Multimodale Interaktion auf einer sozialen Roboterplattform

Multimodal Interaction on a Social Robotic Platform

Jürgen Blume*, Tobias Rehrl, Gerhard Rigoll, Technische Universität München

* Korrespondenzautor: blume@tum.de

Zusammenfassung Dieser Beitrag beschreibt die multimodalen Interaktionsmöglichkeiten mit der Forschungsroboterplattform ELIAS. Zunächst wird ein Überblick über die Roboterplattform sowie die entwickelten Verarbeitungskomponenten gegeben, die Einteilung dieser Komponenten erfolgt nach dem Konzept von wahrnehmenden und agierenden Modalitäten. Anschließend wird das Zusammenspiel der Komponenten in einem multimodalen Spieleszenario näher betrachtet. ►►► **Summary** This paper presents the multimodal interaction capabilities of the robotic research platform ELIAS. An overview of the robotic platform as well as the developed processing components is presented, the classification of the components follows the concept of sensing and acting modalities. Finally, the interplay between those components within a multimodal gaming scenario is described.

Schlagwörter Mensch-Roboter-Interaktion, Multimodalität, Gesten, Blick ►►► **Keywords** Human-robot interaction, multimodal, gestures, gaze

1 Einleitung

Eine intuitive und natürliche Bedienbarkeit der zunehmend komplexeren Technik wird für den Menschen immer wichtiger, da im heutigen Alltag eine Vielzahl an technischen Geräten mit wachsendem Funktionsumfang anzutreffen ist. Unterschiedliche Aktivitäten in der Forschungsgemeinschaft haben sich schon seit längerer Zeit mit verbalen sowie nonverbalen Kommunikationsformen (bspw. Emotions- und Gestenerkennung) in der Mensch-Maschine-Interaktion beschäftigt. Gerade in der jüngeren Zeit trugen auf diesem Forschungsfeld unterschiedliche Innovationen (bspw. Touchscreen, Gestensteuerung im Fernseher) dazu bei, dass intuitive und natürliche Bedienkonzepte mehr und mehr im Alltag Verwendung finden. Auch Möglichkeiten zur Sprach- und Gestensteuerung von Konsolen und Mobiltelefonen finden heute vermehrten Einsatz in der Gerätebedienung. Diese natürlicheren und multimodalen Benutzerschnittstellen sind dem Nutzer schnell zugänglich und erlauben eine intuitivere Interaktion mit komplexen technischen Geräten.

Auch für Robotersysteme bietet sich eine multimodale Interaktion an, um die Benutzung und den Zugang zu den Funktionalitäten zu vereinfachen. Der Mensch soll in seiner Kommunikation idealerweise vollkommene Entscheidungsfreiheit bei der Wahl der Modalitäten haben, um sein gewünschtes Ziel zu erreichen. Dafür werden in diesem Beitrag die wahrnehmenden und agierenden Modalitäten einer, rein auf Kommunikationsaspekte reduzierten, Forschungsroboterplattform beispielhaft in einer Spieleanwendung untersucht.

1.1 Struktur des Beitrags

In diesem Beitrag wird zunächst ein kurzer Überblick über die multimodale Interaktion im Allgemeinen gegeben, hierbei erfolgt eine Betrachtung nach wahrnehmenden und agierenden Modalitäten. Im nächsten Abschnitt werden Arbeiten vorgestellt, die sich auch mit multimodalen Robotersystemen beschäftigen. Im darauf folgenden Abschnitt wird die Roboterplattform ELIAS mit den wahrnehmenden, verarbeitenden und agierenden

Komponenten vorgestellt. Im fünften Abschnitt werden die entwickelten Komponenten zur Wahrnehmung bzw. Interaktion näher erläutert. Die beispielhafte Spieleanwendung sowie die Beteiligung der Wahrnehmungs- bzw. Interaktionskomponenten wird im sechsten Abschnitt präsentiert. Der Beitrag schließt mit einer kurzen Zusammenfassung.

2 Multimodale Interaktion

Zur Kommunikation nutzen Menschen eine Mischung von audio-visuellen Signalen, um anhand von Sprache, Gesten, Blick, Mimik und Körperhaltung (Pose) ihre Stimmungen, Emotionen und Zuwendungen auszudrücken [1]. Eine Übersicht über diese Kommunikationskanäle ist auch in Bild 1 zu sehen.

Die Modalitäten des Menschen lassen sich nach [2] in zwei unterschiedliche Kategorien einteilen: Modalitäten, mit denen der Mensch seine Umgebung wahrnimmt, dazu zählen Sehen, Hören, Berühren, Riechen und Schmecken. Die zweite Kategorie beschreibt Modalitäten, mit denen der Mensch mit seiner Umgebung interagieren kann, hierzu zählen die Stimme, der Blick, die Mimik sowie alle Arten von Bewegungen.

Die Mensch-Mensch-Interaktion bildet die Zielvorstellung für eine natürliche und intuitive Mensch-Maschine-Interaktion. Deswegen soll so eine Maschine (in diesem Fall ein Roboter) in der Lage sein, die Gesamtheit der menschlichen Interaktion zu erfassen sowie auf multimodale Art und Weise natürlich mit dem Menschen über verschieden Kanäle zu kommunizieren.

Aufgrund dieser Anforderung muss eine Maschine sowohl eine Vielzahl der menschlichen agierenden Modalitäten wahrnehmen können, als auch selbst unterschiedliche Modalitäten zur Interaktion anbieten können.

Einige wahrnehmende Modalitäten von Maschinen versuchen die Wahrnehmung des Menschen nachzubilden, beispielsweise visuell mithilfe von Kameras sowie auditiv mithilfe von Mikrofonen. Zudem können Maschinen auch ihre Umwelt haptisch sowie olfaktorisch erfassen. Weiterhin gibt es aber auch noch einige speziell für die Maschinenbedienung entwickelte wahrnehmende Eingabegeräte: z. B. Tastatur, Maus und Stifteingabe [1].

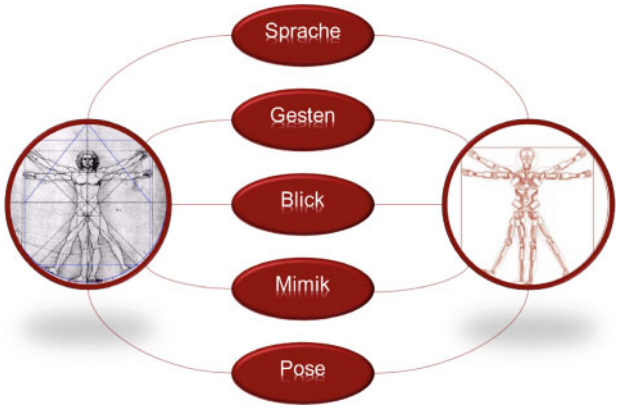

Bild 1 Kommunikationskanäle bei der Mensch-Roboter-Interaktion.

Zusätzlich zum Umfang der beteiligten Modalitäten spielt auch die Art und Weise der Verarbeitung selbst eine entscheidende Rolle bei der Ausgestaltung der Mensch-Maschine-Interaktion. Zunächst können drei wesentliche Verarbeitungsebenen für die Modalitäten unterschieden werden: Datenebene, Merkmalsebene sowie Entscheidungsebene [2]. Daneben spielt aber auch die Tatsache, dass der Mensch seine Modaliäten zur Kommunikation meist komplementär verwendet [3] eine Rolle bei der Gestaltung von Mensch-Maschine-Interaktionen.

3 Verwandte Arbeiten im Bereich der Mensch-Roboter-Interaktion

Eine multimodale Bedienbarkeit besonders im Bereich von komplexen technischen Systemen und Geräten wie Robotern kann dem Nutzer eine natürlichere Interaktion bieten. Deshalb arbeiten Wissenschaftler und Unternehmen auf der ganzen Welt an Robotern zu diesem Themengebiet. Im Folgenden seien einige dieser Entwicklungen kurz aufgezeigt. Der Roboter ARMAR III [4] aus Karlsruhe soll beispielsweise in der Küche helfen und kann seine menschlichen Partner in diesem Umfeld am Gesicht erkennen sowie diese auch verfolgen. Zur Interaktion erkennt das System neben Sprache auch Zeigegesten seines Kommunikationspartners. Auch der Fraunhofer Care-o-Bot[1] bietet in der dritten Version eine multimodale Benutzerschnittstelle zur Interaktion mit der Serviceplattform. Diese verfügt über einen ausklappbaren berührungssensitiven Bildschirm, des Weiteren können Sprache und Gesten erkannt werden, zusätzlich gibt es noch eine weitere Schnittstelle für die Anbindung mobiler Endgeräte. Die Forschungsplattform Maggie [5] verfügt über multimodale Eingabekanäle, mit denen Berührungen, Sprache, Gesichtsausdrücke und Körperbewegungen wahrgenommen werden können. Barthoc [6] aus Bielefeld ist ein antropomorpher Roboter, welcher eine am Menschen orientierte und natürliche Kommunikation ermöglicht und ebenfalls Sprache, Emotionen und Gesten, verstehen kann. Eine weitere herausragende Forschungsplattform ist der Roboter Justin des DLR [7]. Der Oberkörper ist menschenähnlich konstruiert und auf einer mobilen Plattform montiert. Dank der im Kopf verbauten Sensorik kann er Objekte im Raum verfolgen sowie greifen und des Weiteren Kommandos per Sprache empfangen. Der Fokus dieser Forschungsplattform liegt auf den Manipulationsfähigkeiten für Servicerobote. Justin war beispielsweise in der Lage, Eistee zuzubereiten und zu servieren.

Auch im Bereich von kommerziell verfügbaren Robotern für den Einsatz zu Hause – meist als Unterhaltungsgerät konzipiert – gibt es schon multimodale Bedienschnittstellen. Der Roboterhund Aibo [8] von

[1] http://www.care-o-bot.de/

Sony war ursprünglich als Haustier für Allergiker konzipiert worden, wurde dann aber zunehmend als Spiel- und Forschungsobjekt wahrgenommen. Hierzu trug sicherlich auch die Möglichkeit bei, den Roboter selbst zu programmieren. Der Roboter verfügt über Kameras, Mikrofone und taktile Eingabemöglichkeiten, während für die Ausgabe Lautsprecher und LEDs zur Stimmungsanzeige verwendet werden können.

NAO[2] ist ein kleiner humanoider Roboter von Aldebaran-Robotics und verfügt über zwei Kameras zur Gesichts- und Objektdetektion sowie vier Mikrofone zur Spracherkennung und Audioquellenlokalisierung. Des Weiteren kann über zwei Lautsprecher in acht Sprachen Text synthetisiert werden. Zusätzlich verfügt NAO auch über farbige LEDs, um Rückmeldungen an den Nutzer zu geben. Auch taktile Sensoren können benutzt werden, um mit dem Roboter zu interagieren.

4 Roboterplattform ELIAS

In diesem Abschnitt werden die verwendete Roboterplattform ELIAS (Enhanced LIving ASsistant) sowie die wahrnehmenden und agierenden Modalitäten näher beschrieben. Aufgrund der modularen Bauweise, der einfachen Erweiterbarkeit und der für Forschungsplattformen hohen Robustheit wurde als Basis des Roboters ELIAS die kommerzielle Plattform SCITOS G5 der Firma MetraLabs GmbH[3] verwendet. Die Plattform wurde speziell für den Einsatz in der Nähe von Menschen entworfen und ist dafür auch vom TÜV zertifiziert worden. Zusätzlich umgibt eine Sicherheitsleiste in Bodennähe die Plattform, um bei unerwartetem Kontakt sofort stehen bleiben zu können. Der differentielle Antrieb erlaubt Geschwindigkeiten bis zu 1,4 m/s. Dabei wiegt die Basis 60 kg und kann weitere 50 kg Last transportieren. Der Antrieb schafft eine komplette Umdrehung der Plattform in einer Sekunde und kann Steigungen von bis zu 10° überwinden. Der Industrie-PC ist ein Intel Core Duo 2,0 GHz Prozessor und hat standardmäßig das Fedora Betriebssystem installiert. Auf die Basis wurde die Mensch-Maschine-Schnittstelle montiert. Zu dieser Schnittstelle zählen ein berührungsempfindlicher Monitor sowie ein Roboterkopf mit beweglichen Augen, wie auch in Bild 2 zu sehen ist.

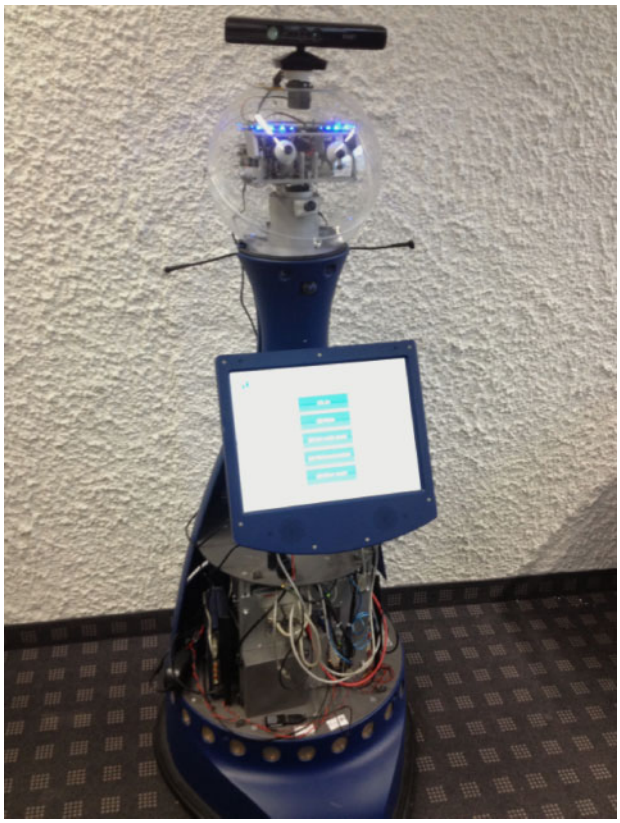

Bild 2 Hardwareübersicht der Roboterplattform ELIAS.

Für mehr Rechenkapazität wurde ein Mac Mini an der Plattform angebracht. Je nach Auslastung und Bewegung des Roboters können bis zu 8 Stunden Betriebszeit erreicht werden, bevor die Batterie wieder geladen werden muss. Die für die Mensch-Maschine-Interaktion relevanten Komponenten der Roboterplattform können nach dem Konzept von wahrnehmenden und agierenden Modalitäten wie folgt eingeteilt werden.

4.1 Wahrnehmende Modalitäten

Die wahrnehmenden Modalitäten der Roboterplattform ELIAS sind:

- Mikrofone: Für die Lokalisierung der Audioquelle wurden zwei AKG-Mikrofone in der Nähe des Kopfes angebracht.
- Kameras: Auf dem Roboter sind Kameras mit Fischaugenobjektiven auf der Vorder- und Rückseite angebracht, um die Umgebung zu erfassen. Weiterhin ist eine Kamera am Kopf angebracht sowie eine auf den Boden gerichtete Kamera. Für die Personenhypothese und die Gestenerkennung wurde ein Kinect Sensor von Microsoft auf der Plattform montiert.
- Sonarsensoren: Der Roboter verfügt in der Basis der Plattform über 24 Sonarsensoren, welche einen Bereich von 20 cm bis zu 3 Meter abdecken können.
- Laserscanner: Der Sick-Laserscanner ist in Fahrtrichtung der Plattform ausgerichtet und dient zur Navigation und Selbstlokalisierung der Plattform.
- Touchscreen: Der resistive Touchscreen ermöglicht die Auflösung einzelner Berührungsereignisse (*single touch display*).

4.2 Agierende Modalitäten

Die agierenden Modalitäten der Roboterplattform ELIAS sind:

- Augen: Die Augen werden von je zwei Piezo-Aktoren angetrieben, welche das menschliche Auge

[2] http://www.aldebaran-robotics.com/en/

[3] http://www.metralabs.com/

in Geschwindigkeit und Beschleunigung sogar übertreffen [9]. Dadurch können auch sehr schnellen Sakkadenbewegungen realistisch wiedergegeben werden. Durch die parallele Kinematik und die spezielle Aufhängung wurden zwei Freiheitsgrade pro Auge realisiert.

- Kopf: Der Kopf der Roboterplattform hat menschenähnliche Proportionen. Vertikal kann der Kopf von –7° bis +20° bewegt werden. In der horizontalen Achse sind 350° möglich.
- LEDs: Der Kopf verfügt weiterhin über 32 LEDs. Diese haben eine blaue Farbe und die möglichen Zustände: *aus*, *an*, *blinkend*, *laufend*. Auch die Intensität sowie die Blinkgeschwindigkeit kann gesteuert werden, um dem Nutzer unterschiedliche Systemzustände zu signalisieren.
- Bildschirm: Auf dem Bildschirm mit 15 Zoll Diagonale können Menüs und Medien mit einer nativen Auflösung von 1024 × 768 Pixel zur Information und Interaktion dargestellt werden.
- Lautsprecher: Im Bildschirmrahmen sind Stereolautsprecher für die Wiedergabe der Sprachsynthese sowie weiterer Audiosignale integriert.

5 Komponenten

Im Folgenden werden entwickelte Komponenten zur Wahrnehmung bzw. Interaktion vorgestellt. Die Roboterplattform ist in der Lage, mittels Sonarsensoren und der vom Kinect Sensor gelieferten Personenhypothesen – und anschließender Verifikation – die Anwesenheit eines Menschen zu erkennen. Für die Wahrnehmung der menschlichen Interaktion können Gesten, Mimik, Sprache sowie Bildschirmberührung erkannt werden. Die Interaktion beruht auf Blick, Sprachsynthese sowie Bildschirmdarstellungen. Eine Ablaufsteuerung ermöglicht das Zusammenspiel der entwickelten Komponenten. Für die Spracherkennung sowie -synthese werden auf der Roboterplattform kommerzielle Lösungen eingesetzt.

5.1 Personenhypothesenverifikation

Der Kinect Sensor ist auf der mobilen ELIAS-Plattform montiert, wodurch die Gefahr von fehlerhaft detektierten Personenhypothesen größer ist, als im Wohnzimmerbereich, welcher das ursprüngliche Einsatzgebiet bei der Sensorentwicklung dargestellt hat. Dies liegt daran, dass der Sensor mit einer Hintergrundsegmentierung arbeitet, welche im relativ statischen Wohnzimmerbereich sehr gut funktioniert. Auf der bewegten Plattform liefert dieser Ansatz aber aufgrund des bewegten Hintergrundes fehlerhafte Personenhypothesen, wie Bild 3 auf der rechten Seite zeigt. Deswegen wurde versucht, die gelieferten Personenhypothesen mittels einer gewichteten Summe über Winkel in der Skelettstruktur zu verifizieren.

Die Winkel zwischen den verschiedenen Körperbereichen (Kopf, Nacken, linke und rechte Schulter, Körpermitte, linke und rechte Hüfte) werden von oben nach unten durchnummeriert und sind auf der linken Seite von Bild 3 dargestellt. Nach der Durchnummierung wird anschließend ein Wert σ berechnet und mit einem Schwellwert zur Entscheidung darüber, ob die gegebene Personenhypothese verifiziert werden kann, verglichen. Für die Berechnung von σ wird folgende gewichtete Summe über die Winkel verwendet:

$$\sigma = \sum_{i=1}^{i=9} w_i b_i \tag{1}$$

wobei w_i ein empirisch ermitteltes Gewicht für den jeweiligen Winkel angibt (wie zuverlässig die Winkelinformation ist) und b_i ein Binärwert ist, der angibt, ob der Winkel innerhalb eines gewissen Winkelbereichs liegt oder nicht. Tabelle 1 zeigt eine Übersicht über die Winkel i, deren Gewicht w_i und die jeweiligen plausiblen Winkelbereiche als Grundlage für die Binärwertberechnung von b_i.

Bei Experimenten mit $N = 45$ Personenhypothesen, von denen 38 falsch und 7 korrekt waren, wurde das beste Ergebnis für einen Schwellwert von 0,4 für die gewich-

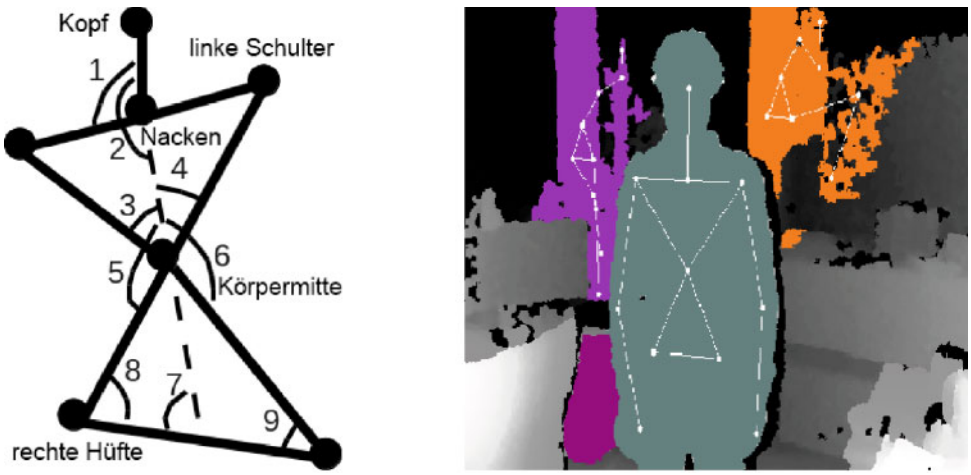

Bild 3 Skelettwinkel (links) und zwei falsche Personenhypothesen im Hintergrund (rechts).

Tabelle 1 Gewicht und Bereiche der Winkel eins bis neun in der Skelettstruktur.

Winkel-Index i	Gewicht w_i	Winkelbereich
1	0,1	$90° \pm 20°$
2	0,2	$180° \pm 20°$
3	0,1	$45° \pm 15°$
4	0,1	$45° \pm 15°$
5	0,1	$135° \pm 15°$
6	0,1	$135° \pm 15°$
7	0,2	$90° \pm 20°$
8	0,05	$90° \pm 10°$
9	0,05	$90° \pm 10°$

Tabelle 2 Konfusionsmatrix für Personenverifikation mit Schwellwert 0,4.

	P	$\overline{P}$
V	100%	2,7%
$\overline{V}$	0%	97,3%

tete Summe aus Gleichung (1) ermittelt. Die sich daraus ergebende Konfusionsmatrix ist in Tabelle 2 dargestellt.

Dabei steht P für die korrekte Personenhypothese und $\overline{P}$ für eine falsche Personenhypothese. Äquivalent steht V für die Verifikation der Personenhypothese, also $\sum_{i=1}^{i=9} w_i b_i = \sigma \geq 0{,}4$ und $\overline{V}$ für die Ablehnung der Hypothese durch den Verifizierer.

Der Verifizierer kann dadurch die korrekten Personenhypothesen bestätigen. Allerdings wurden mit diesen Einstellungen auch 2,7% der falschen Hypothesen irrtümlich als korrekt verifiziert. Zusammenfassend liefert der Verifizierer ein brauchbares Ergebnis, um falsche Hypothesen zu verwerfen. Die Personenhypothesenverifikation hilft als ein erster Vorverarbeitungsschritt für die Ausgestaltung der multimodalen Mensch-Maschine-Interaktion.

5.2 Gestenerkennung

Gesten bilden neben Mimik und Blick eine wesentliche Komponente für die nonverbale Kommunikation in der Mensch-Maschine-Interaktion. Im Gegensatz zur Emotionserkennung aus Mimik, bei der sich sechs universelle Gesichtsausdrücke zur Emotionsdarstellung unterscheiden lassen (siehe [10]), lassen sich Gesten in der Regel nur unvollständig beschreiben und erlauben auch Mehrdeutigkeiten. Gesten können unbewusst (Gestikulation) sowie bewusst (Zeichensprache) in der Kommunikation eingesetzt werden und umfassen dabei sowohl einfache als auch komplexe Bewegungen.

Die Roboterplattform ELIAS wurde mit zwei unterschiedlichen Gestenerkennungssystemen ausgestattet. Das erste System ist in der Lage, unterschiedliche statische Handgesten zu erkennen, das zweite System kann für die Spieleanwendung (vgl. Abschnitt 6) eine bestimmte

Bild 4 Übersicht über die Darstellung der Zahlen *Eins* bis *Fünf* mittels statischer Handgesten.

Menge an dynamischen Gesten lernen (weitere Details zu diesem Verfahren finden sich in [11]).

Mit den statischen Handgesten können die Zahlendarstellung von *Eins* bis *Fünf* sowie zwei weitere Gesten für die linke und rechte Hand erkannt werden. Das System benutzt Merkmale basierend auf dem Histogramm orientierter Gradienten (HOG) [12]. Die HOG-Merkmale werden, ausgehend von der ermittelten Handkontur, aus dem Tiefenbild des Kinect Sensors berechnet. Bild 4 zeigt Beispiele für solche aus dem Tiefenbild gewonnenen Handkonturen[4]. Die statischen Handgesten von *Eins* bis *Fünf* können für die Auswahl einer der fünf Antwortmöglichkeiten des Spiels von Abschnitt 6 verwendet werden. Die Gesten zum Trainieren und Testen des statischen Handgestenerkenners wurden von zehn verschiedenen Personen aufgenommen.

Zur Klassifikation der statischen Handgesten kommen *Support Vector Machines* (SVMs) [13] zum Einsatz, es sind sowohl lineare als auch polynomiale Kernel-Funktionen verwendet worden.

Im Folgenden werden die wesentlichen Schritte – Bilderfassung, Handkonturextraktion, Rotationskompensation, Merkmalsextraktion, Merkmalsreduktion, Klassifikation – für die Erkennung der statischen Handgesten vorgestellt. Bild 5 zeigt eine Übersicht über die wesentlichen Schritte.

Nach einer Bilderfassung sowie einer erfolgreichen Personenhypothesenverifikation mittels des Microsoft Kinect Sensors wird die Handkontur bestimmt. Ausgangsgrundlage für die Handkonturextraktion bildet der Handmittelpunkt $\mathbf{p}_h = (x_h, y_h, z_h)^T$. Die Bestimmung der z-Komponente erweist sich manchmal als fehlerhaft. Der Fehler tritt auf, wenn der Handmittelpunkt zwischen zwei geöffnete Finger gesetzt wird, somit bekommt die z-Komponente des Handmittelpunkts $\mathbf{p}_h$ den Wert des Hintergrunds zugewiesen, anstatt den tatsächlichen Wert der Entfernung der Hand zum Kinect. Zur Korrektur dieses möglichen Fehlers werden anhand einer 2-D Gauß-Verteilung 31 Punkte um den Handmittelpunkt in der x, y-Ebene ausgewählt. Eine anschließende Median-Berechnung über die z-Komponenten der 31 Punkte bestimmt nun die neue z_h^s-Komponente für den geglätteten Handmittelpunkt $\mathbf{p}_h^s = (x_h, y_h, z_h^s)^T$. Mithilfe der

[4] Die gezeigten Handkonturen wurden auf eine einheitliche Größe skaliert. Dieser Schritt ist notwendig für die Bestimmung der HOG-Merkmale

Bild 5 Übersicht über die wesentlichen Schritte für die statische Handgestenerkennung.

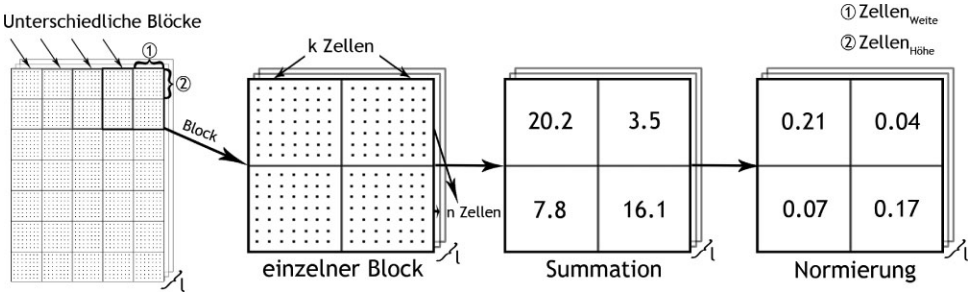

Bild 6 Übersicht über die Berechnung der HOG-Merkmale [12].

Position von $\mathbf{p}_h^s$ wird nun ein Bild mit der Handkontur extrahiert. Dieses Bild wird anschließend auf eine feste Größe von 64×128 Pixel skaliert. Morphologische Operationen auf dem zweidimensionalen Bild dienen sowohl der Rauschreduzierung als auch der Vereinfachung der Darstellung des neuen Handkonturbildes $I_c(x, y)$.

Aufgrund der Tatsache, dass die verwendeten HOG-Merkmale nicht rotationsinvariant sind, werden die Neigungsunterschiede in den Gesten zwischen den verschiedenen Personen anhand einer Rotationskompensation ausgelichen. Für die Rotationskompensation wird das Handkonturbild $I_c(x, y)$ um seinen Schwerpunkt $\mathbf{p}_{cg}$ mit dem Rotationswinkel α gedreht. Für den Schwerpunkt $\mathbf{p}_{cg}$ bzw. Rotationswinkel α ergibt sich folgender Zusammenhang mit den Momenten μ_{pq} des Bildes:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q I_c(x, y) \tag{2}$$

somit ergibt sich der Schwerpunkt p_{cg} zu

$$\mathbf{p}_{cg} = (\bar{x}, \bar{y})^T \tag{3}$$

$$\bar{x} = \frac{\sum_x \sum_y x \cdot I_c(x, y)}{\sum_x \sum_y I_c(x, y)} \tag{4}$$

$$\bar{y} = \frac{\sum_x \sum_y y \cdot I_c(x, y)}{\sum_x \sum_y I_c(x, y)} \tag{5}$$

sowie der Rotationswinkel α

$$\alpha = \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right). \tag{6}$$

Ausgehend von dem rotationsinvarianten Bild der Handkontur werden nun die HOG-Merkmale berechnet. HOG-Merkmale verwenden Gradienten, welche sich beispielsweise mittels Sobel-Filtern in x- bzw. y-Richtung über sogenannte *Zellen* berechnen lassen. Diese *Zellen* stellen eine kleine zusammenhängende Pixelmenge dar. Ein *Block* beschreibt eine Einheit, bestehend aus mehreren Zellen. *Blöcke* werden zur Kontrast-Normalisierung verwendet. In [12] sind verschiedene Verfahren zur *Block*-Normalisierung zu finden. Die notwendigen Schritte zur Bestimmung der HOG-Merkmale sind in Bild 6 dargestellt.

Prinzipiell bieten die HOG-Merkmale aufgrund der Ausgestaltung von *Zellen*, *Blöcken* und *Bins* (diskrete Intervalle in einem Histogramm) viele unterschiedliche Parameterkonfigurationen. Je nach Wahl der Parameter kann eine Merkmalsreduktion sinnvoll sein. In dem gewählten Ansatz wurden die Merkmale anhand der Hauptkomponentenanalyse reduziert. Die Parameter, welche das beste Ergebnis für die Klassifikation der Gesten erzielt haben, sind folgende:

- Zelle: 8×16 Pixel
- Block: 2×2 Zellen
- Anzahl der Bins des Histogramms: 4
- Maximum-Norm
- 150 Eigenvektoren

Mit dieser Konfiguration ließ sich eine erste Erkennungsrate von 87,20% erzielen. Es können noch weitere Modifikationen vorgenommen werden. Zunächst kann die Auswahl der Merkmalsvektoren modifiziert werden. Mit einer geordneten Liste für die größten Werte des Merkmalsvektors können unterschiedliche Anzahlen von Vektoren ausgewählt werden. Mit diesem Vorgehen kann die Erkennungsrate auf 89,73% gesteigert werden. Durch den Einsatz eines polynomialen Kernels in der SVM kann die Erkennungsrate nochmals auf 90,32% gesteigert werden.

Der vorgestellte Erkenner liefert sehr gute Ergebnisse, des Weiteren beträgt die Rechenzeit des Algorithmus 10,53 ms für den linearen Kernel und 11,68 ms für den polynomialen Kernel. Somit kann dieser Ansatz für eine echtzeitfähige Erkennung der Handgesten verwendet werden.

5.3 Mimik

Der Mensch ist in der Lage, mithilfe von Mimik unterschiedliche emotionale Zustände auszudrücken, dabei können verschiedene Kategorien zur Einteilung verwendet werden[5]. Ekman [10] identifizierte in den 1970er Jahren sechs universelle Gesichtsausdrücke, die

[5] http://changingminds.org/explanations/emotions/emotions.htm

Tabelle 3 Übersicht über die Erkennungsraten für die statischen Handgesten.

SVM-Kernel	Erkennungsrate
Linear	89,73%
Polynomial	90,32%

bestimmten emotionalen Zuständen entsprechen und unabhängig von Alter, Geschlecht sowie kulturellem Hintergrund sind. Mit der Mimikanalyse ist es möglich, auf bestimmte emotionale Zustände des Menschen schließen zu können, somit können anhand eines Mimikerkennungssystems weitere Informationen für die Mensch-Maschine-Interaktion gewonnen werden.

Es gibt unterschiedliche Datenbanken, die Mimiksequenzen für die sechs universellen Emotionsgesichtsausdrücke nach Ekman (Freude, Traurigkeit, Wut, Ekel, Angst, Überraschung) enthalten. Für das hier vorgestellte Mimikerkennungssystem wurde die Cohn-Kanade Datenbank [14; 15] verwendet, einige Bildbeispiele finden sich in Bild 7.

Der Aufbau des Mimikerkennungssystems (siehe Bild 8) hat nach [16] drei wesentliche Schritte für die automatische Mimikanalyse: Gesichtsdetektion (engl. *face detection*), Mimikdatenextraktion (engl. *facial expression data extraction*) und Mimikklassifikation (engl. *facial expression classification*).

Zur Detektion des Gesichts sowie zur Merkmalsextraktion wird der Microsoft Kinect Sensor verwendet. Die Merkmale für die Erkennung der sechs universellen Emotionsgesichtsausdrücke beruhen auf einem modellbasierten Ansatz, der das CANDIDE-3-Modell [17] zur Merkmalsextraktion verwendet. Bild 9 zeigt das CANDIDE-3-Modell.

Das CANDIDE-3-Modell besteht insgesamt aus 113 3-D-Knoten, die 168 Oberflächen bilden. Die Parameter des Modells können zwei verschiedenen Arten zugeordnet werden: Formparameter (engl. *Shape Parameters*) und Bewegungsparameter (engl. *Action Parameters*). Die einzig relevanten Merkmale für die Mimikerkennung sind die Modellparameter, die die Mimik beschreiben. Deswegen wird nur eine Teilmenge $\vec{f}_t^{\,o}$ der Modellparameter extrahiert, somit hat der Merkmalsvektor, der zur Klassifikation verwendet wird, nur elf Merkmale. Diese elf Merkmale beschreiben Aktivitäten in der Augen- sowie Mundregion, beispielsweise gibt es ein Merkmal, das die Öffnung des oberen Lids beschreibt (*upper lid raiser*). Der Merkmalsvektor ist personenspezifisch und beschreibt eine statische Mimik. Für die Mimikerkennung ist aber eine personenunabhängige Beschreibung sowie dynamische Mimikbeschreibung besser geeignet, um dies zu erreichen wird folgende neue differentielle Merkmalsvektorkonfiguration erzeugt.

$$\vec{f}_t^{\,d} = \vec{f}_t^{\,o} - \vec{f}_{t=1}^{\,o} \tag{7}$$

Der Vektor $\vec{f}_{t=1}^{\,o}$ beschreibt die neutrale Mimik der Person zu Beginn des Emotionsgesichtsausdrucks. Es ist auch möglich beide Merkmalsvektoren – $\vec{f}_t^{\,o}, \vec{f}_t^{\,d}$ – in einem Vektor $\vec{f}_t^{\,c}$ zu kombinieren. Alle drei Merkmalsvektorkonfigurationen wurden auf der Cohn-Kanade Datenbank [14; 15] getestet, hierbei wurden 331 Mimiksequenzen verwendet. Die Ergebnisse einer, auf Hidden

Bild 7 Beispielbilder unterschiedlicher Mimiken aus der Cohn-Kanade Datenbank [14; 15].

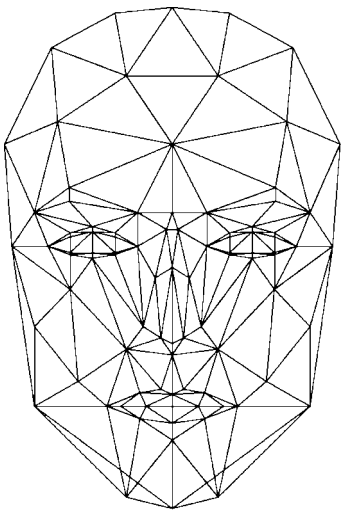

Bild 9 CANDIDE-3-Modell entnommen aus [17].

Bild 8 Übersicht über die wesentlichen Schritte für die Mimikanalyse.

- 10.1515/auto.2013.1062

Tabelle 4 Übersicht über die Erkennungsraten für die Mimikerkennung. Es wurden drei unterschiedliche Merkmalskonfigurationen verwendet. Die Daten zum Testen stammen aus der Cohn-Kanade Datenbank [14; 15].

Merkmalsvektor	Erkennungsrate
$\vec{f}_t^{\,o}$	64,24%
$\vec{f}_t^{\,d}$	76,36%
$\vec{f}_t^{\,c}$	75,15%

Markov Modellen basierten, Klassifikation sind in Tabelle 4 zu sehen.

Mit der Mimikerkennung steht neben der Gestenerkennung nun ein weiteres System auf der ELIAS-Plattform zur Verfügung, um die nonverbale Kommunikation des Menschens zu erfassen.

5.4 Blick

Für eine natürlichere Interaktion mit der Roboterplattform ELIAS wurde eine Blicksteuerung entwickelt. Die Steuerung wurde basierend auf Studien zum menschlichen Blickverhalten bezüglich Gesichtern konstruiert. Neben dieser bewussten Aufmerksamkeitsfokussierung auf den Gesprächspartner beinflussen auch äußere Reize das Blickverhalten. Um auf diese Reize außerhalb des Dialogpartners reagieren zu können, verwendet die Steuerung zusätzlich eine Kombination von visuellen und auditiven Salienzen (Auffälligkeit von Objekten, welche aus ihrer Umgebung herausstechen). Diese Salienzen können entweder vom Reiz induziert sein, beispielsweise zieht eine gelbe Fläche in einer ansonsten blauen Umgebung die Aufmerksamkeit auf sich, oder von einem Ziel oder Aufgabe beeinflusst werden. Hier wird beispielsweise die Suche nach einer Büroklammer den Blick mehr über den Schreibtisch wandern lassen als auf dem Boden oder Fenster.

In der Studie von Bindemann et al. [18] wird das Blickverhalten auf zufällig präsentierte Gesichter untersucht. Der sog. Schwerpunkt-Effekt dauert etwa 250 ms und sorgt dafür, dass man bei einem plötzlich präsentierten Bildreiz fast automatisch den Blick auf das Zentrum des Reizes richtet. Wenn man diesen Effekt vernachlässigt, betrachten die meisten Personen insbesondere Augen und Nase des präsentierten Gesichts. Daraus abgeleitet kann man für ein Blickmodell die Fokussierung auf das Gesicht zu 60% auf die Augen und zu 40% auf die restlichen Gesichtsmerkmale wie Mund und Nase verteilen. Im Gegensatz zu dieser Blickmessung auf statische Bilder von Gesichtern wird in [19] der Blick in einer zwischenmenschlichen Dialogsituation gemessen. Hierbei wird bei gegenseitigem Anblicken der Probanden primär das Gesicht des Gegenübers fokussiert, ohne spezielle Merkmale zu betrachten. Bei den Gesichtsmerkmalen wiederum lag der Mund vorne, welches für eine Dialogsituation plausibel erscheint, gefolgt vom Blick auf die Augen. Allerdings wurden in der Studie so gut wie keine direkten Blickkontakte (Auge-zu-Auge) festgestellt. Sollte dieses Ergebnis nicht an Trackingfehlern, der relativ kleinen Probandengruppe ($N = 5$ mit Trackingergebnissen) oder der Einwirkung der Blickerfassung auf das Blickverhalten der Probanden liegen, könnte dies ein wirklich bedeutsames Ergebnis sein, weil es dem bisher vermuteten Blickverhalten widerspricht.

Neben dem gesichtsrelevanten Blickverhalten wurden auch auditive und visuelle Salienzen für das Modell betrachtet. Für die auditiven Reize wurde die Richtung der Quelle mittels interauraler Zeitdifferenz ermittelt und die Energie als Intensität für den Reiz ausgewertet. Anhand der Richtung der Audioquelle und der visuellen Information, wo das Gesicht des Gesprächspartners liegt, kann ein Audiosignal zudem von der Sprache des Dialogpartners unterschieden werden und somit als eine auditive Salienz identifiziert werden. Für die Ermittlung der visuellen Salienz wurde die Fast-SUN Implementierung [20] gewählt. Hierbei wird mittels einer schnellen Implementierung des SUN-Algorithmus ein menschenähnliches Suchverhalten, trainiert auf Bildsequenzen, nachgebildet. Hierfür wird der Ansatz verwendet, dass ein Objekt im Bild gesucht wird, dessen Klassenzugehörigkeit für einen Punkt $\mathbf{z}$ mit der Zufallsvariablen (ZV) $c_\mathbf{z}$ angegeben wird, die ZV $l_\mathbf{z}$ beschreibt die Position und die ZV $f_\mathbf{z}$ die Merkmale des Punktes $\mathbf{z}$. Damit kann die visuelle Salienz $s_\mathbf{z}$ für den Punkt $\mathbf{z}$ wie folgt definiert werden:

$$s_\mathbf{z} = \frac{p(f_\mathbf{z}, l_\mathbf{z}|c_\mathbf{z})p(c_\mathbf{z})}{p(f_\mathbf{z}, l_\mathbf{z})} \tag{8}$$

Mit der Annahme, dass Merkmale und Position unabhängig und bedingt unabhängig bei gegebener Klassenzugehörigkeit $c_\mathbf{z}$ sind sowie anschließender Logarithmierung ergibt sich aus Gleichung (8):

$$\log s_\mathbf{z} = -\log p(f_\mathbf{z}) + \log p(f_\mathbf{z}|c_\mathbf{z}) + \log p(c_\mathbf{z}|l_\mathbf{z}) \tag{9}$$

Der erste Summand aus Gleichung (9) ist unabhängig von Klasse und Position. Somit führt eine geringe Auftrittswahrscheinlichkeit für Merkmale an einem Punkt, zu einem hohen Beitrag bei der Salienzberechnung. Der zweite Summand beschreibt die Wahrscheinlichkeit für Merkmale für eine gegebene Klassenzugehörigkeit und der letzte Summand gibt das Vorwissen für die Auftrittswahrscheinlichkeit der Klasse bei gegebener Position an. Da in diesem Fall keine speziellen Objekte gesucht werden, reduziert sich die Salienzberechnung im Wesentlichen auf den ersten Summanden, wobei Merkmale, welche im Gesamtbild eher selten sind, einen hohen Wert für die Aufmerksamkeit erhalten und häufig auftretende Merkmale niedrig bewertet werden. Dies entspricht auch der obigen Definition reizinduzierter Salienzen.

Mittels obiger Formel kann nun für jedes Pixel in einem Bild die visuelle Salienz berechnet werden. Mittels der Betrachtung der Bildfolge über der Zeit kann zudem auch eine zeitliche Abschwächung der visuellen Reize erfolgen, so dass neuere Reize stärker gewichtet werden. Mit diesen Berechnungen ergibt sich eine Salienzkarte,

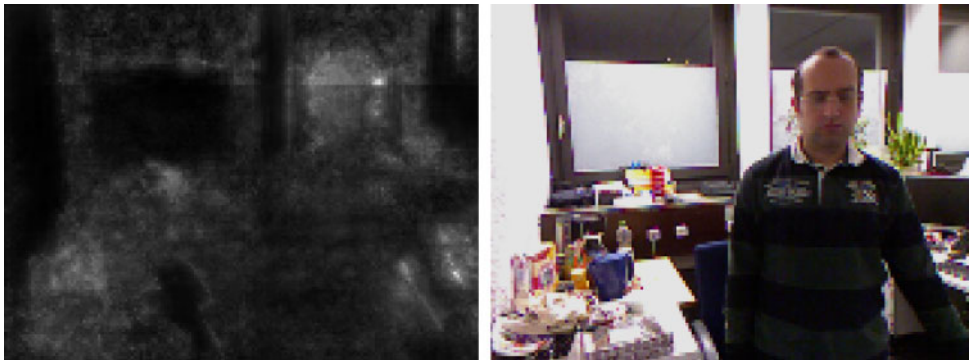

Bild 10 Salienzkarte (links) des auf 160×120 Pixel skalierten Farbbildes (rechts).

Tabelle 5 Übersicht über Blicksteuerung im Dialogmodus.

Modus	Dialog
Gesichtsverfolgung	80%
$\begin{pmatrix}\text{Augen (R/L)}\\ \text{Mund/Nase}\end{pmatrix}$	$\begin{pmatrix}30\%/30\%\\ 20\%\end{pmatrix}$
Visuelle Salienz	20%
Auditive Salienz	Energie > 10× Umgebung

welche in Bild 10 zu sehen ist, in welcher die Salienz mit der Helligkeit der Pixel codiert ist. Aus dieser Salienzkarte wird mittels Maximumssuche der hellste und somit momentan salienteste Punkt $\mathbf{p}_{max(s_z)}$ berechnet.

Mit den berechneten visuellen Salienzen, den auditiven Salienzen und den Gesichtsmerkmalen des Dialogpartners kann nun das in Tabelle 5 dargestellte Modell für die Blicksteuerung im Dialogmodus realisiert werden, welches auf [18] basiert.

Im Dialogmodus wird zu 60% auf die Augenregion fokussiert und zu 20% auf die Mund- sowie Nasenregion. Zu 20% wird auf die berechneten visuell salienten Punkte $\mathbf{p}_{max(s_z)}$ außerhalb des Gesichts reagiert, während auditive Salienzen den Fokus bekommen, wenn diese um den Faktor 10 über dem Grundpegel liegen und nicht aus der Richtung des Gesprächspartners kommen. Neben der unbewussten visuellen und auditiven Salienz gibt es weitere intrinsische Faktoren, welche das Blickverhalten beeinflussen können. Solche weiteren Einflussfaktoren können die aktuelle Rolle im Dialog sein (Zuhörer oder Sprecher) oder das Interesse am Thema bzw. am Dialogpartner sowie andere Personen oder Objekte im Raum. Die ermittelten Punkte im Sensorbild müssen in Winkel für die Blicksteuerung der beiden Augen umgerechnet werden. Hierfür ist in der linken Seite von Bild 11 eine schematische Darstellung der oberen Hälfte der Roboterplattform skizziert. Auf der rechten Seite im Bild ist das 3-D-Gesichtsmodell dargestellt.

Für die Ansteuerung der Roboteraugen müssen nun die relevanten Punkte $\mathbf{p}_s$ (Gesichtsmerkmale oder visuelle Salienzen) vom Sensorkoordinatensystem in Punkte bezogen auf das linke $\mathbf{p}_{la} = (x_{p_{la}}, y_{p_{la}}, z_{p_{la}})^T$ und rechte $\mathbf{p}_{ra} = (x_{p_{ra}}, y_{p_{ra}}, z_{p_{ra}})^T$ Augenkoordinatensystem transformiert werden, um dann in den vertikalen Winkel a_{lv} und horizontalen Winkel a_{lh} für das linke Auge sowie a_{rv} und a_{rh} für das rechte Auge umgerechnet zu werden. Dies geschieht mittels folgender zwei Gleichungen:

$$\mathbf{p}_{la} = \mathbf{t}_{la} + R_{k_h,k_v}(\mathbf{t}_k + R_{s_h,s_v}(\mathbf{t}_s + \mathbf{p}_s)) \tag{10}$$

$$\mathbf{p}_{ra} = \mathbf{t}_{ra} + R_{k_h,k_v}(\mathbf{t}_k + R_{s_h,s_v}(\mathbf{t}_s + \mathbf{p}_s)) \tag{11}$$

Die obigen Gleichungen berücksichtigen die Verschiebung ins Sensorzentrum $\mathbf{t}_s$ sowie die Rotation um die Winkel s_h und s_v des Sensors, Verschiebung ins Kopfzentrum $\mathbf{t}_k$ sowie Rotationen um die Winkel des Kopfes k_h und k_v und die Translation ins linke $\mathbf{t}_{la}$ bzw. rechte $\mathbf{t}_{ra}$ Augenzentrum.

Aus $\mathbf{p}_{la}$ und $\mathbf{p}_{ra}$ lassen sich nun die eigentlich gesuchten Winkel a_{lv}, a_{lh}, a_{rv}, a_{rh} berechnen:

$$a_{lv} = \arctan\left(\frac{y_{p_{la}}}{z_{p_{la}}}\right) \tag{12}$$

$$a_{rv} = a_{lv} \tag{13}$$

$$a_{lh} = \arctan\left(\frac{x_{p_{la}}}{z_{p_{la}}}\right) \tag{14}$$

$$a_{rh} = \arctan\left(\frac{x_{p_{ra}}}{z_{p_{ra}}}\right) \tag{15}$$

$a_{lv} = a_{rv}$ gilt, da sich $\mathbf{p}_{la}$ und $\mathbf{p}_{ra}$ nur in der x Komponente unterscheiden und diese für die Winkelberechnung nicht benötigt wird.

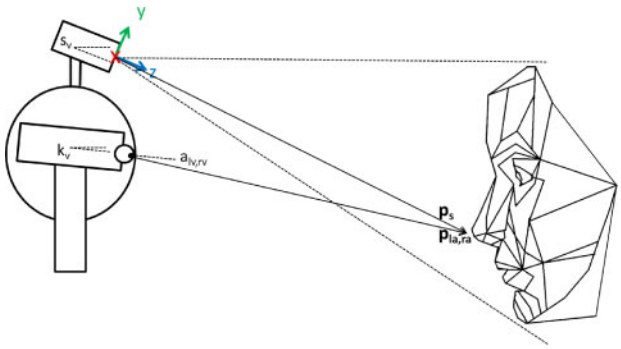

Bild 11 Schematische Seitenansicht des Kopfbereichs der Plattform mit 3-D-Modell des Gesichtes.

- 10.1515/auto.2013.1062

5.5 Ablaufsteuerung

Mit der Ablaufsteuerung soll die Interaktion zwischen Mensch und Maschine (in diesem Fall die ELIAS-Plattform) realisiert werden, dabei lassen sich im Grunde drei Komponenten unterscheiden: Eingabe, Verarbeitung und Ausgabe.

Für die Eingabe kann das System unterschiedliche menschliche Modalitäten verarbeiten. Die ELIAS-Plattform ist in der Lage Sprache, Gesten und Mimik zu erkennen, dafür werden die entwickelten Komponenten von Abschnitt 5.1 bis Abschnitt 5.4 verwendet. Eine kommerzielle Lösung des Microsoft Kinect SDK kommt für die Spracherkennung[6] zum Einsatz, des Weiteren können über den Touchscreen Eingaben vorgenommen werden.

Die Verarbeitung der wahrgenommenen menschlichen Modalitäten wird anhand eines Expertensystems realisiert. Dieses System benutzt Regeln und Fakten um verschiedene Eingaben auf Entscheidungsebene zu verarbeiten, dafür ist das Java Expert System Shell (JESS)[7] verwendet worden. Mittels der Regeln kann nun die Ablaufsteuerung die agierenden Komponenten der ELIAS-Plattform ansteuern.

Unterschiedliche Komponenten stehen der ELIAS-Plattform als Ausgabe zur Verfügung. Ähnlich wie bei der Spracherkennung kommt auch für die Sprachsynthese eine kommerzielle Lösung zum Einsatz[8]. Des Weiteren können Informationen direkt auf dem Bildschirm angezeigt werden. Unterschiedliche Systemzustände (z. B. aktive Spracherkennung) können anhand des LED-Rings dargestellt werden. Aufgrund der mit zwei Piezo-Aktoren ausgestatteten Roboteraugen ist die ELIAS-Plattform in der Lage, komplexe sowie schnelle Blickbewegungen für die Gestaltung der Interaktion zu nutzen. Auch die Mobilität der Plattform kann genutzt werden, um auf Personen zufahren zu können und proaktiv zum Spielen aufzufordern.

6 Akinator-Spiel als eine Anwendung für die Roboterplattform

Der Einsatz von Robotik bietet ein breites Anwendungsspektrum, dabei können Roboter eingesetzt werden, um verschiedene Haushaltstätigkeiten (staubsaugen, Böden wischen etc.) zu erledigen, sowie im Gesundheitswesen einfache Aufgaben (Botendienste, Getränkeverteilung etc.) zu übernehmen. Im Folgenden wird eine Unterhaltungsanwendung betrachtet für den Einsatz der ELIAS-Plattform zur multimodalen Interaktion in einem Spieleszenario. Im Gegensatz zu gewöhnlichen Computersystemen aus dem Heimbereich (PC, Laptop, Tablet etc.) ist die Roboterplattform in der Lage aufgrund der Mobilität proaktiv mit dem Menschen zu interagieren. Des Weiteren sind aufgrund der technischen Ausstattung viele unterschiedliche Ausgestaltungen der

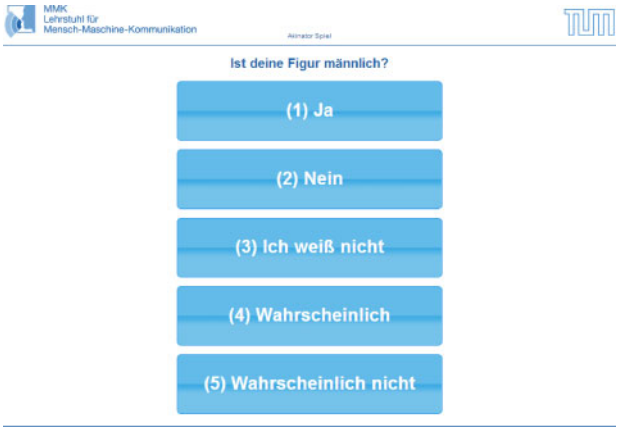

Bild 12 Die erste Frage des Akinator-Spiels.

Interaktion möglich, beispielsweise kann anhand des optionalen dynamischen Gestentrainings das Spiel mit einer physischen Ertüchtigung kombiniert werden. Dieses Konzept der Kombination von Spiel mit der physischer Ertüchtigung wird als *Exergames* bezeichnet und wird unter anderem Bereich von *Ambient Assisted Living* angewandt.

Das Spiel Akinator[9] wurde als eine Beispielanwendung für die multimodale Interaktion mit der ELIAS-Plattform verwendet. Die Ansteuerung des Spiels erfolgte über die von der Firma *Elokence* bereitgestellte Schnittstelle, somit kann das Spiel auf multimodale Art und Weise gespielt werden. Das Akinator-System versucht eine bekannte Figur zu erraten, an die der Benutzer denkt. Hierfür stellt es geschickt Fragen, um durch Ausschlussverfahren in einem Entscheidungsbaum zu navigieren und idealerweise zu einem Blatt am Ende des Baumes zu kommen, welches die Lösung enthält. Eine typische Frage zu Beginn ist beispielsweise, ob die Figur weiblich ist. Antwortmöglichkeiten des Nutzers sind immer *Nein*, *Wahrscheinlich nicht*, *Ich weiß nicht*, *Wahrscheinlich*, *Ja*. Für die Interaktion mit dem Roboter kann der Nutzer diese fünf Antwortmöglichkeiten dann entweder per Sprache, Geste oder als Feld am Bildschirm auswählen. Bild 12 zeigt die graphische Benutzeroberfläche für das Akinator-Spiel.

Der Ablauf eines Akinator-Spiels mit der ELIAS-Plattform besteht aus vier Phasen: *Beginn*, *Gestentraining*, *Spielen* und *Ende*.

1. *Beginn*: In der ersten Phase des Spiels erkennt die ELIAS-Plattform die mögliche Anwesenheit eines Menschen aufgrund der installierten Sonarsensoren sowie Kameras und verifiziert danach die Personenhypothese anhand des in Abschnitt 5.1 beschriebenen Verfahrens. Per Sprachsynthese fordert die Plattform den Menschen auf, eine Runde des Spiels Akinator zu spielen. Der Mensch kann nun via Sprache seine Zustimmung bzw. Ablehnung ausdrücken. Während dieser Phase wird die Grammatik der Spracherkennung nur für Aussagen der Zustimmung (z. B. ja, okay

[6] http://www.microsoft.com/en-us/kinectforwindows/develop/
[7] http://herzberg.ca.sandia.gov/
[8] http://www.ivona.com/en/
[9] http://www.akinator.com

etc.) bzw. der Ablehnung (z. B. nein, nicht etc.) angepasst, um die Erkennungsleistung zu verbessern.
2. *Gestentraining*: In dieser optionalen Phase können für die fünf Antwortmöglichkeiten dynamische Gesten der oberen Körperhälfte dem System beigebracht werden. Weitere Details zu diesem Verfahren finden sich in [11].
3. *Spielen*: Per Sprachsynthese fordert die Plattform nun den Menschen auf, an eine bekannte Person zu denken. Das System beginnt damit, die erste Frage zu stellen. Dies geschieht sowohl per Sprachsynthese als auch auf dem Bildschirm. Der Mensch hat die Möglichkeit sich eine gestellte Frage erneut vorlesen zu lassen, die Anweisung kann per Sprache mit unterschiedlichen Phrasen (bitte wiederholen, Frage erneut vorlesen etc.) erteilt werden. Die Auswahl der fünf Antwortmöglichkeiten kann sowohl per Spracheingabe, statischer Gesten (*Eins* bis *Fünf*, siehe Abschnitt 5.2), dynamischer Gesten oder Touchscreen erfolgen.
4. *Ende*: Wenn das Akinator-System ein gewisses Maß an Konfidenz in Bezug auf die Identität der gesuchten Person erreicht hat, wird die wahrscheinlichste Vermutung sowohl per Sprachsynthese als auch per Bildschirm ausgegeben. Der Mensch hat nun die Möglichkeit ein Urteil über die Hypothese abzugeben. Bei einem richtigen Rateversuch seitens des Systems hat der Mensch die Möglichkeit, das Spiel zu beenden oder eine neue Runde Akinator zu spielen. Bei einem falschen Rateversuch besteht die Möglichkeit das Spiel abzubrechen oder anhand von weiteren Fragen das Spiel fortzusetzen.

Die Mimikerkennung sowie die Blicksteuerung bilden interessante Ansätze zur weiteren Forschungsarbeit. Hierbei kann die Mimikerkennung beispielsweise Zusatzinformation liefern, wie unterhaltsam das momentane Spiel ist. Auch das Interesse der Person am Dialog oder dem Spiel ist eine nützliche Information, hierfür müssen aber Modifikationen am Erkennungssystem gemacht werden, da sich dies nicht direkt aus den sechs Basisemotionen ableiten lässt. Auch für die Blicksteuerung bieten sich weitere Untersuchungen sowie ein Vergleich mit menschlichem Blickverhalten an.

7 Zusammenfassung

Dieser Beitrag beschreibt die multimodalen Interaktionsmöglichkeiten der ELIAS-Plattform im Akinator-Spiel. Aufgrund der technischen Ausstattung ist die ELIAS-Plattform in der Lage, unterschiedliche menschliche Modalitäten (Stimme, Gesten und Mimik) wahrzunehmen sowie selbst mehrere Möglichkeiten zur Interaktion (Bildschirmdarstellung, Sprachsynthese und Blick) anzubieten. Ein erstes Zusammenspiel von Mensch und Roboter im multimodalen Kontext wurde im Akinator-Spiel realisiert.

Durch die entwickelten Komponenten wirkt die Interaktion mit der Plattform viel natürlicher und somit auch das Spiel immersiver als bei der Interaktion an einem gewöhnlichen PC. Ältere Personen haben die Möglichkeit, den für sie besten Kanal für die Interaktion zu selektieren. Die verwendete Personenhypothesenverifikation hilft dabei, nicht mit falschen Hypothesen eine Interaktion zu initiieren. Des Weiteren kann der Nutzer seine Antworten über verschiedene Kanäle geben, wie bei einem Spiel mit einem echten Menschen. Dies trägt dazu bei, dass die Interaktion zwischen Mensch und Maschine einfacher vonstattengehen kann. Jedoch ist es noch eine Herausforderung, ein gutes und robustes Modell für die Erfassung sowie die Wahrnehmung des Menschen zu entwickeln (eine Verbesserung kann beispielsweise mithilfe des Kinect 2.0 Sensors erfolgen). Die Wahrnehmung des Menschen sollte kontextsensitiv in geeignete Signale für die vorhandenen agierenden Komponenten umgewandelt werden. Meist geht dies, wie bei unserem Ansatz auch, über ein konkretes Anwendungsszenario nicht hinaus.

Danksagung

Die Autoren möchten sich bei den folgenden Personen für die Unterstützung und den geleisteten Arbeiten an der Forschungsplattform ausdrücklich bedanken. Für die Bereitstellung des Zugangs zum Akinator bei Arnaud Megret und den Support durch Valentin Ferriere von http://www.elokence.com. Für die Arbeiten an der Roboterplattform bei Herrn Erich Lingfeng Jiang, Yigit Arin, Philipp Tiefenbacher, Erich Schneider, Stefan Kohlbecher, Klaus Bartl, Christoph Mayer und Frank Wallhoff. Ein weiterer Dank geht an Alexander Bannat und Martina Römpp für das Korrekturlesen. Diese Forschungsarbeit wurde unterstützt durch das Projekt ALIAS (AAL-2009-2-049) im Rahmen des AAL Joint Programme mit Mitteln der EU und nationalen Förderungsinstitutionen aus Österreich, Frankreich und Deutschland sowie das Projekt CustomPacker (PPP-FoF-260065) gefördert von der EU im 7. Rahmenprogramm in der ‚Factories of the Future'-Initiative.

Literatur

[1] *A. Jaimes und N. Sebe*. Multimodal human computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1–2):116–134, 2007.
[2] *R. Sharma, V. Pavlovic, und T. Huang*. Toward multimodal human-computer interface. *Proceedings of the IEEE*, 86(5):853–869, 1998.
[3] *S. Oviatt*. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.
[4] *R. Stiefelhagen, H. K. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel*.Enabling Multimodal Human-Robot Interaction for the Karlsruhe Humanoid Robot. *IEEE Transactions on Robotics*, 23(5):840–851, 2007.
[5] *V. Gonzalez-Pacheco, A. Ramey, F. Alonso-Martin, A. Castro-Gonzalez, und M. Salichs*. Maggie: A Social Robot as a Gaming Platform. *International Journal of Social Robotics*, 3:371–381, 2011.
[6] *T. Spexard, M. Hanheide and G. Sagerer*. Human-Oriented Interaction With an Anthropomorphic Robot. *IEEE Transactions on Robotics*, 23(5):852–862, 2007.

[7] *C. Borst, T. Wimböck, F. Schmidt, M. Fuchs, B. Brunner, F. Zacharias, P. R. Giordano, R. Konietschke, W. Sepp, S. Fuchs, C. Rink, A. Albu-Schäffer, and G. Hirzinger*. Rollin' Justin – Mobile platform with variable base. In *ICRA*, pages 1597–1598. IEEE, 2009.

[8] *M. Fujita*. On activating human communications with pet-type robot AIBO. *Proceedings of the IEEE*, 92(11):1804–1813, 2004.

[9] *E. Schneider, S. Kohlbecher, K. Bartl, F. Wallhoff, and T. Brandt*. Experimental platform for Wizard-of-Oz evaluations of biomimetic active vision in robots. In *Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on*, pages 1484–1489, 2009.

[10] *P. Ekman*. Universals and Cultural Differences in Facial Expressions of Emotion. University of Nebraska Press, 1971.

[11] *T. Rehrl, J. Blume, A. Bannat, G. Rigoll, und F. Wallhoff*. On-line Learning of Dynamic Gestures for Human-Robot Interaction. In *35th German Conference on Artificial Intelligence, KI 2012*, 2012.

[12] *N. Dalal und B. Triggs*. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, Seiten 886–893, 2005.

[13] *C. J. C. Burges*. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[14] *T. Kanade, J. F. Cohn, und Y. Tian*. Comprehensive Database for Facial Expression Analysis. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, Seiten 46–53, 2000.

[15] *P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, und I. Matthews*. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis*, Seiten 94–101, 2010.

[16] *M. Pantic und L. Rothkrantz*. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[17] *J. Ahlberg*. CANDIDE-3 – An Updated Parameterised Face. Technischer Bericht, Deptartment of Electrical Engineering, Linköping University, 2001.

[18] *M. Bindemann, C. Scheepers, and A. Burton*. Viewpoint and center of gravity affect eye movements to human faces. *Journal of Vision*, 9(2):1–16, 2009.

[19] *F. Broz, C. L. Nehaniv, Dautenhahn, K., and H. Kose Bagci*. Towards Automated Human-Robot Mutual Gaze. In *Proceedings The Fourth International Conference on Advances in Computer-Human Interactions, ACHI 2011*, Gosier, Guadeloupe, France, February 2011.

[20] *N. Butko, L. Zhang, G. Cottrell, and J. Movellan*. Visual saliency model for robot cameras. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 2398–2403, May 2008.

Manuskripteingang: 31. März 2013

Dipl.-Inf. Jürgen Blume ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für Mensch-Maschine-Kommunikation im Fachbereich Elektrotechnik der Technischen Universität München. Hauptarbeitsgebiete: Mensch-Roboter-Interaktion und multimodale Kommunikation.

Adresse: Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, D-80290 München, Fax: +49-(0)89-289-28535, E-Mail: blume@tum.de

Dipl.-Ing. Tobias Rehrl ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für Mensch-Maschine-Kommunikation im Fachbereich Elektrotechnik der Technischen Universität München. Hauptarbeitsgebiete: Ambient Assisted Living und Mustererkennung.

Adresse: Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, D-80290 München, Fax: +49-(0)89-289-28535, E-Mail: tobias.rehrl@tum.de

Univ.-Prof. Dr.-Ing. habil. Gerhard Rigoll ist Ordinarius des Lehrstuhls für Mensch-Maschine-Kommunikation im Fachbereich Elektrotechnik der Technischen Universität München.

Adresse: Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, D-80290 München, Fax: +49-(0)89-289-28535, E-Mail: rigoll@tum.de