Natürliche Mensch-Roboter Interaktion mittels Sprache, Blickrichtung und Gestik

R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, A. Waibel, Karlsruhe

Kurzfassung

Für eine natürliche Kommunikation zwischen Mensch und Roboter ist die Erkennung menschlicher Interaktionsmodalitäten wie Sprache, Gestik und Blickrichtung erforderlich. Diese müssen unter dem gegebenen Kontext gemeinsam interpretiert werden. In dieser Arbeit präsentieren wir unsere derzeitigen Forschungsarbeiten auf dem Gebiet der Mensch-Roboter Interaktion. Diese beinhalten: Spracherkennung, Dialogverarbeitung, visuelles Erkennen und Verfolgen von Personen, Erkennung von Zeigegesten und Blickrichtung, sowie die gemeinsame Fusion und Interpretation dieser Modalitäten.

1. Einleitung

Eine wesentliche Herausforderung bei der Entwicklung intelligenter, natürlicher Mensch-Roboter Schnittstellen stellt die automatische Erfassung und Interpretation des menschlichen Verhaltens dar. Dies ist insbesondere bei der Entwicklung von zukünftigen Haushaltsrobotern wichtig. Solche Roboter sollten in der Lage sein, die Handlungen und Intentionen der mit ihnen interagierenden Personen zu erkennen und deren Interaktionsmodalitäten richtig interpretieren. Diese umfassen neben der Sprache, als wohl wichtigstes Kommunikationsmittel, auch die Verwendung von Gesten, Blicken, Gesichtsausdruck, sowie emotionaler oder hyperartikulierter Sprache.

In unseren Forschungsgruppen an der Universität Karlsruhe und der Carnegie Mellon University, Pittsburgh, USA beschäftigen wir uns seit langem mit der Erkennung dieser menschlichen Interaktionsmodalitäten, sowie mit der Entwicklung multimodaler Mensch-Maschine-Schnittstellen. Im Rahmen des Sonderforschungsbereiches 588 „Humanoide Roboter" arbeiten wir nun daran, diese Technologien für die natürliche Mensch-Roboter Interaktion einzusetzen und weiterzuentwickeln.

In diesem Beitrag sollen der derzeitige Stand unserer Arbeiten sowie Perspektiven aufgezeigt werden. Es werden Komponenten für Spracherkennung, multimodale Dialogverarbeitung, visuelle Erkennung und Modellierung von Benutzern vorgestellt. Diese

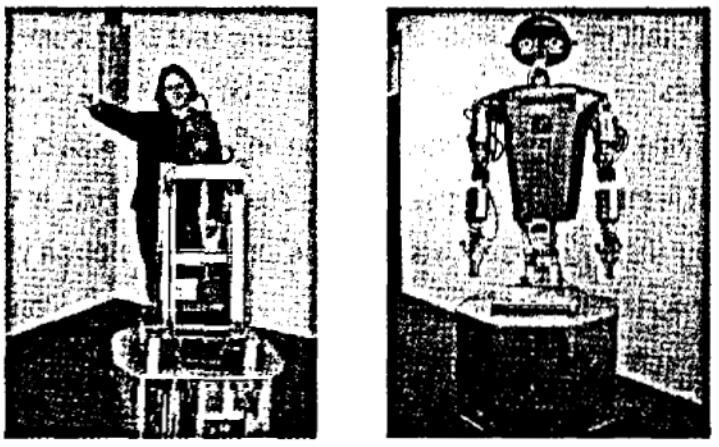
beinhalten die Erkennung von Zeigegesten und der Kopfdrehung von Personen. Die Teilkomponenten wurden auf einer mobilen Roboterplattform integriert und können für multimodale Interaktion mit dem Roboter in Echtzeit eingesetzt werden.

Als Szenario wurde dabei eine Situation in einer Küche betrachtet: Eine Person kann hier dem Roboter beispielsweise Fragen über den Inhalt des Kühlschranks stellen, ihn nach Rezeptvorschlägen befragen, oder den Roboter anweisen, den Tisch zu decken und Ähnliches.

Bisher wurden folgende Komponenten in den mobilen Roboter integriert:

- Ein Spracherkennungssystem
- 3D Verfolgung von Kopf und Händen
- Erkennung von Zeigegesten
- Erkennung der Kopfdrehung
- Dialogverarbeitung
- Sprachsynthese

Bild 1a) zeigt ein Benutzer, der mit unserem Entwicklungsroboter interagiert. Die Komponenten für das visuelle Verfolgen von Benutzern wurden inzwischen auch auf einen zweiarmigen humanoiden Roboter [3] integriert (Abbildung 1 b).



a) b)

Abbildung 1: a) Interaktion mit der Entwicklungsplattform. Folgende Komponenten wurden integriert: Spracherkennung, Sprachsynthese, Erkennung und Verfolgung von Personen und Zeigegesten, Dialogverarbeitung und Fusion von Sprache und Gestik. b) Einige Komponenten wurden bereits in einen zweiarmigen humanoiden Roboter integriert.

2. Spracherkennung

Die Sprache ist wohl das wichtigste Kommunikationsmittel des Menschen. Insofern ist es für eine möglichst natürliche Mensch-Maschine Kommunikation geradezu unerlässlich, Spontansprache zu erkennen und zu verstehen.

Zur Spracherkennung wurde der Ibis-Decoder [4] verwendet, welcher an der Universität Karlsruhe als Teil des Janus Recognition Toolkits (JRTk) [5] entwickelt wurde. Unter Verwendung von JRTk wurde ein sprecherunabhängiger Spracherkenner für spontane Mensch-Roboter-Interaktion entwickelt.

2.1.Kontextfreie Grammatiken

Der Ibis-Decoder besitzt die Möglichkeit nicht nur klassische n-gramm Sprachmodelle zu verwenden, sondern auch entlang kontextfreier Grammatiken (CFG) zu dekodieren. Dies ist speziell in kleinen Domänen, wie z.B. dem hier zugrundeliegenden Haushaltsszenario von Vorteil, da wir hierbei mit ohne jegliche statistische Informationen arbeiten können.

Ein großes Problem bei der Verwendung von CFGs in der Spracherkennung ist jedoch die Modellierung von Umgebungsgeräuschen und Spontansprache zusammen mit ihren grammatikalisch fehlerhaften Satzkonstrukten verbunden mit Hesitationen und Wort-wiederholungen. Solche Effekte, die quasi jederzeit innerhalb eines Satzes auftreten können, lassen sich kaum in einer Grammatik fassen. Aus diesem Grund verwenden wir sogenannte Filler-Wörter im Decoder zur Modellierung solcher Effekte. Sie bestehen aus speziell trainierten akustischen Modellen von bspw. Geräuschen oder Hesitationen und können prinzipiell immer zwischen zwei Terminalen eingefügt werden.

Bei der Verwendung von Spracherkennung im Zusammenhang mit Dialogsystemen, kann man sich den Dialogkontext für die Spracherkennung zunutze machen und somit den Spracherkenner verbessern. So sind z.B. am Anfang eines Dialoges einfache Antworten auf Systemfragen, wie „Ja" oder „Nein" weniger wahrscheinlich. Da Spracherkenner und Dialogsystem auf derselben linguistischen Wissensbasis aufbauen, kann der Dialogmanager den Suchraum des Spracherkenners direkt beeinflussen.

2.2.1 Experimente und Ergebnisse

Die Experimente wurden auf 360 Benutzeranfragen, d.h.15 Minuten Sprache durchgeführt, wobei der Echtzeitfaktor auf einem 800MHz PIII ermittelt wurde. Unser HMM-basiertes Spracherkennungssystem zur Mensch-Roboter Interaktion besteht aus etwa 34.000 Gaussmodellen und wurde auf ungefähr 300 Stunden Telefonkonversationen trainiert. Zur

Kompensation verschiedener Sprecher, Kanäle und Hintergrundgeräusche werden inkrementelle Adaptionsverfahren eingesetzt.

Tabelle 1 zeigt die Performanceverbesserungen, die sich durch den Einsatz der Filler-Wörter ergeben, und vergleicht n-gramm basierte mit grammatikbasierte Spracherkennung. Der Vorteil der grammatikbasierten Erkennung ist neben ihrer höheren Geschwindigkeit die sehr viel geringere Satzfehlerrate, die gerade für das Sprachverstehen sehr wichtig ist.

Tabelle 1: Grammatikbasierter und n-gramm-basierte Spracherkennung von Spontansprache

	Wortfehlerrate	Satzfehlerrate	Echtzeitfaktor
CFG	25,55%	48,21%	-,-
+ Filler-Wörter	23,05%	45,18%	0,759
n-gramm	22,95%	51,79%	0,801

2.2 Entfernte Mikrophone

Ein weit verbreitetes und bisher noch ungelöstes Problem der Spracherkennung ist die Spracherkennung unter Verwendung von Raummikrophonen bei variablem Sprecherabstand. Da es unser Ziel ist, benutzerfreundliche Roboter zu entwickeln, die quasi überall in unserem Leben zu finden sind, ist es einem Benutzer nicht zuzumuten ständig ein Mikrophon mit sich herumzutragen. Aus diesem Grund müssen Technologien entwickelt werden, die die Spracherkennung unter solchen Bedingungen verbessern.

2.2.1 Experimente und Ergebnisse

Bislang wurden nur einige Experimente durchgeführt, die die Sensibilität des eingesetzten Spracherkenners bei Verwendung eines einzigen Raummikrophons evaluieren sollten. Aufgrund der geringen Anzahl an Evaluations- und Adaptionsmaterial für die Haushaltsdomäne wurden 2 Stunden gelesene Sprache zur Adaption und 15 Minuten gelesene Sprache zur Evaluation gesammelt.

Für jeden Mikrophonabstand wurden nun die Codebücher adaptiert, ohne jedoch Sprecheradaption zu betreiben. Wie aus Tabelle 2 ersichtlich lässt sich hierdurch eine Reduktion der Wortfehlerrate um 10-15% erreichen, jedoch sind die Ergebnisse in einem normalen Handlungsradius zum Roboter von 1-3 m immer noch inakzeptabel.

Um zu analysieren, wie stark ein einmal adaptierter Spracherkenner auf Varianzen innerhalb des Entfernungsradius reagiert, wurde der auf 1,5 m adaptierte Erkenner auf den Aufnahmen aus einer Entfernung von 1,2 bzw. 1,5 m getestet. Vergleicht man die Ergebnisse von Tabelle 3 mit denen aus Tabelle 2 fällt auf, dass der Spracherkenner relativ robust

gegenüber sich bewegenden Sprechern zu sein scheint. Wir arbeiten zur Zeit an der Übertragung von Adaptionsverfahren aus dem Bereich der Navigation im Auto [7].

Tabelle 2: Wortfehlerraten eines adaptierten und unadaptierten Systems

	Nah	Lapel	1,2 m	1,5 m	1,8 m	2,4 m
Unadaptiert	26,6%	29,7%	47,7%	51,9%	66,1%	69,3%
Adaptiert	26,5%	28,4%	42,5%	44,7%	59,7%	60,1%

Tabelle 3: Wortfehlerraten eines bereits adaptierten System bei verschiedenen Distanzen

	1,2 m	1,5 m	1,8 m
Adaptiert auf 1,5 m	42,4%	44,7%	60,1%

3. Visuelle Erfassung von Personen

Information über den Aufenthaltsort, die Körperhaltung sowie den Aufmerksamkeitsfokus des Benutzers sind wichtige Schritte auf dem Weg zum Verständnis der Absicht des Benutzers im Dialog mit einem Roboter. Das hier beschriebene System kann aus den Videobildern, die von einer Stereokamera im Kopf des Roboters geliefert werden, folgende Informationen in Echtzeit gewinnen: a) die räumlichen Positionen von Kopf und Händen, b) die Kopfdrehung, und c) die Richtung von Zeigegesten, die der Benutzer ausführt.

3.1 Verfolgung von Hand- und Kopf in 3D

a) Hautfarbklassifikation b) Disparitätenbild c) Ballung

Abbildung 2: Hautfarbene Bildpunkte werden über das Disparitätenbild mit 3D-Koordinaten versehen und mittels eines k-Mittelwerte-Verfahrens räumlich geballt. Die entstehenden Ballungszentren sind mögliche Kopf-/Hand-Positionen.

Kopf und Hände können im Bild anhand ihrer Farbe gefunden werden, da sich menschliche Hautfarbe in einer eng begrenzten Region des chromatischen Farbraumes befindet [8]. Im vorliegenden System wird die Hautfarbverteilung durch Histogramme modelliert. Aufgrund der Mobilität des Roboters ändert sich die Beleuchtungssituation häufig. Daher ist es notwendig, das Hautfarbmodell automatisch zu initialisieren und fortlaufend zu adaptieren. Zu diesem Zweck wird das aus der Stereobildverarbeitung gewonnene und vergleichsweise beleuchtungsinvariante Disparitätenbild (Abb. 2.b) herangezogen und mit einem mit [9] vergleichbaren Verfahren nach einem Kopf durchsucht. Bei einem positiven Ergebnis werden alle Bildpunkte innerhalb der Kopfregion dem Hautfarbmodell zugeführt.
Die Aufgabe des *Trackings* besteht darin, die beste Hypothese s_t bezüglich die Position von Kopf und Händen zu einem jeden Zeitpunkt t zu finden. Die Entscheidung basiert auf der aktuellen Beobachtung O_t (entspricht der Ballung der Hautfarbpixel in Abb. 2.c) sowie den zeitlich vorangegangenen Hypothesen s_{t-1}, s_{t-2}, ... Mit jedem neuen Bild werden alle möglichen Kombinationen der Ballungszentren ausgewertet, um diejenige Hypothese s_t zu finden, die im Hinblick auf die drei folgenden Maße die höchste Bewertung erzielt:

- Die Beobachtungswahrscheinlichkeit $P(O_t \mid s_t)$. Sie erhöht sich mit jedem Hautfarbpixel, welches mit der Hypothese übereinstimmt.
- Die a-priori Wahrscheinlichkeit $P(s_t)$ drückt aus, wie wahrscheinlich die durch s_t repräsentierte Körperhaltung ist.
- Die Übergangswahrscheinlichkeit $P(s_t \mid s_{t-1}, s_{t-2}, ...)$ drückt aus, wie gut s_t dem Bewegungspfad folgt, der durch die vergangenen Hypothesen gegeben ist.

In unseren Versuchen hat sich gezeigt, dass mit dem hier beschriebenen Verfahren eine Person auch dann robust verfolgt werden kann, wenn sich die Kamera bewegt und der Bildhintergrund unruhig ist.

3.2 Bestimmung der Kopfdrehung

Zur Schätzung der Kopfdrehung wird ein bildbasiertes Verfahren verwendet: Zwei neuronale Netze (das eine für den horizontalen, das andere für den vertikalen Drehwinkel) verarbeiten zugleich das Intensitäts- als auch das Disparitätenbild des Kopfes, der auf eine feste Größe von 24x32 Pixeln skaliert wurde. Da die Drehwinkel des Kopfes direkt aus jedem Bild gewonnen werden, ist es nicht notwendig, die initiale Kopfdrehung zu kennen.
Die Netze sind dreischichtig und bestehen jeweils aus 1597 Neuronen. Sie wurden mit Beispielbildern gedrehter Köpfe von 6 verschiedenen Personen trainiert, die sich in einer

Entfernung von ca. 2-3m frei im Sichtfeld der Kamera bewegen durften. Die tatsächliche Kopfdrehung wurde mit einem Magnetsensor aufgenommen. Tabelle 4 zeigt die Ergebnisse zweier Auswertungen: Das Mehrpersonensystem wurde auf allen 6 Benutzern trainiert und mit anderen Bildern derselben Benutzer ausgewertet. Im Fall „unbekannte Person" wurden die Netze mit den Bildern von 5 Personen trainiert und auf der 6. Person ausgewertet. Wie gut zu erkennen ist, konnte durch das Hinzufügen des vergleichsweise beleuchtungsinvarianten Disparitätenbildes die Qualität der Schätzung signifikant verbessert werden.

Tabelle 4: Durchschnittlicher Fehler (horizontal/vertikal) bei der Kopfdrehungsschätzung

	Mehrpersonensystem	Unbekannte Person
Intensität	4,6° / 2,4°	15,5° / 6,3°
Disparität	8,0° / 3,3°	11,0° / 5,7°
Intensität + Disparität	**4,3° / 2,1°**	**9,7° / 5,6°**

3.3 Erkennen von Zeigegesten

In unserem Mensch-Roboter Interaktionsszenario verstehen wir unter einer Zeigegeste eine Bewegung der Hand in Richtung eines Ziels. Um Zeigegesten von anderen natürlichen Handbewegungen abgrenzen zu können, wird das typische Bewegungsmuster der Hand in drei Phasen zerlegt, die jeweils durch ein dediziertes Hidden-Markov-Modell repräsentiert werden: in der Beginn-Phase bewegt sich die Hand in Richtung des Ziels, in der Halte-Phase verweilt die Hand bewegungslos (häufig sehr kurz) und in der Ende-Phase kehrt die Hand auf eine (beliebige) Ruheposition zurück. Als Merkmale für die Modelle dienen die 3D-Trajektorien der zeigenden Hand.

In früheren Experimenten [11] wurde deutlich, dass Menschen üblicherweise auch auf das Ziel blicken, auf das sie gerade zeigen. Um diesen Sachverhalt auszunutzen, werden zusätzlich zur Position der Hand auch die Drehwinkel des Kopfes als Eingabe für die Modelle verwendet.

In einer Auswertung mit 12 Versuchspersonen erkannte das System ca. 80% aller ausgeführten Gesten (recall) mit einer Zuverlässigkeit von 74% (precision). Durch die Hinzunahme der Kopfdrehung verbesserte sich die Zuverlässigkeit signifikant: die Anzahl fälschlich erkannter Gesten (false positives) verringerte sich von ca. 26% auf 13% bei gleich bleibender Erkennungsrate. In unseren Experimenten hat sich die räumliche Linie ausgehend vom Kopfmittelpunkt durch den Mittelpunkt der zeigenden Hand als zuverlässiger Schätzer für die Zeigerichtung erwiesen. Mit einem durchschnittlichen Fehler von 20° war es in den meisten Fällen möglich, unter 8 im Raum verteilten Zielen das richtige auszuwählen.

4. Multimodale Dialogverarbeitung

Die Ausgabe des Spracherkenners und die des Gestenerkenners werden an den multimodalen Dialogmanager weitergegeben, wo der eigentliche Verstehensprozess stattfindet. Hier wird ermittelt, was der Roboter tun soll. Momentan kann der Roboter dem Benutzer in der Küche zur Hand gehen: Er kann Geschirr oder Besteck holen, es irgendwohin bringen, etwas zu trinken oder zu essen holen, Tee oder Kaffee machen, das Licht an- bzw. ausschalten, im Kühlschrank nachschauen, dem Benutzer Rezepte mitteilen, usw. Dafür werden die Resultate der Spracherkennung und der Gestenerkennung an den Dialogmanager geschickt, der beides im Diskurskontext auswertet. Die multimodale Fusion basiert auf der Semantik der beiden Eingabemodalitäten [12].

4.1 Dialogmanagement

Unser Dialogmanager basiert auf den Ansätzen des sprach- und domänenunabhängigen Dialogmanagers ARIADNE [13]. Für den domänenabhängigen Teil haben wir die folgenden Ressourcen entwickelt: Eine Ontologie, eine Spezifikation der Dialogziele, eine Datenbank mit dem Umweltmodell, eine kontextfreie Grammatik und Vorlagen zur Generierung der Sprachausgabe.

Die Spracheingabe wird geparst mit Hilfe einer kontextfreien Grammatik, die angereichert ist mit Informationen aus der Ontologie, in der alle Aufgaben, Objekte und deren Eigenschaften, über die der Benutzer sprechen kann, definiert sind. Die semantische Repräsentation, die im Parsing aufgebaut wird, wird dann mit den Dialogzielen verglichen. Wenn bereits alle Informationen, um ein Dialogziel auszuführen, vom Benutzer gegeben wurden, wird der entsprechende Dialogservice aufgerufen. Wenn allerdings noch Informationen fehlen, so werden Klärungsfragen vom Dialogmanager gestellt. Dafür werden so genannte Generierungsschablonen benutzt, die dafür verantwortlich sind, die sprachliche Ausgabe zu erstellen.

Die Gesteneingabe wird mit Hilfe von einem Umweltmodell aufgelöst. Die hereinkommenden Zeigegesten werden mit den Positionen der Objekte im Umweltmodell verglichen. Alle Zeigegesten, die auf ein potentielles Zielobjekt deuten, werden in eine N-besten Liste von Zeigehypothesen mit ihrer semantischen Repräsentation aufgenommen. Diese Hypothesen werden dann im sprachlichen Kontext benutzt, um zwischen verschiedenen Sprachhypothesen disambiguieren zu können. Dabei erfolgt die Disambiguierung durch Verschmelzung von Sprache und Gestik im multimodalen Parsingprozess.

4.2 Multimodales Parsing

Es wird ein Constraint-basierter Ansatz benutzt, um Sprache und Gestik zu fusionieren. Die Parserregeln definieren zeit-, kontext- und inhaltsbasierte Constraints, als auch Fusionsanweisungen. Der Multimodale Parser ist in das Dialogsystem integriert und wird auf semantischer Ebene auf das Resultat aus Sprach- und Gestikparser angewandt. Die ursprünglichen Ereignisse, aus denen die semantischen Tokens durch Sprach- und Gestikparser erzeugt werden, definieren Zeitstempel, nach denen die semantischen Tokens partiell geordnet werden können.

Zur Disambiguierung von Spracheingaben, werden Gestikereignisse den zugehörigen Spracheingaben zugeordnet. Der Begriff Disambiguierung umfasst u.a. Deixis, Spracherkennungsfehler, N-besten Listen als Resultat der Auflösung von Zeigegesten im Umweltmodell, N-besten Listen als Ausgabe des Spracherkenners, die mit Gestik kombiniert werden

Wir betrachten die Spracheingabe als Hauptkomponente und verwenden Gestik zur Disambiguierung. Daher ist es wichtig, möglichst alle tatsächlich aufgetretenen Gesten zu erkennen (Recall), während falsche Detektionen eher toleriert werden können. Die meisten falschen Detektionen können durch fehlende zeitliche Korrelation zu einer Spracheingabe aussortiert werden. Weitere Constraints basieren auf Inhaltlichen Informationen. So müssen z.B. Objekte, die durch Gestik referenziert werden der Subkategorierung der Spracheingabe übereinstimmen. Betrachtet man z.B. die Eingabe "bitte bring mir diese Tasse", kommen nur Objekte in Frage, die (i) von dem Roboter getragen werden können, (ii) vom Typ Tasse sind. Die Objekteigenschaften werden hierzu aus einer Ontologie entnommen.

5. Zusammenfassung und Ausblick

In der vorliegenden Arbeit haben wir einige unserer derzeitigen Forschungsarbeiten auf dem Gebiet der Mensch-Roboter Interaktion präsentiert. Es wurden Systeme zur Spracherkennung, zur Dialogverarbeitung und zur visuellen Erfassung und Modellierung von Benutzern vorgestellt. Die beschriebenen Komponenten wurden auf einer mobilen Roboterplattform integriert und für die Mensch-Roboter Interaktion in einem Küchenszenario genutzt.

Im Rahmen des Sonderforschungsbereiches 588 „Humanoide Roboter" arbeiten wir nun daran, die Robustheit der einzelnen Systemkomponenten zu verbessern. Desweiteren wurde bereits damit begonnen, wichtige neue Komponenten für die Mensch-Roboter Interaktion, wie beispielsweise audio-visuelle Personenerkennung, zu entwickeln. Schließlich sollen

diese Komponenten in eine neue humanoide Roboterplattform mit zwei Armen integriert werden.

Danksagung

Die vorgestellten Forschungsarbeiten wurden durch die Deutsche Forschungsgemeinschaft *im Rahmen des Sonderforschungsbereiches 588 „Humanoide Roboter" unterstützt.*

Literatur

[1] Proceedings of the Third IEEE International Conference on Humanoid Robots - Humanoids 2003. Karlsruhe, Germany: IEEE, 2003.

[2] Special Issue on Human-Friendly Robots. Journal of the Robotics, Society of Japan, 1998, vol. 16, no. 3.

[3] T. Asfour, A. Ude, K.Berns, and R. Dillmann, "Control of armar for the realization of anthropomorphic motion patterns", HUMANOIDS 2001, Tokyo, Japan

[4] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," Proc. Of ASRU, Italy, December 2001.

[5] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe-VERBMOBIL Speech Recognition Engine," ICASSP, Munich, Germany, 1997.

[6] E. Seemann, K. Nickel, R. Stiefelhagen, „Head Pose Estimation Using Stereo Vision For Human-Robot Interaction", Int. Conf. On Face & Gesture Recognition, Korea, 2004.

[7] M. Westphal and A. Waibel, "Model-Combination-Based acoustic mapping," ICASSP 2001, Salt Lake City, May 2001.

[8] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaption,", Carnegie Mellon University, School of Computer Science, Tech. Rep. CMU-CS-97-146, 1997.

[9] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," CVPR , Santa Barbara, CA, 1998.

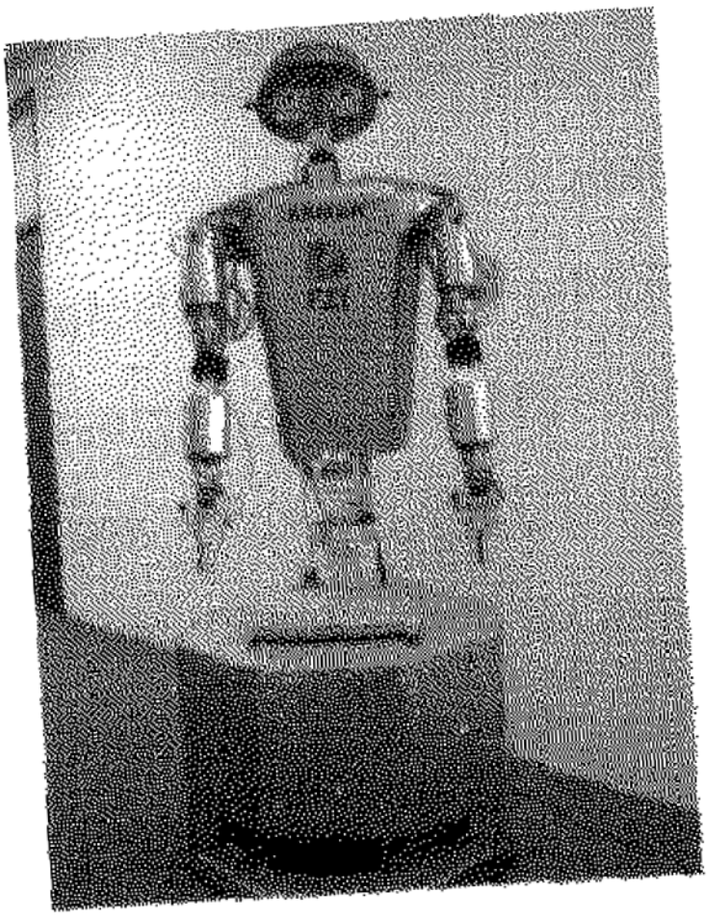
[10] R. Stiefelhagen, J. Yang, and A. Waibel, "Tracking focus of attention for human-robot communication," Int. Conf. on Humanoid Robots - Humanoids 2001, 2001.

[11] K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation," Int. Conf. on Multimodal Interfaces, Canada, 2003.

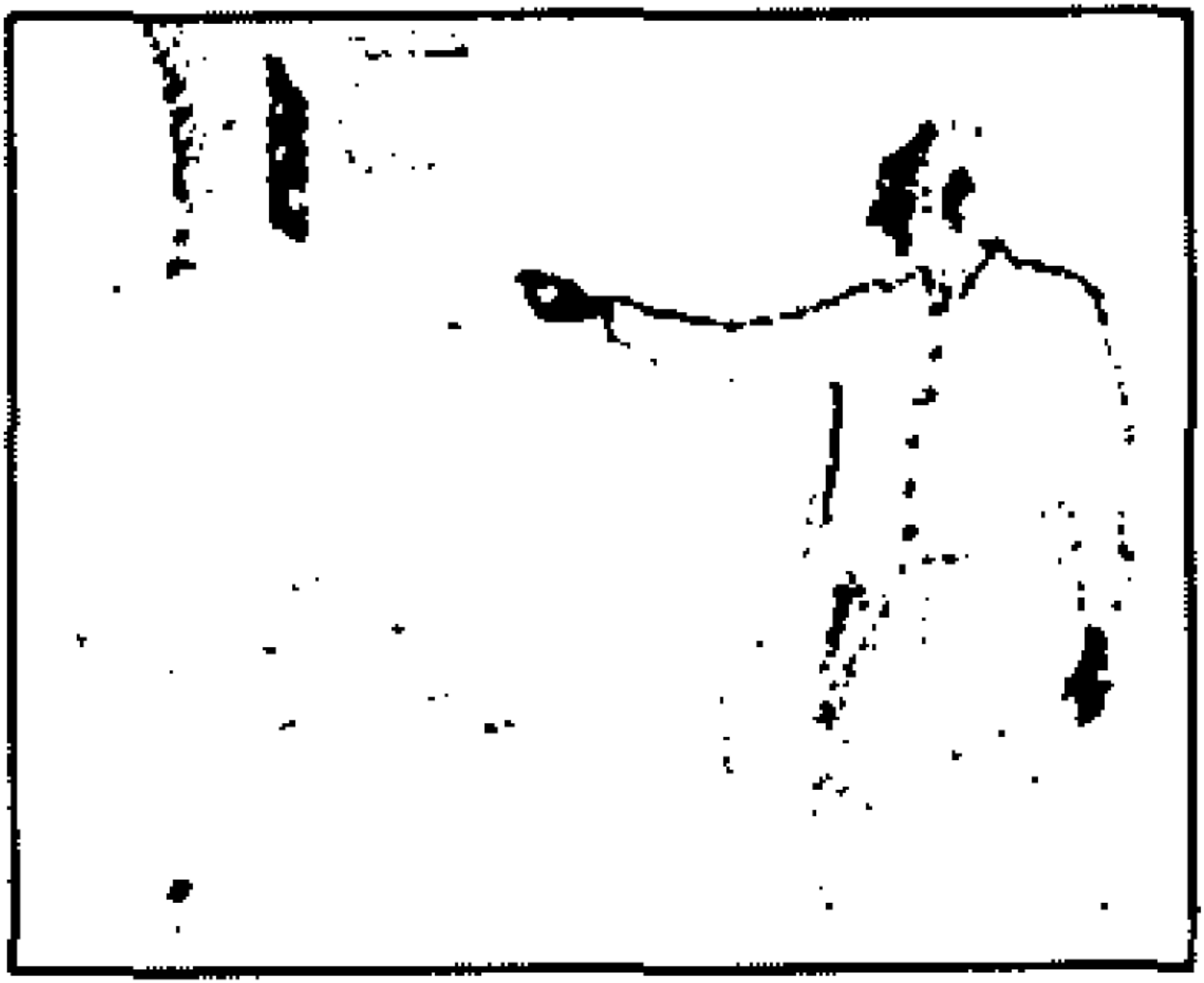
[12] P. Gieselmann and M. Denecke, "Towards multimodal interaction with an intelligent room," in Proceedings of Eurospeech, Geneva, 2003.

[13] M. Denecke, "Rapid prototyping for spoken dialogue systems," Proceedings of the 19th International Conference on Computational Linguistics, 2002.

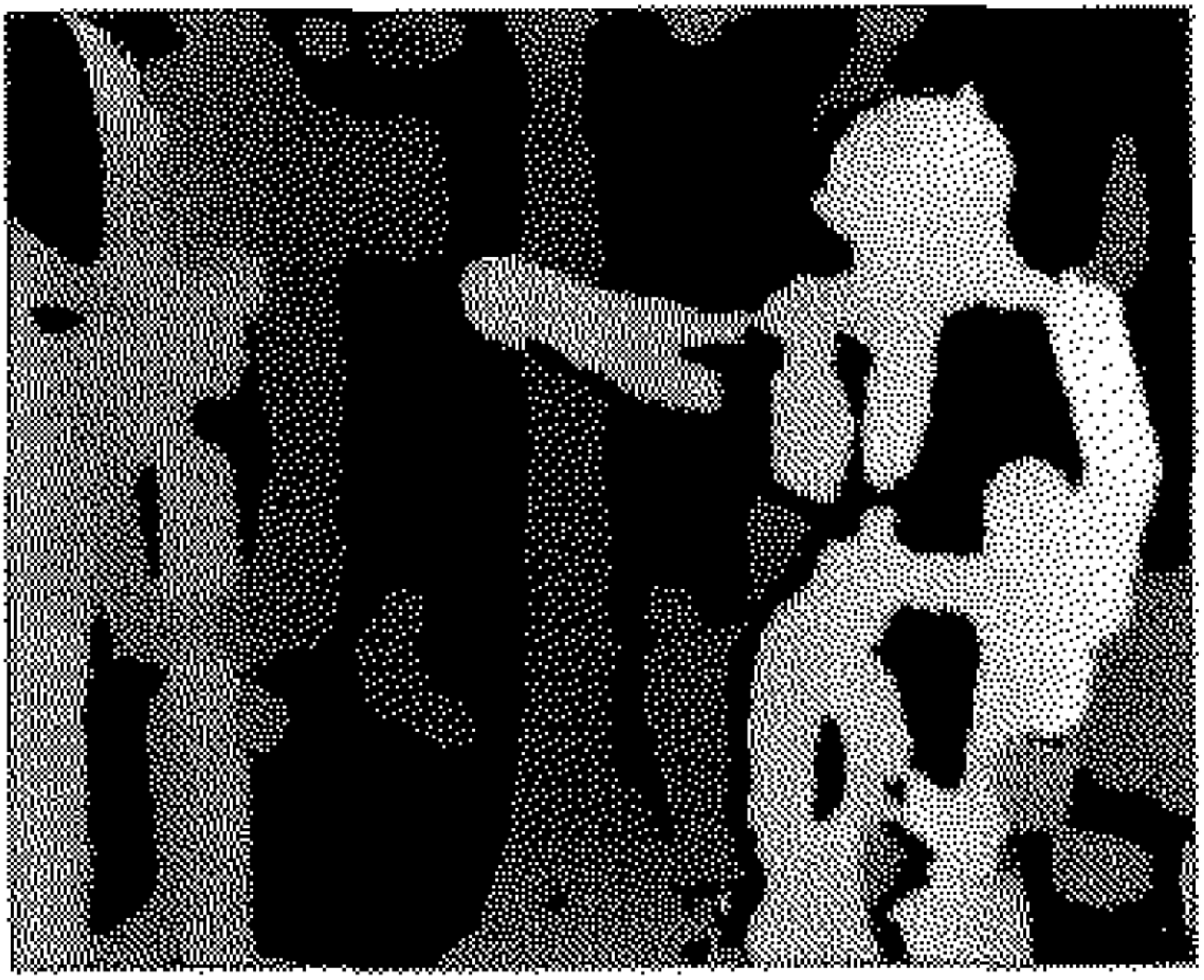
a)



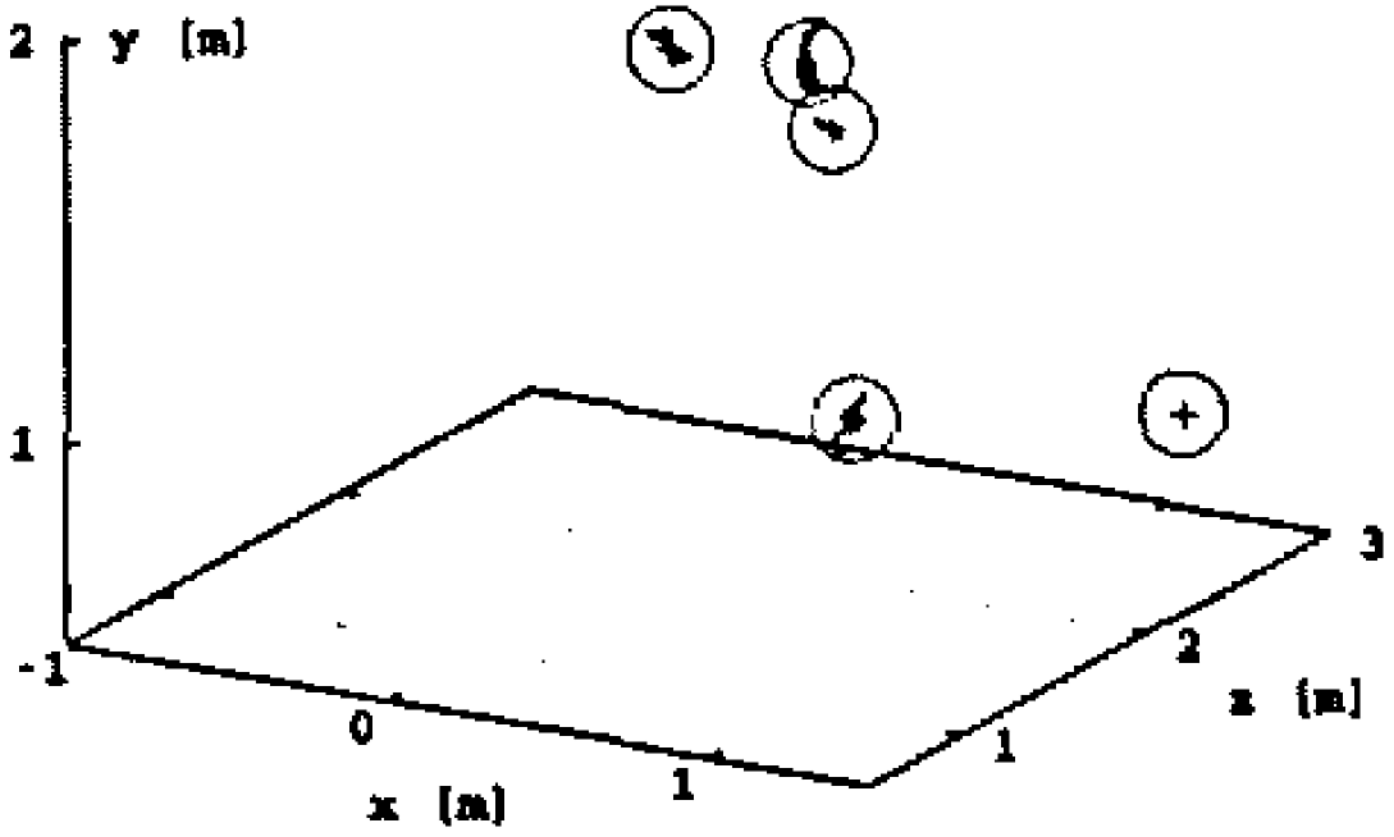
b)



a) Hautfarbklassifikatio



b) Disparitätenbild



c) Ballung