17. Fachgespräch Autonome Mobile Systeme AMS'01,
Stuttgart, S. 113-119, Informatik aktuell, Springer Verlag

Multimodale Mensch-Maschine-Interaktion für Servicerobotik-Systeme*

H.-J. Böhme, T. Hempel, Ch. Schröter, T. Wilhelm, J. Key & H.-M. Gross

Fachgebiet Neuroinformatik, Technische Universität Ilmenau,
98684 Ilmenau (Thüringen)
email: hans@informatik.tu-ilmenau.de

Zusammenfassung Die Frage der Mensch-Roboter-Interaktion wird für zukünftige erfolgreiche Applikationen von Servicerobotern eine wesentliche, wenn nicht *die zentrale* Rolle spielen. Vor diesem Hintergrund beschreibt der vorliegende Beitrag ein Konzept zur multimodalen Mensch-Maschine-Kommunikation und dessen Realisierungsstand anhand eines konkreten Einsatzszenarios, der Entwicklung eines intelligenten, interaktiven und mobilen Informationskiosks im Baumarkt. Obwohl sich das vorgestellte Konzept an den konkreten Erfordernissen dieses Szenarios orientiert, stellt es einen generischen Ansatz dar, der sich auf eine Vielzahl weiterer Applikationen übertragen lässt.

1 Einleitung

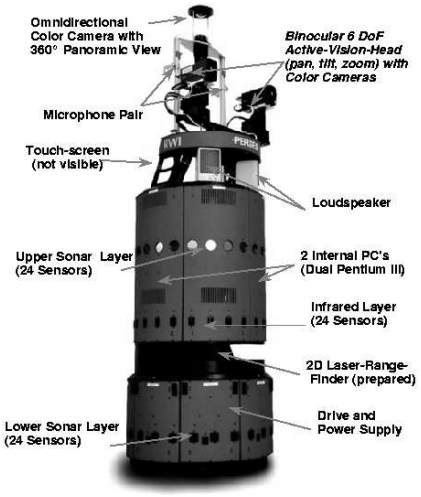

Abb. 1. Experimentalroboter PERSES.

Der Themenkreis Mensch-Roboter-Interaktion (MMI) gewinnt im Kontext angestrebter Serviceroboter-Applikationen zunehmend an Bedeutung [6]. Dies trifft insbesondere dann zu, wenn der Serviceroboter in einem relativ gering technisierten Bereich eingesetzt werden und mit uneingewiesenen Nutzern interagieren soll. Die notwendige Interaktionsfähigkeit hängt dabei auch von der konkret zu erbringenden Serviceleistung des Roboters ab. Für einen Reinigungsroboter, der einen Supermarkt während der Öffnungszeiten reinigen soll, stellen Menschen in erster Linie dynamische Hindernisse dar, die es zu umfahren gilt. Demgegenüber muss ein System, dass als mobiler Informationskiosk in der gleichen Einsatzumgebung operieren soll, hinsichtlich der Mensch-Roboter-Interaktion über völlig andere Möglichkeiten verfügen, da von dieser Interaktion maßgeblich die Erfüllung der Serviceaufgabe abhängt.

* gefördert durch das Projekt PERSES (TMWFK, Nr. B 611-98041)

Der in Abb. 1 dargestellte B21-Roboter PERSES dient als Experimentalplattform. Er verfügt über die technische Ausstattung, die zur multimodalen MMI notwendig ist. Zunächst werden in Abschnitt 2 die einzelnen Stufen der Interaktion mit ihren entsprechenden Methoden erläutert, bevor in Abschnitt 3 ein exemplarischer Interaktionszyklus skizziert und anhand experimenteller Ergebnisse demonstriert wird.

2 Methoden zur MMI

2.1 Aufmerksamkeitssteuerung

Um Hypothesen über das Vorhandensein von Personen in der Umgebung des stehenden Roboters zu generieren, erfolgt eine Bewegungsanalyse im omnidirektionalen Kamerabild, die unbewegte und bewegte Bildregionen segmentiert. Eine Auswertung der Bewegungsrichtung der detektierten Segmente sichert, dass sich der Roboter bevorzugt solchen Regionen zuwendet, die sich auf ihn zu bewegen.

Ein Verfahren zur Geräuschlokalisation, welches auf der Modellierung der menschlichen auditorischen Informationsverarbeitung basiert, bietet dem Interaktionspartner die Möglichkeit, auch akustisch die Aufmerksamkeit auf sich zu ziehen. Damit stehen dem Roboter zwei alternative Aufmerksamkeitstrigger zur Verfügung, die in einem Fusionsmodul integriert wurden, welches letztlich die Richtung determiniert, welcher sich der Roboter dann zuwendet.

2.2 Personenverifikation

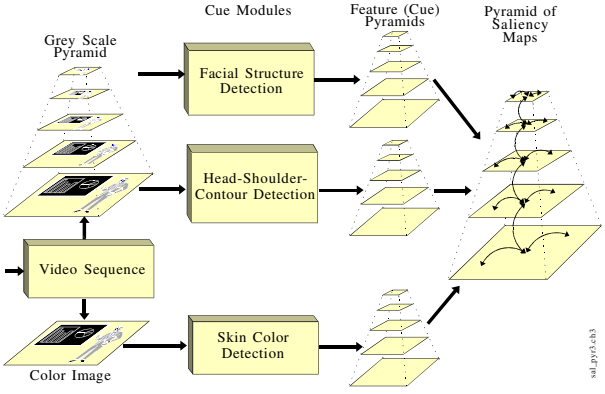

Abb. 2. Multi-Cue-Architektur zur Personenverifikation.

Die Architektur des Moduls zur Personenverifikation zeigt Abb. 2. Weder Bewegung noch Geräusche sind ausreichend, um wirklich sicher auf das Vorhandensein einer Person schließen zu können. Deshalb erfolgt nach der Zuwendung zu einer interessanten Richtung eine abschließende Personenverifikation, die die visuellen Cues Gesicht, Kopf-Schulter-Silhouette und Hautfarbe miteinander kombiniert und die im Vergleich zu [1, 3] hinsichtlich Robustheit und Performanz deutlich weiterentwickelt wurde. Dies betrifft insbesondere die Gesichtsdetektion, die in der aktuellen Implementierung über ein Cascade-Correlation-Netzwerk (CCNW) [?] realisiert wird. Die Vorteile beim Einsatz eines Neuronalen Netzes zur Gesichtsdetektion (siehe auch [7])

liegen darin begründet, dass die Gesichtsstruktur direkt als Verteilung von Intensitäts(Grau)werten aufgefasst werden kann, und dass durch das vorliegende Zweiklassenproblem (Gesicht vs. kein Gesicht) die Parameter des Klassifikators mittels eines Trainingsdatensatzes adjustiert werden können. Für das konkrete

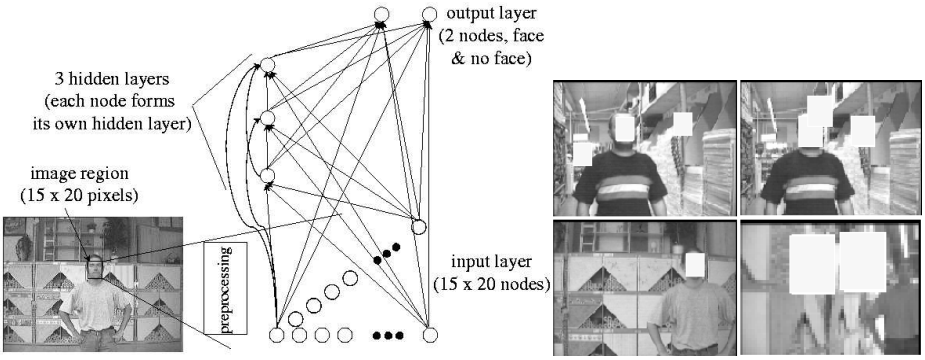

Abb. 3. Links: Topologie des CCNW, das für die Gesichtsdetektion eingesetzt wird. Rechts: Typische Gesichtsdetektionsergebnisse mittels Cascade-Correlation-Netzwerk, vor teilweise extrem strukturiertem Hintergrund. Die obere Reihe zeigt zwei Bilder, die sowohl korrekte als auch falsch-positive Detektionen enthalten, darunter ein Beispiel für ausschließlich korrekte Detektion und ein Bild, welches nur falsch-positive Detektionen aufweist.

CCNW spricht, dass es im Gegensatz zum Multilayer-Perceptron eine gleichzeitige Parameter- und Toplogieadaption gestattet, was letztlich zu einem optimal an die Problemkomplexität angepassten, minimalen Netzwerk führt. Dies ist ein entscheidender Aspekt hinsichtlich der Gesamtperformanz der Personenverifikation. Der Trainingsprozess startet mit einem minimalen (linearen) Netzwerk und generiert sukzessive neue Hiddenknoten, die so trainiert werden, dass sie maximal zur Fehlerreduktion am Netzwerkausgang beitragen.

Das Problem der Gewinnung von Negativbeispielen, also von Bildausschnitten, die typisch für „kein Gesicht“ sind, kann durch die Verwendung eines Bootstrap-Algorithmus [7] gelöst werden. Dabei nutzt man für den Beginn des Trainingsprozesses zunächst eine Menge zufällig ausgewählter Bildausschnitte, die kein Gesicht beinhalten. Im fortschreitenden Trainingsprozess wird dann das Netzwerk auf Bildern getestet, die keine Gesichter beinhalten, und alle falschpositiv detektierten Regionen werden der Menge der Negativbeispiele hinzugefügt, mit der dann der Trainingsprozess fortgesetzt wird. Abb. 3 zeigt die Topologie des gegenwärtig verwendeten CCNW sowie einige typische Detektionsergebnisse des Gesichtsdetektors.

Um die Sicherheit der Personenverifikation zu erhöhen, wird für die Fusion der Beiträge der einzelnen Cues und die daran anschließende Selektion der entsprechenden Position gefordert, dass an der jeweiligen 3D-Position innerhalb der Auflösungspyramide (Abb. 2) mindestens 2 der 3 Merkmalsdetektoren ein signifikantes Ergebnis beitragen. Für eine detailliertere Beschreibung sei auf [1, 2] verwiesen.

Das Verifikationsmodul operiert sowohl auf den Bildern der omnidirektionalen Kamera als auch auf Bildern der Frontalkamera(s). Dabei decken die Frontalkameras einen Entfernungsbereich von ca. 0.8m bis 3m (5 Auflösungsebenen), die omnidirektionale Kamera (3 Auflösungsebenen) von ca. 0.2m bis 1.5m ab. Die Personenverifikation liefert die Initialisierung für das sich anschließende Personentracking.

2.3 Personentracking

Auch das Personentracking nutzt alle zur Verfügung stehenden Kamerasysteme und operiert wiederum auf Auflösungspyramiden, um einen möglichst großen Entfernungsbereich abdecken zu können. Es basiert auf dem in [5] vorgeschlagenen CONDENSATION-Algorithmus, der in Abb. 4 veranschaulicht wird.

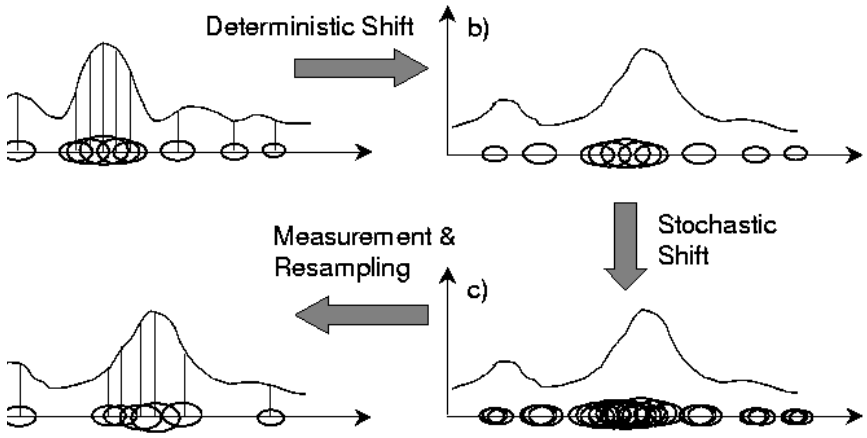

Abb. 4. Teilschritte des Condensation-Algorithmus für eine eindimensionale Verteilungsfunktion. Die Größe der Ellipsen korrespondiert mit dem Funktionswert der Verteilungsfunktion und damit mit der Gewichtung der entsprechenden Stützstelle.

Ziel des Verfahrens ist es, eine unbekannte, möglicherweise multimodale Verteilungsfunktion $f_t(x, y, z)$ zu approximieren, ohne diese im gesamten Definitionsbereich, d.h. an jeder Position innerhalb der Auflösungspyramide, berechnen zu müssen. Diese Verteilungsfunktion repräsentiert die Wahrscheinlichkeit, dass sich das zu verfolgende Objekt zum Zeitpunkt t an Position (x, y, z) befindet. Begonnen wird mit einer ausreichend großen Anzahl an Stützstellen, an denen der Funktionswert $f_t(x, y, z)$ berechnet wird, wobei die Dichte der Stützstellen an der Initialposition, die von der Personenverifikation vorgegeben wurde, am größten ist (Abb. 4 a)). Die deterministische Verschiebung berücksichtigt die vermutete (geschätzte) Bewegung des Objekts (Bewegungsmodell) bis zur Aufnahme des nächsten Bildes (Abb. 4 b)). Dieser wird eine stochastische Verschiebung

überlagert, die neue Stützstellen generiert, um Ungenauigkeiten des Bewegungsmodells zu kompensieren (Abb. 4 c)). Im finalen Schritt (Abb. 4 b)) werden die Funktionswerte $f_{t+1}(x, y, z)$ an allen generierten Stützstellen neu berechnet und entsprechend ihres Funktionswertes bewichtet. Stützstellen, die einen vorgegebenen Wichtungswert unterschreiten, werden gelöscht und in der Umgebung von Stützstellen mit hohem Gewichtswert oder zufällig neu eingefügt. Die zufällige Initialisierung neuer Stützstellen sichert, dass auch multimodale Verteilungsfunktionen approximiert werden können.

Die zum Tracking verwendeten Merkmale (Kontur, einfaches Farbmodell) wurden aus der Personenverifikation abgeleitet. Da die Personenverifikation kontinuierlich weiter berechnet wird, können deren Ausgaben in den Trackingprozess eingekoppelt werden, was zu einer enormen Erhöhung der Robustheit führt. Weiterhin erfolgt zur Verifikation des Trackingprozesses eine einfache sonarbasierte Auswertung der Entfernungsinformation in Richtung des aktuell verfolgten Objekts.

2.4 Sprachausgabe und grafisches Interface

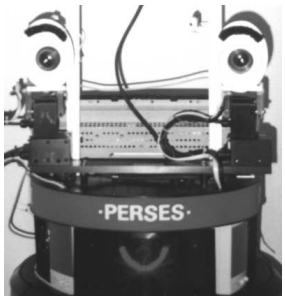

Abb. 5. Gesicht des PERSES-Roboters.

Für die unmittelbare MMI finden sowohl eine Sprachausgabe als auch ein grafisches Benutzerinterface Verwendung. Mittels Sprache ist der Roboter in der Lage, sowohl seine eigentlichen Serviceleistungen zu offerieren als auch dem Interaktionspartner seinen aktuellen Status intuitiv mitzuteilen. Ein grafisches Nutzerinterface ist notwendig, um dem Benutzer die Auswahl verschiedener Serviceleistungen auf einfache Weise zu ermöglichen. Inspiriert vom smarten „Gesicht“ des MINERVA-Roboters [8], wurde PERSES ebenfalls mit einem Gesicht (siehe Abb. 5) versehen. Durch die augenähnliche Gestaltung der Kamerafronten und einen ansteuerbaren Mund kann der „emotionale“ Zustand des Roboters intuitiv besser verdeutlicht werden.

3 Experimentelle Ergebnisse

Die Interaktion startet mit der in Abb. 6 dargestellten Bewegungsanalyse im omnidirektionalen Videodatenstrom. Person P2 bewegt sich auf den Roboter zu, Person P1 geht am Roboter vorbei (oben links). Beide Personen werden detektiert (oben rechts), und nach der Analyse der Bewegungsrichtung beider Personen (unten) wendet sich der Roboter Person P2 zu.

Einige exemplarische Ergebnisse zur sich daran anschließenden Personenverifikation zeigt Abb. 7. Nach erfolgreicher Verifikation begrüßt der Roboter den potentiellen Interaktionspartner mit einer typischen Sprachsequenz und bietet seine Dienste an. Anschließend hat der Kunde die Möglichkeit, über das grafische

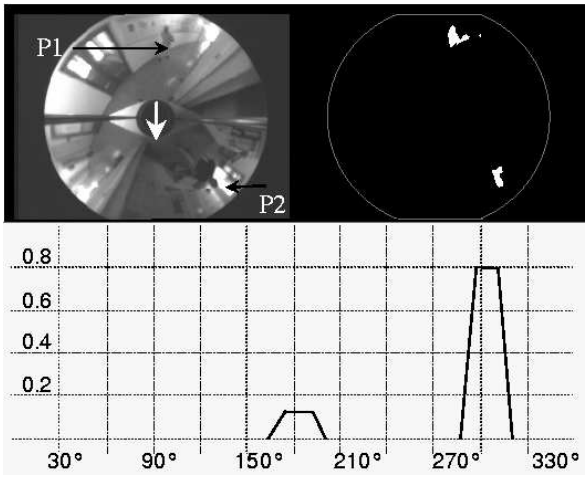

Abb. 6. Bewegungsanalyse im omnidirektionalen Videodatenstrom. Die sich aus etwa 280^o-Richtung direkt auf den Roboter zu bewegende Person erzeugt aufgrund der Auswertung der Bewegungsrichtung im Vergleich mit der in ca. 180^o-Richtung detektierten Person eine deutlich höhere Aufmerksamkeit, die im unteren Diagramm dargestellt ist. Der weiße Pfeil markiert die Orientierung des Roboters, der Winkel läuft im Uhrzeigersinn.

Abb. 7. Ergebnisse der Personenverifikation für verschiedene Situationen im Baumarkt: in Gängen zwischen den Regalreihen (1. bis 3. von links) sowie vor einem Regal, welches eine sehr schwierige Hintergrundstruktur darstellt. Hervorgehoben werden soll die hohe Spezifität des Verfahrens, die in einigen Fällen zur Nicht-Detektion einer vorhandenen Person (rechtes Beispielbild) führt. Damit soll verhindert werden, dass der Roboter aufgrund fehlerhafter Verifikation auch unbelebte Objekte anspricht.

Interface beispielsweise den gewünschten Marktbereich oder Artikel auszuwählen, zu dem der Roboter ihn dann lotst. Um sicherzustellen, dass der Kontakt zum aktuellen Kunden möglichst kontinuierlich bestehen bleibt, erfolgt ein visuelles Personentracking (siehe Abb. 8), welches durch die Personenverifikation initialisiert wird und dessen Robustheit durch die parallel dazu fortgesetzte Verfifikation enorm verbessert werden kann. So lange der Kontakt problemlos aufrecht erhalten werden kann, erfolgt keine Artikulation seitens des Roboters. Detektiert der Roboter einen drohenden Kontaktverlust, bittet er den Kunden via Sprachausgabe, einen kürzeren Abstand zu wahren, oder der Roboter unterbricht alternativ seine Fahrt zur aktuell gültigen Zielregion und versucht seinerseits, dem Kunden hinterherzufahren, um einen Kontaktverlust aktiv zu vermeiden.

4 Zusammenfassung und Ausblick

Es wurde ein Ansatz zur multimodalen Mensch-Roboter-Interaktion vorgestellt, der sich zwar stark am Szenario eines interaktiven Einkaufsassistenten orientiert,

Abb. 8. Ausschnitte aus einer mehrminütigen Trackingsequenz (die Rahmen indizieren die wahrscheinlichste Position des zu verfolgenden Objekts). Der Trackingprozess ist derzeit noch beschränkt auf Personen, die sich dem Roboter in etwa frontal zuwenden.

prinzipiell jedoch für den Einsatz verschiedenster interaktiver Serviceroboter geeignet ist. Die zukünftigen Arbeiten werden die Erweiterung des Trackingprozesses um stereobasierte Verfahren umfassen, um über einen größeren Entfernungsbereich das Kontakthalten zum Benutzer sicherstellen zu können und die derzeitige Limitierung auf in etwa frontal zugewandte Personen zu überwinden.

Literatur

1. Boehme, H.-J., Braumann, U.-D., Corradini, A., and Gross, H.-M. Person Localization & Posture Recognition for Human-Robot Interaction. In *GW'99 - The 3rd Gesture Workshop, Gif-sur-Yvette, France*, pages 105–116. Springer, Lecture Notes in Artificial Intelligence 1739, 1999.
2. Boehme, H.-J., Wilhelm, T., Key, J., Schroeter, Ch., Hempel, T., and Gross, H.-M. An Approach to Multimodal Human-Machine Interaktion for Intelligent Service Robots. In *EUROBOT'01 - the fourth Euromicro Workshop on Advanced Mobile Robots*. IEEE Computer Society Press, 2001.
3. Böhme, H.-J. and Gross, H.-M. Ein Interaktives Mobiles Service-System für den Baumarkt. In *14. Fachgespräch Autonome Mobile Systeme (AMS'99), München*, pages 344–353. Springer, 1999.
4. Fahlmann, S.E. and Lebiere, Ch. The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems 2*, pages 524–532. Morgan Kaufmann Publishers, Inc., 1990.
5. Isard, M. and Blake, A. CONDENSATION – conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.
6. Lawitzky, G. Mensch-Maschine-Interaktion bei Servicerobotik-Systemen. In *15. Fachgespräch Autonome Mobile Systeme, AMS'99*, pages 2–7. Springer Verlag, 1999.
7. Rowley, H. A., Baluja, S., and Kanade, T. Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.
8. Thrun, S., Beetz, M., Bennewitz, M., Burgard, W., Cremers, A.B., Dallaert, F., Fox, D., Hähnel, D., Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. Probabilistic algorithms and the interactive museum tour-guide robot minerva. *International Journal of Robotics Research*, 19(11):972–999, 2000.