Matthias Zobel, Joachim Denzler, Benno Heigl, Elmar Nöth, Dietrich Paulus,
Jochen Schmidt, Georg Stemmer

Demonstration von Bildverarbeitung und Sprachverstehen in der Dienstleistungsrobotik

erschienen in:
Autonome Mobile Systeme 2001, 17. Fachgespräch, Stuttgart, 11./12. Oktober 2001
Stuttgart, Deutschland
S. 141–147
2001

Demonstration von Bildverarbeitung und Sprachverstehen in der Dienstleistungsrobotik

Matthias Zobel, Joachim Denzler, Benno Heigl, Elmar Nöth, Dietrich Paulus,
Jochen Schmidt, Georg Stemmer

Lehrstuhl für Mustererkennung, Institut für Informatik
Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen
info@immd5.informatik.uni-erlangen.de,
URL: http://www5.informatik.uni-erlangen.de

Zusammenfassung Die typischerweise gewünschten Einsatzgebiete für Dienstleistungsroboter, z. B. Krankenhäuser oder Seniorenheime, stellen sehr hohe Anforderungen an die Mensch-Maschine-Schnittstelle. Diese Erfordernisse gehen im Allgemeinen über die Möglichkeiten der Standardsensoren, wie Ultraschall- oder Infrarotsensoren, hinaus. Es müssen daher ergänzende Verfahren zum Einsatz kommen. Aus der Sicht der Mustererkennung sind die Nutzung des Rechnersehens und des natürlichsprachlichen Dialogs von besonderem Interesse.
Dieser Beitrag stellt das mobile System MOBSY vor. MOBSY ist ein vollkommen integrierter autonomer mobiler Dienstleistungsroboter. Er dient als ein automatischer dialogbasierter Empfangsservice für Besucher unseres Instituts. MOBSY vereinigt vielfältige Methoden aus unterschiedlichsten Forschungsgebieten in einem eigenständigen System. Die zum Einsatz kommenden Methoden aus dem Bereich der Bildverarbeitung reichen dabei von Objektklassifikation über visuelle Selbstlokalisierung und Rekalibrierung bis hin zu multiokularer Objektverfolgung. Die Dialogkomponente umfasst Methoden der Spracherkennung, des Sprachverstehens und die Generierung von Antworten. Im Beitrag werden die zu erfüllende Aufgabe und die einzelnen Verfahren dargestellt.

1 Motivation

Die Entwicklung von Dienstleistungsrobotern erfordert das Zusammenspiel zahlreicher Forschungsrichtungen, z. B. Sensorik, Regelungstechnik, künstliche Intelligenz und neuerdings auch Rechnersehen und automatisches Sprachverstehen. Die beiden letztgenannten Disziplinen erlangten in der jüngsten Vergangenheit eine größere Bedeutung, da Dienstleistungsroboter dem Menschen in Bereichen wie zum Beispiel der Versorgung pflegebedürftiger Personen als persönlicher Assistent dienen sollen. Das bedeutet, dass sich Dienstleistungsroboter von anderen mobilen Robotersystemen hauptsächlich durch deren intensive Interaktion mit Menschen in natürlicher Umgebung unterscheiden. In den typischen Bereichen, in denen man Dienstleistungsroboter in Zukunft antreffen wird und teilweise schon antrifft, beispielsweise in Krankenhäusern oder

Diese Arbeit wurde durch die DFG gefördert im Rahmen des Sonderforschungsbereichs SFB 603/TP B2 und durch die BFS im Projekt DIROKOL. Die Verantwortung für den Inhalt dieses Beitrags liegt bei den Autoren.

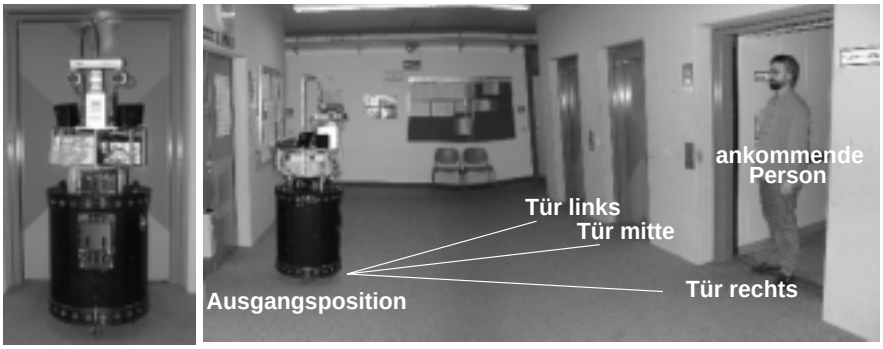

Abbildung 1. Der autonome mobile Dienstleistungsroboter MOBSY (links); Test- und Entwicklungsumfeld (rechts).

Altenpflegeeinrichtungen, übersteigen die Anforderungen an die Mensch-Maschine-Schnittstelle die Möglichkeiten klassischer Robotersensoren, wie Ultraschall-, Laser- oder Infrarotsensoren. Rechnersehen und natürliche Kommunikation und Dialogführung stellen somit eine notwendige Ergänzung der Sensorik solcher Systeme dar.

Dieser Beitrag konzentriert sich deshalb auf die beiden genannten Aspekte: Rechnersehen und natürlichsprachliche Kommunikation mit Dialogführung. Anhand des Anwendungsszenarios „automatischer Empfangsservice“ wird die erfolgreiche Integration aktueller Forschungsergebnisse aus beiden Bereichen in ein prototypisches System demonstriert. Angemerkt sei, dass im Gegensatz zu Arbeiten bei anderen Systemen, z. B. [2, 8], die technische Konstruktion eines Dienstleistungsroboters hier nicht im Vordergrund steht. Auch spielt der Aspekt des automatischen Wissenserwerbs und Lernens, obwohl hierzu bereits eigene Untersuchungen (u. a. [6, 10]) vorliegen, im präsentierten Stadium des Systems im Vergleich z. B. zu [4] eine untergeordnete Rolle.

Im nächsten Abschnitt wird die Aufgabe spezifiziert, die das mobile System MOBSY zu lösen hat. In Abschnitt 3 werden kurz die einzelnen eingesetzten Techniken aus dem Bereich des Rechnersehens und der Dialogkomponente beschrieben. Der Ansatz zur Selbstlokalisation wird dabei genauer vorgestellt, da dieses Problem typischerweise mittels klassischer Robotiksensoren gelöst wird. Der Beitrag schließt mit Ergebnissen und einem Ausblick auf zukünftige Verbesserungen und Anwendungen.

2 Anwendungsszenario

Das gewählte Umfeld, in dem MOBSY arbeitet, ist in Abbildung 1 dargestellt. Dabei handelt sich um einen Bereich vor den Aufzügen in unserem Institut. In dieser Umgebung agiert MOBSY als mobiler Empfangsservice für Besucher und Gäste. Näher spezifiziert bedeutet dies die Ausführung der folgenden Schritte:

– MOBSY wartet in seiner Ausgangsposition darauf, dass sich eine der drei Aufzugstüren öffnet. Dazu bewegt MOBSY seinen Kamerakopf so, dass die Türen in der Reihenfolge *Links*, *Mitte*, *Rechts*, *Mitte*, ... gesehen werden können.

– Wenn eine Person ankommt, nähert sich MOBSY dieser auf den in Abbildung 1 als Linien eingezeichneten Pfaden. Während dieser Annäherung stellt MOBSY sich als mobiles Empfangssystem vor und bittet die Person stehen zu bleiben. Gleichzeitig beginnt das System mit dem Kamerakopf das Gesicht der Person zu verfolgen, um einen ersten Kontakt mit der Person herzustellen.
– Nachdem MOBSY vor der Person angekommen ist, beginnt MOBSY mit dem natürlichsprachlichen Informationsdialog. Dabei wird weiterhin das Gesicht der Person verfolgt.
– Nach Beendigung des Dialogs dreht MOBSY sich um und fährt in seine Ausgangsposition zurück. Dort angekommen muss sich MOBSY auf Grund von Fehlern in der Odometrieinformation repositionieren.
– Danach fängt MOBSY wieder an, auf eine ankommende Person zu warten.

Diese Schleife wird so lange wiederholt, bis MOBSY extern unterbrochen wird. Die Ausführung der oben genannten Schritte erfordert das koordinierte Zusammenspiel von *Objektdetektion* und *Objektklassifikation*, *visueller Gesichtsverfolgung* und *Kamerasteuerung*, *natürlichsprachlichem Dialog*, *Roboternavigation einschließlich Hindernisvermeidung* und *visueller Selbstlokalisierung und Rekalibrierung*.

Die für diese Gebiete verwendeten Methoden werden detaillierter im folgenden Abschnitt 3 beschrieben. Da die Navigation und Hindernisvermeidung mit klassischen Infrarotsensoren realisiert ist und MOBSY auf vordefinierten Pfaden fährt, wird auf eine Darstellung dieses Moduls im Folgenden verzichtet.

3 Systemdesign und Module

Das eingesetzte mobile System besteht aus der eigentlichen mobilen Plattform, ein XR4000 der Firma Nomadic Technologies, und einem Aufbau zur Aufnahme von zusätzlicher Ausrüstung, z. B. Kamerakopf, Richtmikrofon, etc. Der Kamerakopf besitzt 10 Freiheitsgrade und ist ein Bisight/Unisight binokulares System der Firma HelpMate Robotics. Die gesamte Bild- und Sprachverarbeitung wird auf einem in die Plattform integrierten Dual Pentium II 300 MHz Rechner durchgeführt.

Die im Folgenden beschriebenen Module realisieren die Teilaufgaben der Spezifikation aus Abschnitt 2, in denen Bild- und Sprachverarbeitung verwendet wird. Ein wichtiger Aspekt, der hier aus Platzgründen nicht näher behandelt wird, ist die Integration dieser Module, damit ein koordiniertes Zusammenspiel gewährleistet ist.

Objektklassifikation. In dem gewählten Szenario wird erwartet, dass die Besucher des Instituts mit einem der drei Aufzüge ankommen. Daraus folgt, dass der Ankunft einer Person das Öffnen einer der Aufzugstüren voraus geht. Der Mechanismus, der das Ankommen einer Person anzeigt, basiert daher auf der Unterscheidung zwischen offenen und geschlossenen Aufzugstüren.

Zu diesem Zweck werden von Support Vektor Maschinen (SVM) als Klassifikator eingesetzt, da diese prädestiniert für das Lösen von Zweiklassenproblemen sind (vgl. [13] für eine detaillierte Beschreibung). Die verwendete SVM arbeitet auf Farbbildern der Größe 96×72, die vom Kamerakopf geliefert werden, und klassifiziert diese in die beiden Klassen *offen* und *geschlossen*.

Zum Training der SVM wurde eine Trainingsmenge von 337 Bildern der Aufzugstüren zusammengestellt. Die Trainingsmenge wurde manuell klassifiziert in 130 *geschlossene* und 207 *offene* Fälle. Eine Aufzugstür gilt dabei als *offen* bei einem Öffnungsgrad zwischen komplett offen und halb geschlossen. Im anderen Fall wird die Tür als *geschlossen* behandelt. Als SVM wurde das System SVM$^{\text{light}}$ [9] benutzt.

Eine offene Aufzugstür ist alleine nicht ausreichend, um über die Ankunft einer Person zu entscheiden. Man denke beispielsweise an die Situation, dass sich die Aufzugstüren öffnen und keine Person aussteigt. In der derzeitigen Realisierung von MOBSY führt dies dazu, dass das System auch in diesen Fällen das Ankommen einer Person fälschlicherweise annimmt und mit der Annäherungsphase beginnt; dies wird dann allerdings durch eine Zeitüberschreitung in der Dialogkomponente abgefangen.

Gesichtsverfolgung. Während MOBSY sich einer angekommenen Person nähert und auch während der eigentlichen Dialogphase sollen beide Kameras des Kamerakopfs auf das Gesicht der Person ausgerichtet sein, um den Kontakt zwischen Mensch und Maschine aufrechtzuerhalten. Die Fixation könnte dabei vom System auch dazu benutzt werden, um visuell über das Vorhandensein einer Person zu entscheiden, z. B. wenn die Person während des Dialog weggeht, oder auch zur Erkennung von Gesichtern.

Für die Gesichtsverfolgung müssen zwei Hauptprobleme gelöst werden: Gesichtsdetektion und Bewegungssteuerung der Kameras. Gesichtsdetektion basiert auf der Bestimmung von Hautfarbenregionen in Farbbildern [5] wobei für jeden Bildpunkt ein Farbabstand berechnet wird. Es werden Bilder der Größe 96×72 verwendet. Der Schwerpunkt der bestimmten Hautfarbenregion wird dabei als die Position des Gesichts interpretiert. Ausgehend von diesen Positionen werden Steuerungswinkel für die Neige- und Vergenzachsen des binokularen Kamerakopfs berechnet. Um die Bewegungen möglichst glatt zu halten, werden die Vergenzbewegungen mit der Zeit durch entsprechende Schwenkbewegungen des gesamten Kamerasystems ausgeglichen.

Natürlich ist Hautfarbensegmentierung nicht sehr spezifisch für Gesichter, aber die folgenden Fakten rechtfertigen aus unserer Sicht die Wahl dieses Vorgehens. Erstens ist es sehr wahrscheinlich, dass eine Hautfarbenregion in einer Höhe von ca. 1,7 m in dem gewählten Szenario durch ein Gesicht hervorgerufen ist, und zweitens hat es sich in der Experimenten durch seine Robustheit und Schnelligkeit bewährt.

Dialog. Sobald der Roboter die Person erreicht hat, initiiert das Dialogmodul das Gespräch mit einer Begrüßung und einer kurzen Einführung in die Fähigkeiten des Systems. Das Dialogmodul ist in vier Untereinheiten gegliedert, die eine Verarbeitungshierarchie bilden: Für jede Benutzeräußerung wird vom Spracherkenner eine Hypothese der gesprochenen Wortfolge ausgegeben. Diese Wortfolge wird von einem Parser in eine semantisch-pragmatische Repräsentation umgewandelt. Unter Berücksichtigung des aktuellen Dialogzustands erzeugt der Dialogmanager daraus eine Systemantwort. Diese wird schließlich sprachsynthetisch ausgegeben.

Alle Untereinheiten des Dialogmoduls müssen sowohl mit dem relativ hohen Geräuschpegel als auch mit den unterschiedlichen Benutzeräußerungen zurechtkommen. Der Geräuschpegel ist zum Teil auf die Umgebung des Roboters, z. B. die Aufzugstüren oder unbeteiligte Personen, aber auch auf die Plattform selbst zurückzuführen, da z. B. ständig eingebaute Ventilatoren in Betrieb sind. Auch sind die Äußerungen der Besucher des Instituts entsprechend vielfältig.

Damit die Hintergrundgeräusche vor und nach einer Benutzeräußerung die Erkennung nicht stören, fängt der Spracherkenner nur an zu arbeiten, wenn ein bestimmter Energieschwellwert im Signal für eine Mindestdauer überschritten wird. Sobald der Schwellwert für ein längeres Zeitintervall unterschritten worden ist, wird der Spracherkenner wieder angehalten. Hochfrequente Störungen, etwa durch die Eigengeräusche des Roboters, werden durch einen Tiefpassfilter entfernt. Der Erkenner verarbeitet kontinuierliche Sprache; das Lexikon enthält z. Zt. knapp 100 Wörter. Als akustische Merkmale werden Mel-Cepstrum-Koeffizienten und ihre ersten Ableitungen verwendet. Eine detaillierte Beschreibung des Spracherkenners findet sich in [7].

Die akustischen Modelle des Erkenners wurden mit ca. 900 gelesenen Sätzen an die Empfangsservice-Domäne adaptiert, das Sprachmodell des Erkenners enthält Bigramme. In der erkannten Wortkette erfolgt das Sprachverstehen durch eine Suche nach sinnvollen Phrasen, die bei der Entwicklung des Systems festgelegt wurden (vgl. [12]). Jede Phrase hat eine vordefinierte semantisch-pragmatische Repräsentation, auf die sie abgebildet wird. Dabei werden alle Wörter ignoriert, die keiner sinnvollen Phrase zugeordnet werden können. Diese einfache Strategie erhöht die Robustheit gegenüber falsch erkannten Wörtern und toleriert ein relativ hohes Maß an Variabilität der gesprochenen Eingabe. Der Dialogmanager speichert den aktuellen Dialogzustand und generiert regelbasiert unter Berücksichtigung der Eingabe eine angemessene Antwort. Wenn der Besucher z. B. fragt: „Und wo gibt's das?“, informiert MOBSY über den Ort, an dem die im Satz zuvor nachgefragte Information zu finden ist. Durch den gespeicherten Dialogzustand können Erkennungsfehler gefunden werden, die einen Widerspruch zwischen Dialogzustand und semantisch-pragmatischer Repräsentation verursachen.

Die Phrasen zur Begrüßung und zur Auskunft werden zufallsgesteuert aus einer Menge von gleichwertigen Phrasen ausgewählt. Die Sprachsynthese selbst basiert auf dem German Festival Sprachsynthesesystem [3, 11].

Selbstlokalisierung. Zur Selbstlokalisierung des Roboters wird eine an der Decke montierte Leuchtstoffröhre ausgenutzt. Die Roboterposition und -orientierung kann aus einem einzelnen Bild dieser Lampe berechnet werden, falls die gewünschte Lage des Roboters relativ zur Lampe aus vorhergehenden Messungen bekannt ist. Durch geeignete Korrekturbewegungen wird anschließend die gewünschte Position angefahren.

Abbildung 2 (rechts) zeigt die hier verwendete 3D-Konfiguration. Die Lampenposition sei definiert durch den Endpunkt $\boldsymbol{p}_1$ und einen beliebigen zweiten Vektor $\boldsymbol{p}_2$ auf der Röhre. Eine der beiden Kameras wird so positioniert, dass sie in Richtung $\boldsymbol{p}_1$ blickt. Eine aus dieser Position gewonnene Aufnahme ist in Abbildung 2 (links) zu sehen. Ist die Lampe in dieser ersten Ansicht nicht vollständig sichtbar, führt die Kamera Suchbewegungen durch. Im Bild können die projizierten Punkte $\boldsymbol{q}_1$ und $\boldsymbol{q}_2$ der entsprechenden 3D-Punkte $\boldsymbol{p}_1$ und $\boldsymbol{p}_2$ durch einfache Analyse des binarisierten Bildes ermittelt werden, wobei sich $\boldsymbol{q}_2$ auf einem beliebigen Punkt auf der durch die Leuchtstoffröhre festgelegten Geraden im Bild befinden kann. Diese Gerade wird durch lineare Regression aller hellen Punkte bestimmt. Der sichtbare Endpunkt wird durch einfache Suche entlang dieser Geraden gefunden.

Das 3D-Koordinatensystem wird so positioniert, dass sein Ursprung dem Projektionszentrum der Kamera entspricht, seine z-Achse senkrecht zum Fußboden ist und die y-Achse zur Vorderseite des Roboters zeigt. Außerdem wird angenommen, dass das

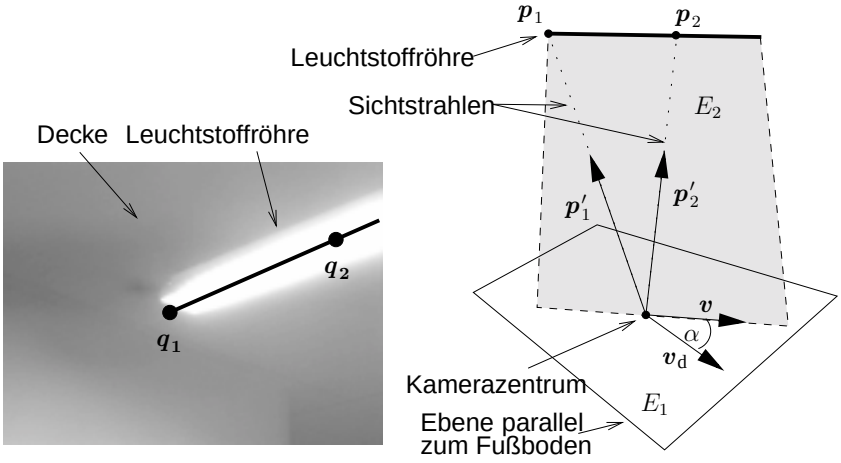

Abbildung 2. Beispielbild zur Selbstlokalisierung (links); die verwendete 3D-Konfiguration mit Bezeichnungen (rechts).

Projektionszentrum der Kamera dem Schnittpunkt von Schwenk- und Neigeachse des binokularen Kamerasystems entspricht und zusätzlich die Rotationsachse des Roboters schneidet. Diese approximierenden Annahmen sind in der Realität nicht exakt erfüllt, sie führen jedoch zu ausreichender Genauigkeit bei den Experimenten.

Die Ebene E_1 sei parallel zum Fußboden. Die Ebene E_2 schneide den Ursprung des Koordinatensystems und die Lampe entlang ihrer Längsachse. Der Vektor $\boldsymbol{v}$ zeige in Richtung der Geraden, die durch Schneiden dieser beiden Ebenen gebildet wird: $\boldsymbol{v} = (\boldsymbol{p}'_1 \times \boldsymbol{p}'_2) \times (0,0,1)^{\mathrm{T}}$.

Die gewünschten Koordinaten $\boldsymbol{p}_{\mathrm{d}}$ des Lampen-Endpunkts relativ zum Koordinatensystem sowie die gewünschte Richtung $\boldsymbol{v}_{\mathrm{d}}$ der Längsachse der Lampe ergeben sich aus der gewählten Konstellation (im gewählten Szenarion gilt $\boldsymbol{v}_{\mathrm{d}} = (0,-1,0)^{\mathrm{T}}$). Stünde der Roboter schon an der gewünschten Position, würde $\boldsymbol{p}_{\mathrm{d}}$ in die gleiche Richtung wie $\boldsymbol{p}'_1$ zeigen und $\boldsymbol{v}_{\mathrm{d}}$ in die gleiche Richtung wie $\boldsymbol{v}$. Ergeben sich Unterschiede, muss der Roboter um den Winkel $-\alpha$ rotiert werden. Die zur Korrektur notwendige Translation wird bestimmt, indem $\boldsymbol{p}'_1$ mit dem Winkel α um die z-Achse rotiert, das Ergebnis auf die Länge von $\boldsymbol{p}_{\mathrm{d}}$ skaliert und letztlich $\boldsymbol{p}_{\mathrm{d}}$ davon abgezogen wird.

4 Ergebnisse und Ausblick

Das vorgestellte System war während der 25-Jahrfeier unseres Instituts für mehr als zwei Stunden ohne Funktionsstörungen oder externe Eingriffe in Betrieb. MOBSY befand sich in dieser Zeit in einer Umgebung, in der permanent neue Besucher ankamen, Besucher sich unterhielten und dadurch ein hohes Hintergrundrauschen entstand, sowohl aus Sicht der Bild- als auch der Sprachverarbeitung (Bilder und Videoclips finden sich im Internet [1]). Es stellte damit seine Robustheit in einer für mobile Systeme typischerweise schwierigen Umgebung unter Beweis.

Auch weiterhin wird MOBSY regelmäßig für Demonstrationen eingesetzt, wobei die Fähigkeiten ständig erweitert werden. Ein erstrebenswertes Szenario ist, dass MOBSY nicht nur Auskunft gibt, sondern die Besucher basierend auf visueller Objektverfolgung und Navigation zu den Büros der Mitarbeiter oder zu anderen interessanten Positionen begleitet.

Der Aspekt der intelligenten Interaktion mit Menschen spielt eine immer wichtiger werdende Rolle im Bereich der Dienstleistungsrobotik. Daher müssen die Bereiche Rechnersehen und natürlichsprachlicher Dialog verstärkt mit der klassischen Sensorik zusammengeführt und integriert werden.

Literatur

1. http://www5.informatik.uni-erlangen.de/~mobsy.
2. R. Bischoff: *Recent Advances in the Development of the Humanoid Service Robot HERMES*, in *3rd EUREL Workshop and Masterclass - European Advanced Robotics Systems Development*, Bd. I, 2000, S. 125–134.
3. A. Black, P. Taylor, R. Caley, R. Clark: *The Festival Speech Synthesis System*, http://www.cstr.ed.ac.uk/projects/festival.html.
4. W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun: *The Interactive Museum Tour-Guide Robot*, in *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998, S. 11–18.
5. D. Chai, K. N. Ngan: *Locating Facial Region of a Head-and-Shoulders Color Image*, in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998, S. 124–129.
6. F. Deinzer, J. Denzler, H. Niemann: *Classifier Independent Viewpoint Selection for 3-D Object Recognition*, in G. Sommer, N. Krüger, C. Perwass (Hrsg.): *Mustererkennung 2000, 22. DAGM-Symposium, Kiel*, Springer, Berlin, September 2000, S. 237–244.
7. F. Gallwitz, M. Aretoulaki, M. Boros, J. Haas, S. Harbeck, R. Huber, H. Niemann, E. Nöth: *The Erlangen Spoken Dialogue System EVAR: A State-of-the-Art Information Retrieval System*, in *Proceedings of 1998 International Symposium on Spoken Dialogue (ISSD 98)*, Sydney, Australia, 1998, S. 19–26.
8. U. Hanebeck, C. Fischer, G. Schmidt: *ROMAN: A Mobile Robotic Assistant for Indoor Service Applications*, in *Proceedings of the IEEE RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1997, S. 518–525.
9. T. Joachims: *Making Large-Scale Support Vector Machine Learning Practical*, in Schölkopf et al. [13], S. 169–184.
10. F. Mattern: *Automatische Umgebungskartenerstellung durch probibilistische Fusion von Sensordaten mit einem autonomen mobilen System*, Studienarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg, 2000.
11. G. Möhler, B. Möbius, A. Schweitzer, E. Morais, N. Braunschweiler, M. Haase: *Speech Synthesis at the IMS*, http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html.
12. E. Nöth, J. Haas, V. Warnke, F. Gallwitz, M. Boros: *A Hybrid Approach to Spoken Dialogue Understanding: Prosody, Statistics and Partial Parsing*, in *Proceedings European Conference on Speech Communication and Technology*, Bd. 5, Budapest, Hungary, 1999, S. 2019–2022.
13. B. Schölkopf, C. Burges, A. Smola (Hrsg.): *Advances in Kernel Methods: Support Vector Learning*, The MIT Press, Cambridge, London, 1999.