arbeiten

1. Yijiong, M.: SLAM Using Active Vision, Juli 2003
2. Zhou, X.: A Visual Guidance System for Tasks in Service Robotics, Juli 2003

# 7 Studienarbeiten

1. Jäger, F.: 3D Object Recognition for Service Robots, Januar 2003
2. Cincarek, T.: Klassifikation von Sprechergruppen, Juni 2003
3. Kähler, O.: Adaptive Sensordatenfusion in der Bewegungsdetektion, Juli 2003
4. Hönig, F.: Kernelmethoden in der Spracherkennung, Juli 2003

# 8 Diplomarbeiten

1. Weiß, R.: Robuste Klassifikation von Anregungsarten der Sprache, Juni 2003
2. Müller, R.: Anwendung des optimalen Designs von Experimenten im aktiven Rechnersehen, August 2003

# 9 Dissertationen

1. Warnke, V.: Integrierte Segmentierung und Klassifikation von Äußerungen und Dialogakten mit heterogenen Wissensquellen, April 2004

2. Reinhold, M.: Robuste, probabilistische, erscheinungsbasierte Objekterkennung, Oktober 2003

3. Buckow, J.: Multilingual Prosody in Automatic Speech Understanding, Oktober 2003

4. Heigl, B.: Plenoptic Scene Modelling from Uncalibrated Image Sequences, November 2003

5. Zobel, M.: Optimale Brennweitenwahl für die multiokulare Objektverfolgung, Dezember 2003

# 10 Habilitationen

1. Denzler, J.: Probabilistische Zustandsschätzung und Aktionsauswahl im Rechnersehen, Juni 2003

# 11 Vorträge

1. Batliner, A.: User States, User Strategies, and System Performance: How to Match the One with the Other, ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems, Chateau d'Oex, Schweiz, 29.08.2003

2. Batliner, A.: We are not amused - but how do you know? User states in a multi-modal dialogue system, European Conf. on Speech Communication and Technology (EUROSPEECH) 2003, Genf, Schweiz, 02.09.2003

3. Batliner, A.: Automatic recognition of emotional states, Journee parole expressive, Grenoble, Frankreich, 20.11.2003

4. Denzler, J.: Optimale Sensordatenauswahl und –verarbeitung im aktiven Rechnersehen, Kolloqium des Fachbereichs Informatik, Universität Kaiserslautern, Kaiserslautern, 15.01.2003.

5. Denzler, J.: Methoden des Rechnersehens für intelligente, interagierende Systeme, Kolloqium des Fachbereichs Informatik, Technische Universität Darmstadt, Darmstadt, 17.01.2003.

6. Denzler, J.: Optimale Sensordatenauswahl für die Zustandsschätzung im Rechnersehen, Kolloqium der Mathematisch–Geographischen Fakultät, Katholische Universität Eichstätt, Eichstätt, 08.04.2003.

7. Denzler, J.: Rekonstruktion und Anwendungung bildbasierter Objekt– und Szenenmodelle im Rechnersehen, Kolloqium der Fakultät für Mathematik und Informatik, Universität Leipzig, Leipzig, 26.06.2003.

8. Denzler, J.: Optimale Sensordatenauswahl für die Objekterkennung mit Anwendungen in der Service–Robotik, Habilitationsvortrag, Technische Fakultät, Universität Erlangen–Nürnberg, Erlangen, 18.06.2003.

9. Denzler, J.: Information Theoretic Focal Length Selection for Real–Time Active 3–D Object Tracking, International Conference on Computer Vision (ICCV), 11.-17. Oktober 2003, Nizza, Frankreich, 14.10.2003.

10. Denzler, J.: Rekonstruktion und Anwendungung bildbasierter Objekt– und Szenenmodelle im Rechnersehen, Kolloqium der Fakultät für Mathematik und Informatik, Universität Jena, Jena, 03.12.2003.

11. Deventer, R.: Using Test Plans for Bayesian Modeling Third International Conference Machine Learning and Data Mining, Leipzig, Germany, 07.07.2003

12. Deventer, R.: Bayesian controller versus traditional controller International Conference on Computational Intelligence For Modelling, Control & Automation - CIMCA 2003, Wien, Austria, 14.02.2003

13. Grzegorzek, M.: Erscheinungsbasierte Lokalisation und Klassifikation von Objekten, Graduiertenkolleg 3D-Bildanalyse und -synthese, Universität Erlangen, Erlangen, 03.11.2003

14. Grzegorzek, M.: How Fusion of Multiple Views Can Improve Object Recognition in Real-World Environments, 8th International Fall Workshop Vision, Modelling and Visualization, 19.-21. November 2003, München, 21.11.2003

15. Hacker, C.: Various Information Sources for HMM with Weighted Multiple Codebooks, Speech Processing Workshop, 09.09.2003, Magdeburg

16. J. Hornegger: Medizinische Bildverarbeitung bei Siemens Medical Solution. Mathematik–Kolloquium an der Universität Duisburg, 19.03.2003

17. J. Hornegger: Digitale Bildverarbeitung in der minimal–invasiven Diagnostik und Therapie. Kolloquium der Elektrotechnik und Informationstechnik an der RWTH Aachen, 17.06.2003

18. J. Hornegger: Interaktive Medizinische Bildverarbeitung. Kolloquium Neue Methoden und Verfahren der Informationsverarbeitung im Gesundheitswesen an der Friedrich–Alexander–Universität Erlangen–Nürnberg, 25.11.2003

19. H. Niemann: Using Lightfields in Image Processing. Dep. of Electrical Engineering, Stanford Univ., Stanford, CA, USA, 17.03.2003

20. H. Niemann: Using Lightfields in Image Processing. Imaging and Visualization Dep., Siemens Corp. Research, Princeton, NJ, USA, 20.03.2003

21. H. Niemann: Stochastische Modellierung für die Objekterkennung. Fakultät für Informatik, Univ. Lübeck, Lübeck, 09.05.2003

22. H. Niemann: Image–based Modeling and its Application in Image Processing. 6th Open German – Russian Workshop on Pattern Recognition and Image Processing (OGRW-6-2003), Katun Village, Altai Republic, Russian Federation, 25.08.2003

23. H. Niemann: Knowledge-based Exploration of Scenes. Dagstuhl-Seminar 03441 Cognitive Vision Systems, Dagstuhl, 29.10.2003

24. Schmidt, J.: Robust Hand-Eye Calibration of an Endoscopic Surgery Robot Using Dual Quaternions, Pattern Recognition, 25th DAGM Symposium, Magdeburg, Germany, 12.09.2003

25. Nöth, E., Prosody and Automatic Speech Recognition - Why not yet a Success Story and where to go from here, Symposium on Prosody and Speech Processing, Universität Tokio, Japan, 5.2.2003

26. Nöth, E., Multimodale Eingabe in der Mensch-Maschine-Kommunikation, Universität Augsburg, Augsburg, 7.7.2003

27. Nöth, E., Context-Sensitive Evaluation and Correction of Phone Recognition Output, Interspeech 2003, Genf, Ch, 4.9.2003

28. Nöth, E., Erkennung vvon UserStates mit Mimik und Prosodie, SmartKom-Projektstandssitzung, Stuttgart 5.9.2003

29. Nöth, E., Multimodal Input in Human-Machine-Communication, TSD '03, Budvar, CZ, 9.9.2003

30. Nöth, E., Audi-natif - Ein AUtomatisches DIalogsystem für NATürliches Interaktives Fremdsprachenlernen, IBM, Mannheim 24.9.2003

31. Nöth, E., Audi-natif - Ein AUtomatisches DIalogsystem für NATürliches Interaktives Fremdsprachenlernen, Klett-Verlag, Stuttgart, 13.11.2003

32. Steidl, S.: Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer, Pattern Recognition, 25th DAGM Symposium, Magdeburg, Germany, 12.09.2003

33. Stemmer, G.: Acoustic Normalization of Children's Speech, European Conf. on Speech Communication and Technology (EUROSPEECH) 2003, Genf, Schweiz, 02.09.2003

34. Stemmer, G.: Context-Dependent Output Densities for Hidden Markov Models in Speech Recognition, European Conf. on Speech Communication and Technology (EUROSPEECH) 2003, Genf, Schweiz, 02.09.2003

35. Stemmer, G.: Robust Speech Recognition for Children, Istituto Trentino di Cultura, Centro per la Ricerca Scientifica e Tecnologica (ITC-irst), Trento, Italien, 26.05.03

36. Vogt, F.: A System for Real-Time Endoscopic Image Enhancement, Medical Image Computing and Computer Assisted Intervention (MICCAI) 2003, Montréal, Kanada, 15.-18.11.2003, Montréal 17.11.2003

37. Vogt, F.: Endoskopische Lichtfelder mit einem kameraführenden Roboterarm, Workshop Bildverarbeitung für die Medizin - Algorithmen, Systeme, Anwendungen, http://www.bvm-workshop.org/, 2003, Lübeck, 11.03.2003

38. Vogt, F.: Qualitätskriterien Teilprojekt B6, Arbeitskreis Visualisierung des Sonderforschungsbereichs 603,

39. Zeissler, V.: Emotionserkennung in einem automatischen Dialogsystem: ist der Mensch besser als die Maschine, 14. Konferenz Elektronische Sprachsignalverarbeitung (ESSV2003), Universität Karlsruhe, 24. -26. Sept. 2003, Karlsruhe, 25.09.2003

40. Zinsser, T.: A Refined ICP Algorithm for Robust 3-D Correspondence Estimation, International Conference on Image Processing, Barcelona, Spanien, 16.09.2003

41. Zinsser, T.: Performance Analysis of Nearest Neighbor Algorithms for ICP Registration of 3-D Point Sets, Vision, Modeling, and Visualization 2003, München, Deutschland, 19.11.2003

# Literatur


[1] J. Adelhardt, R. Shi, C. Frank, V. Zeißler, A. Batliner, E. Nöth, H. Niemann: *Multimodal User State Recognition in a Modern Dialogue System*, in *Proc. of the 26th German Conference on Artificial Intelligence, KI '03*, 2003, S. 591–605.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Nöth: *How to Find Trouble in Communication*, *Speech Communication*, Bd. 40, 2003, S. 117–143.

[3] A. Batliner, C. Hacker, S. Steidl, E. Nöth, J. Haas: *User States, User Strategies, and System Performance: How to Match the One with the Other*, in *Proc. of an ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, ISCA, Chateau d'Oex, Aug. 2003, S. 5–10.

[4] A. Batliner, E. Nöth: *Prosody and Automatic Speech Recognition - Why not yet a Success Story and where to go from here*, in *Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, Tokyo, 2003, S. 357–364.

[5] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, E. Nöth: *We are not amused - but how do you know? User states in a multi-modal dialogue system*, in EUROSPEECH 2003 [14], S. 733–736.

[6] S. Bouattour, B. Heigl, D. Paulus, J. Hornegger: *An Iterative Approach of local 3D-Reconstruction of Non-Rigid Movements of Heart Region from many Angiographics Views*, in Ertl et al. [13], S. 183–190.

[7] F. Deinzer, J. Denzler, H. Niemann: *Viewpoint Selection - Planning Optimal Sequences of Views for Object Recognition*, in Petkov und Westenberg [27], S. 65–73.

[8] J. Denzler, B. Heigl, M. Zobel, H. Niemann: *Plenoptic Models in Robot Vision*, *Künstliche Intelligenz*, 2003, S. 62–68.

[9] J. Denzler, M. Zobel, H. Niemann: *Information Theoretic Focal Length Selection for Real-Time Active 3-D Object Tracking*, in *International Conference on Computer Vision*, IEEE Computer Society Press, Nice, France, October 2003, S. 400–407.

[10] R. Deventer, J. Denzler, H. Niemann: *Bayesian controller versus traditional controllers*, in M. Mohammadian (Hrsg.): *2003 International Conference on Computational Intelligence for Modelling Control and Automation*, 2003.

[11] R. Deventer, J. Denzler, H. Niemann, O. Kreis: *Using Test Plans for Bayesian Modeling*, in P. Perner, A. Rosenfeld (Hrsg.): *Machine Learning and Data Mining in Pattern Recognition*, Springer, Berlin, 2003, S. 307–316.

[12] G. W. Ehrenstein, S. Amesöder, L. F. Diaz, H. Niemann, R. Deventer: *Werkstoff- und prozessoptimierte Herstellung flächiger Kunststoff-Kunststoff und Kunststoff-Metall-Verbundbauteile*, in M. Geiger, G. W. Ehrenstein (Hrsg.): *DFG Sonderforschungsbereich 396 Berichts- und Industriekolloqium 15./16. Oktober 2003*, Meisenbach, Bamberg, 2003, S. 149–178.

[13] T. Ertl, B. Girod, G. Greiner, H. Niemann, H.-P. Seidel, E. Steinbach, R. Westermann (Hrsg.): *Vision, Modeling, and Visualization 2003*, Aka / IOS Press, Berlin, Amsterdam, Munich, Germany, November 2003.

[14] *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, September 2003.

[15] C. Frank, E. Nöth: *Automatic Pixel Selection for Optimizing Facial Expression Recognition using Eigenfaces*, in Michaelis und Krell [25], S. 378–385, Lecture Notes in Computer Science.

[16] C. Frank, E. Nöth: *Optimizing Eigenfaces by Face Masks for Facial Expression Recognition*, in Petkov und Westenberg [27], S. 646–654.

[17] C. Gräßl, F. Deinzer, F. Mattern, H. Niemann: *Improving Statistical Object Recognition Approaches by a Parameterization of Normal Distributions*, in *6th German-Russian Workshop Pattern Recognition and Image Understanding*, Institute of Automation and Electrometry, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Katun, Russian Federation, Aug. 2003, S. 38–41.

[18] C. Gräßl, F. Deinzer, H. Niemann: *Continuous Parametrization of Normal Distributions for Improving the Discrete Statistical Eigenspace Approach for Object Recognition*, in V. Krasnoproshin, S. Ablameyko, J. Soldek (Hrsg.): *PRIP 03*, Bd. 1, Minsk, Mai 2003, S. 73–77.

[19] Gräßl, Ch. and Zinßer, T. and Niemann, H.: *Illumination Insensitive Template Matching with Hyperplanes*, in Michaelis und Krell [25], S. 273–280, Lecture Notes in Computer Science.

[20] M. Grzegorzek, F. Deinzer, M. Reinhold, J. Denzler, H. Niemann: *How Fusion of Multiple Views Can Improve Object Recognition in Real-World Environments*, in Ertl et al. [13], S. 553–560.

[21] C. Hacker, G. Stemmer, S. Steidl, E. Nöth, H. Niemann: *Various Information Sources for HMM with Weighted Multiple Codebooks*, in A. Wendemuth (Hrsg.): *Proceedings of the Speech Processing Workshop, Magdeburg, Germany, September 09, 2003*, Magdeburg, 2003, S. 9–16.

[22] S. Krüger, F. Vogt, W. Hohenberger, D. Paulus, H. Niemann, C. H. Schick: *Evaluation der rechnergestützten Bildverbesserung in der Videoendoskopie von Körperhöhlen*, in Wittenberg et al. [37], S. 293–297.

[23] M. Levit, H. Alshawi, A. Gorin, E. Nöth: *Context-Sensitive Evaluation and Correction of Phone Recognition Output*, in EUROSPEECH 2003 [14], S. 925–928.

[24] T. Maier, M. Benz, G. Häusler, E. Nkenke, F. W. Neukam, F. Vogt: *Automatische Grobregistrierung intraoperativ akquirierter 3D-Daten von Gesichtsoberflächen anhand ihrer Gauß'schen Abbilder*, in Wittenberg et al. [37], S. 11–15.

[25] B. Michaelis, G. Krell (Hrsg.): *Pattern Recognition, 25th DAGM Symposium*, Bd. 2781, Springer-Verlag, Berlin, Heidelberg, New York, Sep. 2003, Lecture Notes in Computer Science.

[26] D. Paulus, J. Hornegger: *Applied pattern recognition: A practical introduction to image and speech processing in C++*, Advanced Studies in Computer Science, Vieweg, Braunschweig, 4. Ausg., 2003.

[27] N. Petkov, M. A. Westenberg (Hrsg.): *Computer Analysis of Images and Patterns - CAIP '03*, Nr. 2756 in Lecture Notes in Computer Science, Springer, Heidelberg, August 2003.

[28] J. Schmidt, F. Vogt, H. Niemann: *Robust Hand-Eye Calibration of an Endoscopic Surgery Robot Using Dual Quaternions*, in Michaelis und Krell [25], S. 548–556, Lecture Notes in Computer Science.

[29] R. Shi, J. Adelhardt, V. Zeissler, A. Batliner, C. Frank, E. Nöth, H. Niemann: *Using Speech and Gesture to Explore User States in Multimodal Dialogue Systems*, in J.-L. Schwartz, F. Berthommier, M.-A. Cathiard, D. Sodoyer (Hrsg.): *Proceedings of the AVSP 2003*, Institut de la Communication Parlée, INP Grenoble, Grenoble, Stendhal, 2003, S. 151–156.

[30] S. Steidl, G. Stemmer, C. Hacker, E. Nöth, H. Niemann: *Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer*, in Michaelis und Krell [25], S. 600–607, Lecture Notes in Computer Science.

[31] G. Stemmer, C. Hacker, S. Steidl, E. Nöth: *Acoustic Normalization of Children's Speech*, in EUROSPEECH 2003 [14], S. 1313–1316.

[32] G. Stemmer, V. Zeissler, C. Hacker, E. Nöth, H. Niemann: *A Phone Recognizer Helps to Recognize Words Better*, in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Bd. 1, Hong Kong, 2003, S. 736–739.

[33] G. Stemmer, V. Zeissler, C. Hacker, E. Nöth, H. Niemann: *Context-Dependent Output Densities for Hidden Markov Models in Speech Recognition*, in EUROSPEECH 2003 [14], S. 969–972.

[34] F. Vogt, S. Krüger, H. Niemann, C. Schick: *A System for Real-Time Endoscopic Image Enhancement*, in R. E. Ellis, T. M. Peters (Hrsg.): *Medical Image Computing and Computer Assisted Intervention (MICCAI) 2003*, Bd. 2879, Springer Verlag, Berlin, Montreal, Kanada, Nov. 2003, S. 356–363, Lecture Notes in Computer Science (LNCS).

[35] F. Vogt, S. Krüger, D. Paulus, H. Niemann, W. Hohenberger, C. H. Schick: *Endoskopische Lichtfelder mit einem kameraführenden Roboter*, in Wittenberg et al. [37], S. 418–422.

[36] S. Weber, T. Schüle, C. Schnörr, J. Hornegger: *A Linear Programming Approach to Limited Angle 3D Reconstruction from DSA-Projections*, in T. Wittenberg, P. Hastreiter, H. Handels, A. Horsch, H.-P. Meinzer (Hrsg.): *Bildverarbeitung für die Medizin 2003*, Springer, Berlin, March 2003, S. 41–45.

[37] T. Wittenberg, P. Hastreiter, U. Hoppe, H. Handels, A. Horsch, H.-P. Meinzer (Hrsg.): *7. Workshop Bildverarbeitung für die Medizin*, Springer Berlin, Heidelberg, New York, Erlangen, März 2003.

[38] V. Zeissler, J. Adelhardt, E. Nöth: *Emotionserkennung in einem automatischen Dialogsystem: ist der Mensch besser als die Maschine?*, in K. Kroschel (Hrsg.): *Elektronische Sprachsignalverarbeitung*, w.e.b. Universitätsverlag, Karlsruhe, Germany, September 2003, S. 114–121, ISBN 3-935712-83-9.

[39] T. Zinßer, J. Schmidt, H. Niemann: *A Refined ICP Algorithm for Robust 3-D Correspondence Estimation*, in *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[40] T. Zinßer, J. Schmidt, H. Niemann: *Performance Analysis of Nearest Neighbor Algorithms for ICP Registration of 3-D Point Sets*, in Ertl et al. [13], S. 199–206.

[41] M. Zobel, J. Denzler, B. Heigl, E. Nöth, D. Paulus, J. Schmidt, G. Stemmer: *MOBSY: Integration of Vision and Dialogue in Service Robots*, *Machine Vision and Applications*, Bd. 14, Nr. 1, 2003, S. 26–34.