Proc. Neural Networks in Applications (NN'98), Magdeburg, pp. 211-218, 1998

Neuronale Verfahren zur Lokalisation und Bewertung von Handgesten in Real-World-Szenen *

Anja Brakensiek, Ulf-Dietrich Braumann, Andrea Corradini, Hans-Joachim Böhme, Horst-Michael Groß

Fachgebiet Neuroinformatik
Technische Universität Ilmenau
PF 10 05 65
D-98684 Ilmenau
Tel.: +49 3677 69 1307, Fax.: +49 3677 69 1665
E-Mail: anja@informatik.tu-ilmenau.de

Zusammenfassung

Für eine natürliche gestenbasierte Mensch-Maschine-Kommunikation in Real-World-Umgebungen ist zum einen die robuste Lokalisation eines potentiellen Nutzers und zum anderen die Detektion und Interpretation der Hände (Gesten) eine elementare Voraussetzung. Der vorliegende Beitrag behandelt die vier zur Nutzerlokalisation verwendeten Merkmale: Hautfarbe, Gesichtsstruktur, Kopf-Schulter-Silhouette und Bewegung. Anhand erster Ergebnisse wird verdeutlicht, daß erst die Kombination der verschiedenen, sich ergänzenden Verfahren die erforderliche *Robustheit* der Nutzerlokalisation unter Real-World-Bedingungen sichert, auf die die Handdetektion und Gesteninterpretation aufbauen können.

Schlüsselwörter: Mensch-Maschine-Kommunikation, visuelle Personenlokalisation, Growing-Neural-Gas-Netzwerk

1 Einleitung

Die Lokalisation von Personen und die Auswertung deren natürlicher Gesten (Handzeichen) ist Teilaspekt des GESTIK-Projektes am Fachgebiet Neuroinformatik, dessen Ziel in der Entwicklung einer neuronalen Architektur zur visuellen, gestenbasierten Interaktion eines Nutzers mit unserem mobilen Robotersystem MILVA[1] in Real-World-Umgebungen besteht ([3], [2], vgl. auch Ansatz in [5]).
Einsatzgebiete des Roboters, der als intelligenter Transportassistent dienen soll, könnten beispielsweise Flughäfen oder Supermärkte darstellen, denn hier ist eine intuitive Mensch-Maschine-Kommunikation erforderlich. Aufgrund der großen Varianz des Einsatzszenarios (unterschiedliche Beleuchtung und Umgebung, variable Distanzen zwischen Nutzer und Roboter) findet die visuelle Personenlokalisation mit Hilfe von mehreren parallel arbeitenden Verfahren statt, die in ihrer Verknüpfung eine robuste Angabe der Position erlauben. Ein weiteres Merkmal zur Nutzerlokalisation ist die akustische Auffälligkeit ([15]), auf die hier jedoch nicht näher eingegangen werden soll.
Die Nutzerlokalisation bildet die Voraussetzung für die eigentliche Handdetektion, bei der es vorerst nicht um 'Fingerzeichen' geht, sondern um die Bestimmung der Geste, die auf den relativen Positionen von Händen und Kopf beruht. Mit Hilfe eines definierten Posenalphabetes soll die Kommunikation (z.B. 'halte an', 'fahre nach links', etc.) mit dem Roboter erfolgen. Die visuell geführte Navigation, die auf einem Multi-Agenten-Ansatz beruht, der den zweiten Schwerpunkt des GESTIK-Projektes bildet, wird in [9] erläutert.

2 Modellarchitektur

Abbildung 1 zeigt den prinzipiellen Aufbau zur Nutzerlokalisation und Gestenerkennung.

* gefördert durch das Thüringer Ministerium für Wissenschaft, Forschung und Kultur (TMWFK), GESTIK-Projekt
[1] http://cortex.informatik.tu-ilmenau.de/technik.html

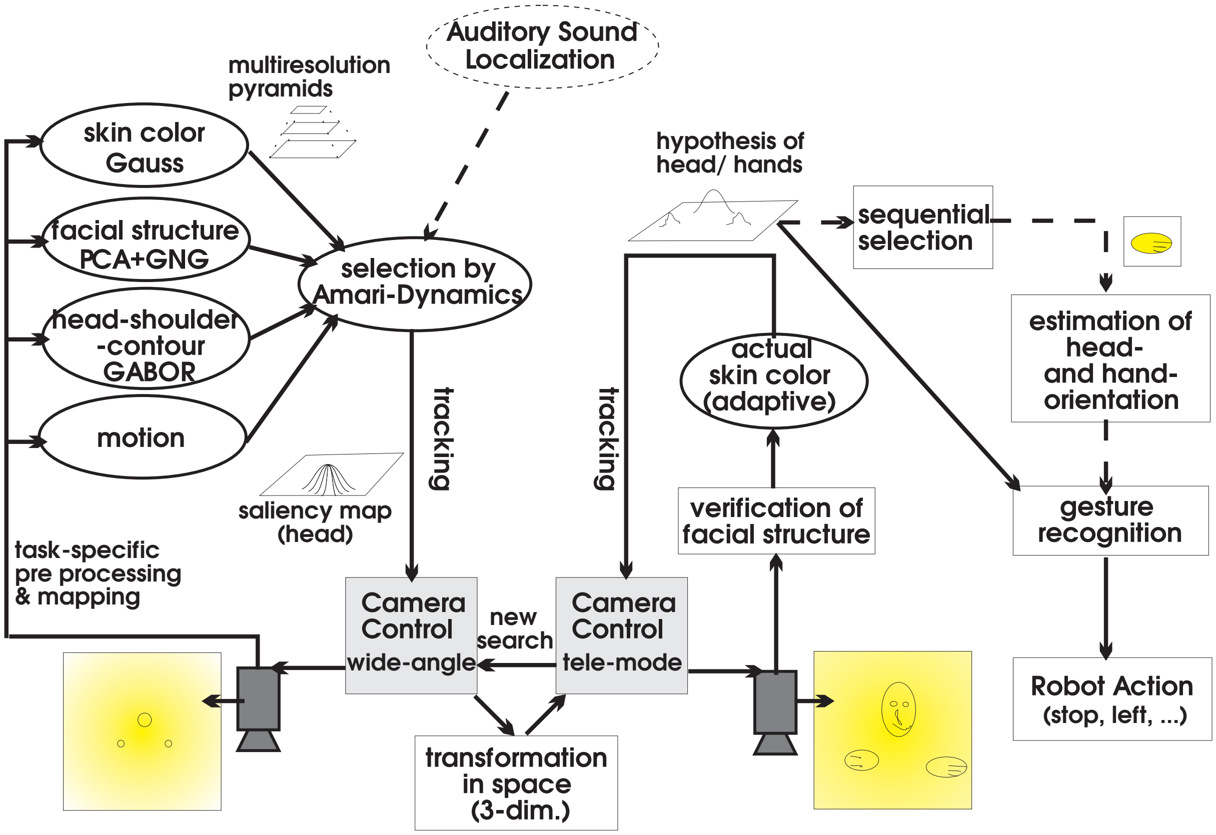

Abbildung 1: Prinzipdarstellung der Modellarchitektur: die gestrichelten Pfeile stellen einen Ausblick dar

Ein aktives Zwei-Kamera-System mit 7 Freiheitsgraden (für jede Kamera jeweils Pan/Tilt/Zoom, zusätzlich Pan für beide Kameras) beobachtet im Weitwinkelmodus die Einsatzumgebung, um so einen möglichst großen Suchbereich abzudecken. Nachdem im Bild einer der Kameras ein potentieller Nutzer detektiert wurde, wird die zweite Kamera so angesteuert, daß sie die Person zentral und in definierter Größe abbildet. In diesem Telebild erfolgt dann die Handdetektion und Gestenauswertung.

3 Verfahren zur Personenlokalisation

Die Lokalisation eines potentiellen Nutzers (Gesicht) findet aufgrund der variierenden Aufnahmebedingungen mit Hilfe von vier parallel arbeitenden und sich ergänzenden Verfahren (Cue-Module) statt: die Hautfarbklassifikation, die Detektion der Gesichtsstruktur, der Kopf-Schulter-Kontur und der Bewegung. Dabei wird vorausgesetzt, daß ein Nutzer, der mit dem Roboter kommunizieren will, zuerst frontal zu ihm steht. Somit treffen unter günstigen Bedingungen alle der zuvor genannten Merkmale zu: das Gesicht ist dem Roboter zugewandt, so daß sowohl die Hautfarbe als auch die Gesichtsstruktur detektiert werden könnte; die Silhouette der Person ist ebenfalls einheitlich. Die Bewegung hingegen ist ein Merkmal, welches zwar nicht zwingend auftreten muß, daß aber doch mit hoher Wahrscheinlichkeit durch eine Person erzeugt wird.
Um neben den xy-Koordinaten im Bild auch die grobe Entfernung des Nutzers zur Kamera schätzen zu können, werden die strukturbasierten Merkmale in einer Auflösungspyramide berechnet, in der das Bild jeweils um den Faktor $1/\sqrt{2}$ unterabgetastet wird. Aus der am besten zutreffenden Ebene und der bekannten Gesichtsgröße kann dann die ungefähre Entfernung berechnet werden.

Die Detektion der **Kopf-Schulter-Kontur** erfolgt mit Hilfe eines Gaborfilterarrangements auf dem Grauwertbild. Die Gaborfilter werden in Form einer durchschnittlichen Kopf-Schulter-Silhouette plaziert. Die Klassifikation wird über ein spezielles Abstandsmaß zu diesem Prototyp berechnet. Eine genaue Erläuterung des Verfahrens findet sich in [4]. Im folgenden werden die weiteren Lokalisationsverfahren konkret beschrieben.

3.1 Hautfarbklassifikation

Die Hautfarbe ist kein für die Personenlokalisation eindeutiges Merkmal und kann somit die weiteren Merkmale nur unterstützen. Zum einen können natürlich hautfarbene Regionen in der gesamten Szene (auch im Hintergrund) auftreten, zum anderen variiert die Hautfarbe verschiedener Personen unter unterschiedlichen und nicht bekannten Beleuchtungsbedingungen stark (vgl. Abb.3 rechts). Aus diesem Grund wird zur Nutzerlokalisation ein relativ grobes Hautfarbmodell verwendet, welches in erster Linie sicherstellen soll, daß Hautpartien möglichst sicher als hautfarben detektiert werden. Ein weiteres spezifischeres Hautfarbmodell, welches auf den aktuellen Nutzer adaptiert wird, wird zur Handlokalisation eingesetzt.
Um die mögliche Variabilität der Hautfarbe einzuschränken, werden zwei Maßnahmen ergriffen: bei jedem aufgenommenen Bild wird eine Farbadaptation durchgeführt und das zur Klassifikation verwendete Hautfarbmodell ist intensitätsnormiert.

Farbadaptation
Ziel der Farbadaptation ist es, weitgehend reproduzierbare Farbwerte bei Beleuchtungsänderungen zu erhalten.
Das hier verwendete Adaptationsverfahren wurde von [11] hergeleitet und bezieht sich auf die Darstellung der Bildpunkte (Farbwerte) in einem neurophysiologisch motivierten *Elementarfarbraum*. Die Farbtripel lassen sich aus dem RGB-Farbraum durch die folgende Lineartransformation in den Elementarfarbraum, der durch eine Rot-Grün (RG)-, Blau-Gelb (BY)- und Weiß-Schwarz (WS)-Achse aufgespannt wird, überführen:

$$\begin{pmatrix} RG \\ BY \\ WS \end{pmatrix} = \begin{pmatrix} 0.5 & -1.0 & 0.5 \\ -0.875 & 0.0 & 0.875 \\ 1.5 & 1.5 & 1.5 \end{pmatrix} * \begin{pmatrix} R \\ G \\ B \end{pmatrix}$$

Die im Farbraum dargestellten Pixel eines Farbbildes ergeben eine Punktwolke, deren größte Ausdehnung bei optimaler Beleuchtung (z.B. Tageslicht) entlang der Unbuntachse (Weiß-Schwarz-Achse) verläuft. Bei der Farbadaptation werden Punktwolken von nicht optimal (z.B. Glühlampenbeleuchtung) beleuchteten Bildern, die aus der Unbuntachse ausgelenkt sind, durch eine Scherung auf diese zurückgeführt. Im Anschluß erfolgt eine Dynamikanpassung entlang der Unbuntachse. Somit wird eine relativ stabile Farbempfindung realisiert. Die folgende Abbildung 2 veranschaulicht den Elementarfarbraum und das Adaptationsverfahren anhand eines Beispiels.

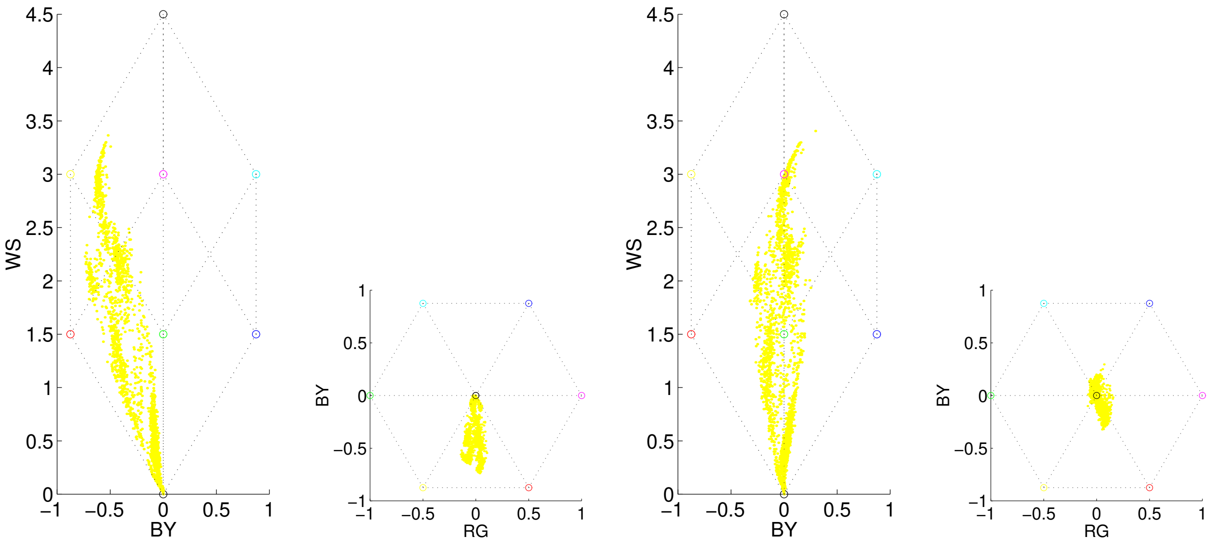

Abbildung 2: Punktwolke eines nicht optimal beleuchteten Bildes (links) und des adaptierten Bildes (rechts) im Elementarfarbraum (Projektion auf die BY-WS- bzw. RG-BY-Ebene)

Der Elementarfarbraum selbst wird in unserer Arbeit nicht zur Hautfarbklassifikation verwendet, er dient hier nur zur Darstellung des von uns angewandten Adaptationsverfahrens.

Farbmodell
Für die Erstellung einer Hautfarbdatenbank wurden per Hand segmentierte Hautpartien von farbadaptierten Portraitaufnahmen verwendet. Als Farbmodell dienen die intensitätsnormierten Farbanteile r' und g' (chromatische

Farben) aller Farbtripel der Hautfarbdatenbank (vgl. [14]).

$r' = \frac{R}{R+G+B}$ $\qquad$ $g' = \frac{G}{R+G+B}$

Aufgrund der Normierung kann die Varianz in der Intensität der Hautfarbe beim Klassifizieren größer sein als die, die bei der Erstellung der Datenbank vorhanden war. Bei der Klassifikation findet nur eine Beschränkung der erlaubten Intensität statt, um extrem helle bzw. dunkle Pixel auszusondern.
Für das verwendete Farbmodell wird das Farbhistogramm (2-dimensional) durch eine 2-dimensionale Gaußfunktion approximiert, indem der Mittelwert und die Kovarianzmatrix des Datensatzes gebildet werden (s. Abb.3 links).

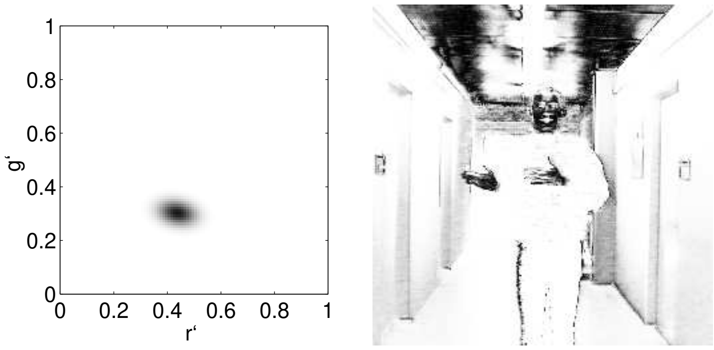

Abbildung 3: links: berechnetes Hautfarbmodell (Gaußfunktion); rechts: Hautklassifikationsbeispiel (Originalbild: s. Abb.5 rechts): Haut und Hintergrundregionen (hölzerne Decke) werden klassifiziert

Wurde ein Nutzer (Gesicht) gefunden, wird die aktuelle Gesichtsfarbe zur Bildung eines spezifischeren Hautfarbmodells herangezogen, sodaß im damit klassifizierten Bild möglichst nur noch Gesicht und Hände des Nutzers detektiert werden. Dabei wird der Mittelpunkt, die Ausdehnung und Orientierung der Gaußglocke an die aktuelle Farbverteilung angepaßt. Zweck der Anpassung des Hautfarbmodells ist zum einen, die Position der Hände möglichst sicher zu erkennen, zum anderen, das Tracken eines Nutzers zu erleichtern.

3.2 Gesichtsstrukturdetektion

Für die Detektion von Gesichtern wird ein Neuronales Netz, das Growing-Neural-Gas-Netzwerk (GNG), verwendet, welches vor der Beschreibung des eigentlichen Verfahrens im folgenden kurz erläutert wird.

Growing-Neural-Gas-Netzwerk (GNG)
Das GNG ist ein topologiebeschreibendes Netzwerk und verbindet Eigenschaften der Growing-Cell-Structures mit denen des Neural-Gas (siehe auch [8]). Die Aktivität y_i eines GNG-Knotens i ist eine Funktion des Abstandes (d_i) zwischen Eingangsvektor $\underline{x}$ und Referenzvektor $\underline{w}_i$.

$d_i = |\underline{x} - \underline{w}_i|$ $\qquad$ $d_b = min(d_i) \rightarrow b$ $\qquad$ $y_i = e^{-\frac{|\underline{x}-\underline{w}_i|^2}{\sigma_i^2}}$

Ein wesentlicher Vorteil ist die selbständige Anpassung der Netztopologie in Abhängigkeit vom lokalen Approximationsfehler. Das Netz wird mit 2 Neuronen, die mit einer Kante verbunden sind, als Minimalkonfiguration initialisiert; der Approximationsfehler steuert das Einfügen und Löschen von Neuronen und Kanten, wobei die Dichte bzw. Wahrscheinlichkeit der Trainingsmuster berücksichtigt wird. Das heißt, bei jeder Repräsentation eines Eingabemusters werden die Gewichte $\underline{w}$ des Best-Matching-Neurons (b) und der Nachbarschaft in Richtung des Eingangsvektors adaptiert. Neue Referenzvektoren werden epochenweise zwischen den Neuronen eingefügt, die am häufigsten adaptiert wurden. Kanten wiederum existieren zwischen Neuronen, deren Zuständigkeitsbereiche σ im Eingangssignalraum benachbart liegen. Die mittlere Länge aller Kanten eines GNG-Knotens bestimmt die Größe seines Zuständigkeitsbereichs. Die Klassifikation erfolgt über das Best-Match-Neuron, d.h. über das Neuron, welches den geringsten euklidischen Abstand zum Eingangsvektor besitzt.
Das GNG kann sowohl im unüberwachten als auch im überwachten Modus, in welchem eine zusätzliche Ausgabeschicht vorhanden ist, verwendet werden. Das überwachte (supervised) GNG ähnelt dann stark einem Radial-Basis-Function-Netzwerk (RBF). Die GNG-Neurone bilden die verdeckte Schicht und werden vollständig mit einer Ausgabeschicht verschaltet. Beim Training erfolgt die Gewichtsänderung der Ausgabeschicht nach der Delta-Regel und zwar parallel zur Adaption der GNG-Neurone.
Für eine genaue Beschreibung des GNG sei auf [7] verwiesen.

Beschreibung des Verfahren

Die Detektion von Gesichtern erfolgt im Grauwertbild mit Hilfe von Eigenfaces (Eigenvektortransformation, Principal Component Analysis, PCA; vgl. [10]). Die Eigenvektoren wurden von den Bildern einer Gesichtsdatenbank (ORL-Datenbank[2]) berechnet. Von diesen Trainingsbildern wurde ein 15x15 Pixel großer Gesichtsausschnitt (frontal) gewählt und bezüglich Mittelwert und Standardabweichung normiert (siehe Abb.4). Die Wahl des Gesichtsausschnittes anstelle des gesamten Kopfes (mit Hintergrund) verringert die mögliche Varianz der Bilder. Die Fitwertvektoren, die sich ergeben, wenn man die drei eigenwertgrößten Eigenvektoren mit den Trainingsbildern faltet, werden einem GNG zum Training übergeben.

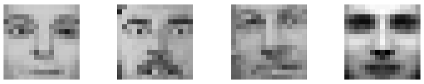

Abbildung 4: 3 beispielhaft ausgewählte Gesichter und Mittelwert der positiven Trainingspattern (normierte Darstellung)

Diese Art der Vorverarbeitung und Merkmalsextraktion wird nicht nur mit den Gesichtern (Positivbeispiele), sondern auch mit ausgewählten Negativbeispielen durchgeführt. Diese Negativbeispiele werden mit Hilfe eines Bootstrap-Algorithmus (siehe auch [13]) ermittelt. Hierbei werden falsch positiv klassifizierte Bildregionen der Menge der Negativbeispiele zugefügt.
Durch das Training des GNG im überwachten Modus mit Positiv- und Negativbeispielen können schärfere Klassengrenzen erzielt werden. Den folgenden Ergebnissen liegen je 174 Positiv- und Negativbeispiele zugrunde.
In der Recall-Phase muß die beschriebene Normierung und Faltung mit der Eigenvektormatrix mit jedem Bildausschnitt durchgeführt werden, ehe der Fitwertvektor vom GNG klassifiziert werden kann. Dabei muß das Fenster schrittweise über das Bild geschoben werden, da das verwendete Verfahren der PCA-Klassifikation nur in geringem Maße verschiebungsinvariant ist. Ist die eigentliche Gesichtsstruktur (Augen, Mund) aufgrund der Beleuchtungsbedingungen (z.B. Gegenlicht) oder der Entfernung nicht deutlich erkennbar oder die Auflösung zu gering, versagt die PCA.
Wie das folgende Beispiel zeigt, werden nicht nur Gesichter klassifiziert (Abb.5). Ein wesentliches Merkmal ist jedoch, wie auch bei der Detektion der Kopf-Schulter-Partie, daß eine korrekte Lokalisation meist auch auf den korrespondierenden Positionen der benachbarten Auflösungsebenen (Auffälligkeitspyramide) stattfindet.

Abbildung 5: Detektion der Gesichtsstruktur: die Quadrate zeigen die wahrscheinlichsten Gesichtsregionen, wobei ihre Größe der jeweiligen Auflösungsebene entspricht

3.3 Bewegungsdetektion

Die Detektion und Auswertung der Bewegung basiert auf Sequenzen von Grauwertbildern, in denen zwei wesentliche Bewegungseigenschaften extrahiert werden. Dies sind zum einen die Regionen, in denen Bewegung aufgrund von variierenden Pixelwerten (Binary Motion Region, BMR) detektiert wird, und zum anderen die zeitliche Charakteristik der Bewegung in diesen Regionen (Motion History Image, MHI, vgl.[6]). Die Detektion von Regionen, in denen Bewegung vorliegt, kann zur Unterstützung der Nutzerlokalisation herangezogen werden, während die Bewegungscharakteristik im weiteren Projektverlauf zur Beschreibung dynamischer Gesten dienen soll.

Zur Bewegungsdetektion verwenden wir die neurophysiologisch motivierten Reichardt-Detektoren ([12]). Die Vereinigung aller Bewegungsregionen einer Bildsequenz ergibt das BMR-Bild (binär), aus dessen Gestalt Rückschlüsse auf die Art der Bewegung gezogen werden können (lokale Bewegung, Größe). Um im BMR-Bild die relevanten

[2] http://www.cam-orl.co.uk/facedatabase.html

Regionen von Rauschen bzw. geringen Körperbewegungen unterscheiden zu können, werden nur die jeweils größten Bereiche berücksichtigt. Dies geschieht über die Dynamik eines neuronalen Feldes (WTA). In jedem Zyklus wird die jeweils größte Region detektiert und aussortiert, bis alle übrigen Regionen kleiner sind als ein gewisser Prozentsatz der größten Region. Dabei werden exzitatorische Verbindungen der Nachbarschaft und des Inputs verwendet (Abb.6).

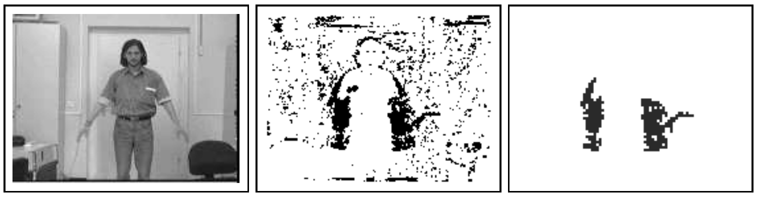

Abbildung 6: Beispielbild einer Bewegungssequenz ('Schmetterling'); BMR-Bild; Ergebnis des WTA-Zyklus

Bei der Verknüpfung mit den anderen Merkmalen zur Nutzerlokalisation ist zu beachten, daß hier keine direkte Korrespondenz im Ort besteht. Ein Winken beispielsweise ist zwar gestenrelevant, markiert aber nicht den Kopf (bzw. das Gesicht) der Person, der anhand der anderen Merkmale detektiert werden sollte. Somit kann die Bewegungsdetektion zur Nutzerlokalisation nur einen unterstützenden Charakter haben, um – gerade in relativ großen Entfernungen – die mögliche Nutzerposition grob zu schätzen.

3.4 Kombination der Module

Abhängig von den Umgebungsbedingungen (Beleuchtung, Szeneninhalt), die bei der Interaktion zwischen Roboter und Nutzer kaum beeinflußbar sind, bewirken die hier beschriebenen Verfahren unterschiedlich sichere Nutzerlokalisationen. Die Ergebnisse aller Cue-Module werden in topographisch organisierte Felder dynamischer Neuronen eingekoppelt (Amari-Dynamik, s.[1], [4]), um so eine eindeutige Nutzerposition (Ort und Entfernung) zu ermitteln (WTA-Verfahren). Dabei werden sämtliche Ebenen der Auffälligkeitspyramiden in den Selektionsprozeß mit einbezogen.
Ein Beispiel zur Nutzerlokalisation anhand mehrerer Merkmale (keine Bewegungsinformation) zeigt die folgende Abbildung (Abb.7). Sowohl die Gesichtsstrukturdetektion als auch die Kopf-Schulter-Partie kennzeichnet die vordere Person als den wahrscheinlichsten Nutzer. Das zur Nutzerlokalisation verwendete grobe Hautfarbmodell klassifiziert beide Personen und einige Hintergrundregionen. Aus der Verknüpfung kann die relevante Person gekennzeichnet und das spezifische Hautfarbmodell adaptiert werden. In der Regel ist die Nutzerlokalisation der einzelnen Merkmale allerdings nicht so eindeutig wie hier dargestellt.

Abbildung 7: farbadaptiertes Originalbild mit Ergebnis der Gesichtsstrukturdetektion (l.o., Quadrate geben die wahrscheinlichsten Gesichtsregionen an), Detektion der Kopf-Schulter-Partie (r.o.), Hautfarbklassifikationsergebnis des groben Modells (l.u., die dunkelsten Punkte besitzen die größte Hautfarbwahrscheinlichkeit), Hautfarbklassifikation des auf die vordere Person adaptierten Modells (r.u.)

Erst die parallele Nutzung der vier, sich gegenseitig ergänzenden Verfahren sichert die notwendige Robustheit der Nutzerlokalisation auch unter hochgradig variablen Umgebungsbedingungen. Durch den Einsatz der parallel operierenden Cue-Module wird das Gesamtsystem damit weniger abhängig vom Vorhandensein *einer bestimmten* Informationsqualität im Bild.

4 Gestenerkennung

Die Gestenerkennung bezieht sich vorerst auf statische Gesten (Posen), die nur durch die relative Position der Hände und des Kopfes zueinander definiert sind. Die sechs Grundposen, die für den Roboter relevant sind, sind im folgenden dargestellt (Abb.8).

Abbildung 8: Posenalphabet: 'hallo', 'halt', 'fahre nach links', 'fahre nach rechts', 'komm links neben mich', 'komm rechts neben mich'

Grundsätzlich sollen die Posen natürlich und selbsterklärend sein. Für den Einsatz in dem von uns angestrebten Szenario müssen die Posen für *jeden* intuitiv auszuführen sein. Die Handdetektion erfolgt dabei rein hautfarbbasiert auf der Basis des nutzerspezifischen Farbmodells.
Eine Erweiterung auf die Bestimmung der Kopf- und Handorientierung als weiteres gestenrelevantes Merkmal wird angestrebt. Diese soll mit Hilfe von Gaborfilterarrangements durchgeführt werden. Der Stand des jetzigen Verfahrens ist der, daß die Handorientierungsschätzung auf der Grundlage des hautfarbsegmentierten Bildes sicher funktioniert, wenn die *gesamte* Hand deutlich klassifiziert wurde, wovon jedoch unter beliebigen Bedingungen nicht immer ausgegangen werden kann.
Ein weiterer Schritt ist die Erkennung *dynamischer* Gesten, deren Ansatz bereits in Kapitel 3.3 kurz beschrieben wurde.

5 Zusammenfassung

Es wurde ein System vorgestellt, welches auf der Basis von vier sich gegenseitig unterstützenden Merkmalen eine robuste Nutzerlokalisation (Gesicht) in Real-World-Umgebungen realisiert. Die Hautfarbklassifikation findet auf dem farbadaptierten Bild mit Hilfe eines intensitätsnormierten Farbmodells statt. Die Detektion der Gesichtsstruktur wird mit Eigenfaces (PCA) und einem GNG auf mehreren Auflösungsebenen durchgeführt, um neben der Position im Bild auch die Entfernung bestimmen zu können. Gleiches gilt für die Detektion der Kopf-Schulter-Silhouette, welche auf der Filterung mit einem Gaborfilterarrangement basiert. Als weiteres Merkmal wurde die Bewegungsdetektion hinzugezogen, deren Verkopplung mit den übrigen Cues noch aussteht.
Durch die Verknüpfung über dynamische neuronale Felder soll die Position einer Person, die mit dem Roboter kommunizieren will, eindeutig bestimmt werden. Die sich anschliessende Handdetektion und Gestenerkennung befindet sich noch im Anfangsstadium und wird zur Zeit rein hautfarbbasiert durchgeführt, indem das Hautfarbmodell auf den aktuellen Nutzer adaptiert wird. Die Gestenerkennung, die sich vorerst auf statische Posen, die durch die Relativpositionen von Kopf und Händen definiert sind, bezieht, soll im weiteren durch eine Orientierungsschätzung der Hände bzw. Verwendung von dynamischen Gesten ergänzt werden.

Literatur

[1] Amari, S., 1977, Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields. *Biological Cybernetics*, **No. 27**, pp. 77-87.

[2] Böhme, H.-J., Brakensiek, A.,Braumann, U.-D., Krabbes, M., Gross, H.-M., 1997, Neural Architecture for Gesture-Based Human-Machine-Interaction. *Proceedings of the Bielefeld Gesture Workshop*, Bielefeld, Germany, September, to appear.

[3] Brakensiek, A., Braumann, U.-D., Böhme, H.-J., Rieck, C., Groß, H.-M., 1997, Farb- und strukturbasierte neuronale Verfahren zur Lokalisierung von Gesichtern in Real-World-Szenen. *Mustererkennung 97*, DAGM, Braunschweig, September, S. 113-120.

[4] Braumann, U.-D., Brakensiek, A., Böhme, H.-J., Groß, H.-M., 1998, Gaborfilterbasierte visuelle Personendetektion mit dreidimensionalen Feldern dynamischer Neuronen. *NN'98*, Magdeburg, Februar, wird erscheinen.

[5] Darrell, T., Basu, S., Wren, C., Pentland, A., 1997, Perceptually-driven Avatars and Interfaces: active methods for direct control. *M.I.T. Media Lab Perceptual Computing Section Technical Report*, No.416, M.I.T., USA.

[6] Davis, J.W., Bobick, A.F., 1997, The Representation and Recognition of Action using Temporal Templates. *IEEE Conference on Computer Vision and Pattern Recognition CVPR 97*.

[7] Fritzke, B., 1995, A Growing Neural Gas Network Learns Topologies. *Proc. of NIPS 7*, pp. 625-632.

[8] Hamker, F., Heinke, D., 1997, Implementation and Comparison of Growing Neural Gas, Growing Cell Structures and Fuzzy Artmap. *Schriftenreihe des Fachgebietes Neuroinformatik der TU Ilmenau*, Report Nr.1/97, TU Ilmenau.

[9] Krabbes, M., Böhme, H.-J., Stephan, V., Gross, H.-M., 1997, Extension of the ALVINN-Architecture for Robust Visual Guidance of a Miniature Robot. *Proc. EUROBOT'97*.

[10] Moghaddam, B., Pentland, A., 1995, Maximum Likelihood Detection of Faces and Hands. *Int. Workshop on Automatic Face- and Gesture-Recognition*, Zürich, pp. 122-128.

[11] Pomierski, T., Gross, H.-M., 1996, Biological Neural Architecture for Chromatic Adaptation Resulting in Constant Color Sensations. *Proc. of ICNN'96*, IEEE, pp. 734-739.

[12] Reichardt, W., Poggio T., Hausen K., 1983, Figure-Ground Discrimination by Relative Movement in the Visual System of the Fly. Part II:toward a neural circuity. *Biological Cybernetics*, pp. 1-30.

[13] Rowley, H. A., Baluja, S. Kanade, T., 1995, Human Face Detection in Visual Scenes. *Technical Report CMU-CS-95-158R*, USA.

[14] Yang, J., Waibel, A., 1996, A Real-Time Face Tracker. *WACV 96*, Sarasota, USA, pp. 142-147.

[15] Zahn, T., Izak, R., Trott, K., Paschke, P., 1997, A paced analog silicon model of auditory attention. *1st European Workshop on Neuromorphic Systems*, Stirling, August.