43rd International Scientific Colloquium
Technical University of Ilmenau
September 21-24, 1998

A. Brakensiek/ U.-D. Braumann/ H.-J. Böhme/ H.-M. Groß

Neuronale Verfahren zur Mensch-Roboter-Interaktion: Personenlokalisation und Gestenerkennung[1]

Zusammenfassung. Für eine natürliche gestenbasierte Mensch-Maschine-Kommunikation in Real-World-Szenen ist neben der robusten Lokalisation eines potentiellen Nutzers die Detektion der Hände eine elementare Voraussetzung. Der vorliegende Beitrag behandelt kurz die zur Nutzerlokalisation verwendeten Merkmale (Hautfarbe, Gesichtsstruktur und Kopf-Schulter-Silhouette), die direkt und indirekt zur Handdetektion herangezogen werden. Die Erkennung intuitiver statischer Gesten erfolgt farbbasiert und soll in Navigationskommandos für ein mobiles Robotersystem umgesetzt werden.

1 EINLEITUNG

Für eine benutzerfreundliche Mensch-Maschine-Interaktion spielen *natürliche* Kommunikationsformen und Eingabemedien, wie Sprache und Gestik, eine immer größere Rolle (vgl. Ansätze in [4, 7]). Das Ziel des GESTIK-Projektes liegt in der Entwicklung einer neuronalen Architektur zur visuellen, gestenbasierten Interaktion eines Nutzers mit unserem mobilen Robotersystem MILVA[2] in Real-World-Umgebungen (s.[1]). Dabei bildet die Lokalisation von Personen und die Auswertung deren Gesten (Körperposen) einen Schwerpunkt.
Einsatzgebiete des Roboters könnten alle Bereiche sein, in denen eine *möglichst natürliche, intuitiv verständliche* Mensch-Roboter-Kommunikation notwendig ist, wie beispielsweise bei einer Funktion als intelligenter Transport- oder Serviceroboter in Supermärkten oder Flughäfen. 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 (Multi-cue approach). Ein weiteres Merkmal zur Nutzerlokalisation ist die akustische Auffälligkeit (s. [14]), auf die hier jedoch nicht näher eingegangen werden soll.
Die Nutzerlokalisation bildet die Voraussetzung für die vorerst rein farbbasierte Handdetektion, bei der es allerdings nicht um 'Fingerzeichen' geht, sondern um die Bestimmung der Körpergeste, 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, wird in [8] erläutert.

2 SYSTEMARCHITEKTUR

Die folgende Abbildung 1 zeigt den prinzipiellen Aufbau der bisherigen Gesamtarchitektur zur Personenlokalisation und Gestenerkennung.
Ein aktives Zwei-Kamera-System beobachtet im Weitwinkelmodus die Einsatzumgebung, um so einen möglichst großen Suchbereich abzudecken. Anhand verschiedener Cues wird eine Person

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

gesucht, die mit dem Roboter in Interaktion treten will. Nachdem im Bild einer der Kameras mittels einer dynamischen Selektion ein möglicher Nutzer des Systems detektiert wurde, wird die zweite Kamera so angesteuert, daß sie die Person zentral und in definierter Größe abbildet. Anhand dieses Telebildes erfolgt dann die Handdetektion und Gesteninterpretation.

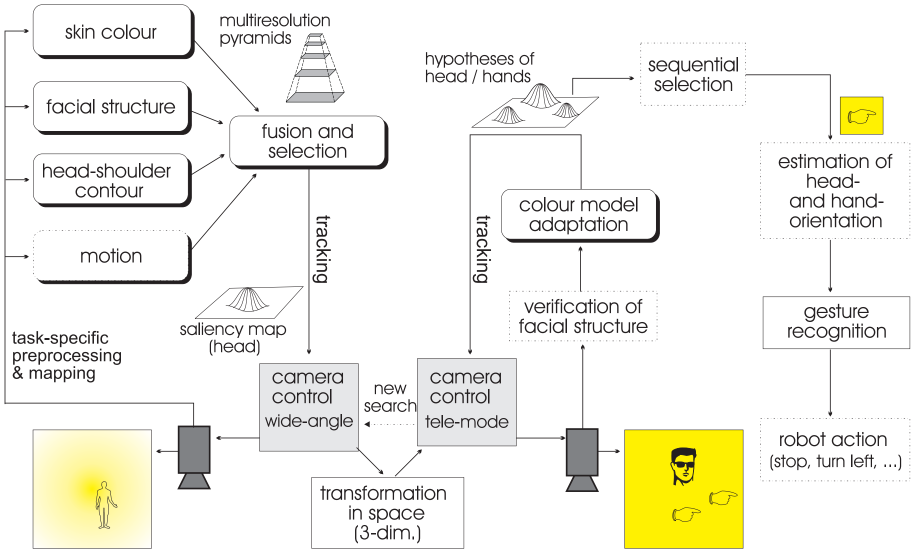


Abb. 1: Prinzipdarstellung der Systemarchitektur (gestrichelte Boxen: Ausblick)

3 PERSONENDETEKTION

Die Lokalisation eines potentiellen Nutzers (Gesicht) findet aufgrund der variierenden Aufnahmebedingungen mit Hilfe von mehreren parallel arbeitenden und sich ergänzenden Verfahren (Cue-Module, vgl. [2]) statt: die Hautfarbklassifikation, die Detektion der Gesichtsstruktur und der Kopf-Schulter-Kontur. Als Ergänzung kann die Bewegungsdetektion mit herangezogen werden, die jedoch nur in geringem Maße personenspezifisch ist. 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 bzw. Kopf-Schulter-Silhouette kann dann eine Entfernungsabschätzung vorgenommen werden.

Die Detektion 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 (Differenzbild) detektiert wird, und zum anderen die zeitliche Charakteristik der Bewegung in diesen Regionen (vgl.[5]). 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. Im folgenden werden die drei weiteren Lokalisationsverfahren konkreter 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, vgl. Abb.2 rechts: schwarz entspricht der größten Hautwahrscheinlichkeit) auftreten, zum anderen variiert die Hautfarbe verschiedener Personen unter unterschiedlichen und nicht bekannten Beleuchtungsbedingungen stark. 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 Handdetektion eingesetzt. Um die möglichen Variationen in der Hautfarbe zuzulassen, werden zwei Maßnahmen ergriffen: zum einen wird bei jedem aufgenommenen Bild eine Farbadaptation durchgeführt, um die Beleuchtungseinflüsse zu kompensieren und zum anderen wird zur Klassifikation ein intensitätsnormiertes Hautfarbmodell verwendet.

Farbadaptation: Ziel der Farbadaptation ist es, weitgehend reproduzierbare Farbwerte bei Beleuchtungsänderungen zu erhalten (s. [10]). Dies geschieht, indem die Punktwolke der Farbwerte eines Bildes im Farbraum so korrigiert wird, daß der Schwerpunkt der Punktwolke unbunt wird. Somit wird eine relativ stabile Farbempfindung realisiert, die allerdings immer noch szenenabhänging ist. Die bei Beleuchtungsänderung (Tageslicht, Neonlicht) auftretenden sprunghaften Farbänderungen können mit diesem Verfahren jedoch hinreichend genau kompensiert werden, so daß das Hautfarbmodell wieder zutrifft.

Farbmodell: Für die Erstellung einer Hautfarbdatenbank wurden per Hand segmentierte Hautpartien von farbadaptierten Portraitaufnahmen unterschiedlicher Hauttypen verwendet. Als Farbmodell dienen die intensitätsnormierten Farbanteile r und g (chromatische Farben) aller Farbtripel dieser Hautfarbdatenbank (vgl. auch [13]), wodurch das Hautfarbmodell unspezifischer wird $(r = R/(R+G+B)\ ,\ \ g = G/(R+G+B))$.

Für das verwendete Farbmodell wird das Farbhistogramm (s. Abb.2 mitte) durch eine 2-dimensionale Gaußfunktion approximiert, indem der Mittelwert und die Kovarianzmatrix des Datensatzes gebildet werden (s. Abb.2 links).

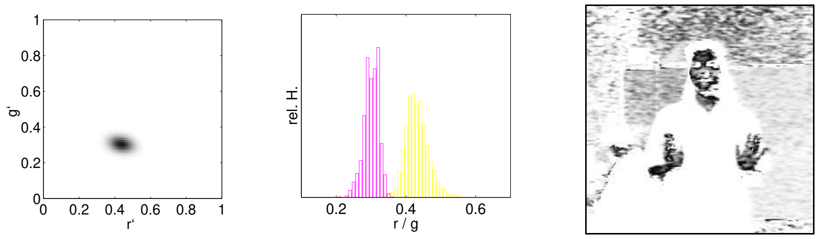


Abb. 2: v.l.n.r.: berechnetes Hautfarbmodell (Gaußfunktion); Histogramm der Farbanteile der Datenbank; Hautklassifikationsbeispiel (Original: s. Abb.3 links)

Wurde eine Person (Gesicht) mit Hilfe der verschiedenen Cues (s. auch Kapitel 3.2 und 3.3) gefunden, wird die aktuelle Gesichtsfarbe zur Bildung eines spezifischeren Hautfarbmodells herangezogen, so daß im damit klassifizierten Bild möglichst nur noch Gesicht und Hände des Nutzers detektiert werden. Dabei wird der Mittelpunkt, die Ausdehnung und Orientierung der Gaußglocke des allgemeinen Farbmodells an die aktuelle Hautfarbverteilung der Person angepaßt. Ziel 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 ein topologiebeschreibendes Netzwerk ist (s. [6]). Das überwachte (supervised) GNG, welches hier verwendet wird, besitzt eine zusätzliche Ausgabeschicht und ähnelt stark einem Radial-Basis-Function-Netzwerk. Die GNG-Neurone bilden die verdeckte Schicht und werden vollständig mit der Ausgabeschicht verschaltet.

Beschreibung des Verfahrens: Die Detektion von Gesichtern erfolgt im Grauwertbild mit Hilfe von Eigenfaces (Principal Component Analysis, PCA; vgl. [9]). Die Eigenvektoren wurden von den Bildern einer Gesichtsdatenbank (ORL-Datenbank[3]) berechnet. Dabei wurde ein 15×15

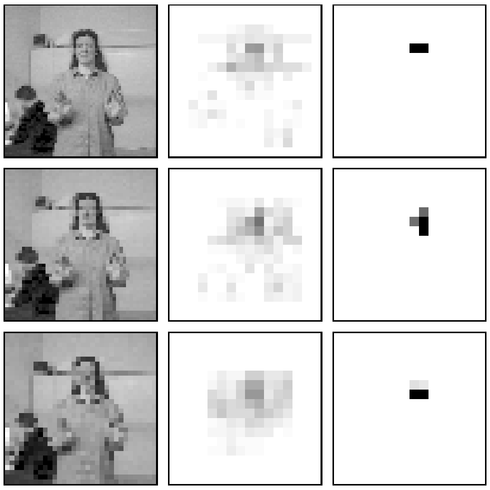
[3] http://www.cam-orl.co.uk/facedatabase.html

Pixel großer frontaler Gesichtsausschnitt gewählt und bezüglich Mittelwert und Standardabweichung normiert. Die Fitwertvektoren, die bei der Faltung der drei eigenwertgrößten Eigenvektoren mit den Trainingsbildern entstehen, werden dem GNG zum Training übergeben (s. [2]).
Wichtig ist, daß das Training nicht nur mit den Gesichtern erfolgt, sondern auch mit ausgewählten Negativbeispielen, die mit Hilfe eines Bootstrap-Algorithmus (vgl. [12]) ermittelt werden. Hierbei werden falsch positiv klassifizierte Bildregionen der Menge der Negativbeispiele zugefügt.
Ist beispielsweise aufgrund der Beleuchtungsbedingungen (z.B. Gegenlicht) oder der Entfernung (geringe Auflösung) die Gesichtsstruktur nicht deutlich erkennbar, ist das Ergebnis nicht sicher. Dieser Fall wird jedoch durch die Kombination mit der Kopf-Schulter-Kontur ausgeglichen.
Wie die folgenden Beispiele zeigen, werden nicht nur Gesichter klassifiziert (Abb.3). Ein wesentliches Merkmal für eine korrekte Lokalisation ist zum einen die lokal begrenzte Klassifikationsantwort und zum anderen die Korrespondenz richtiger Positionen auf benachbarten Ebenen.

Abb. 3: Detektion der Gesichtsstruktur: die Quadrate zeigen die wahrscheinlichsten Gesichtsregionen (>0.7), wobei ihre Größe der jeweiligen Auflösungsebene entspricht

3.3 DETEKTION DER KOPF-SCHULTER-KONTUR

Die Detektion der Kopf-Schulter-Kontur einer Person erfolgt mit Hilfe eines Arrangements steuerbarer Filter auf dem Grauwertbild, die in Form einer durchschnittlichen Kopf-Schulter-Silhouette plaziert werden. Die Klassifikation wird über ein spezielles Abstandsmaß der ermittelten Orientierungen zu diesem Prototyp berechnet. Die hier gezeigten Ergebnisse beruhen auf einer Filterung in 4 definierten Orientierungen (ausführliche Erläuterung in [3], s. Abb.4).



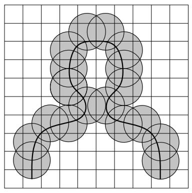


Abb. 4: Detektion der Kopf-Schulter-Kontur: die Lokalisation erfolgt nicht scharf auf einer der Auflösungsebenen (Zeilen), sondern auf korrespondierenden Positionen benachbarter Ebenen; 1.Spalte: Grauwertbild, 2.Spalte: Ergebnis der Filterung mit dem Filterarrangement (oben), 3.Spalte: Ergebnis der Selektion im dreidimensionalen Feld dynamischer Neurone

3.4 KOMBINATION DER VERFAHREN ZUR LOKALISATION

Die hier beschriebenen Verfahren bewirken abhängig von den Umgebungsbedingungen (Beleuchtung, Szeneninhalt, Entfernung zwischen Person und Roboter), die im voraus unbekannt sind, unterschiedlich sichere Ergebnisse zur Nutzerlokalisation. Die Ergebnisse aller Cue-Module werden

in dreidimensionale dynamische neuronale Felder eingekoppelt (s. [3]), um so mittels der nichtlinearen Relaxationsdynamik *eine* eindeutige Nutzerposition (Ort und Entfernung) zu ermitteln. Dabei werden sämtliche Ebenen der Auffälligkeitspyramiden in den Selektionsprozeß mit einbezogen.
Die Personenlokalisation anhand mehrerer Merkmale wird anhand des Beispiels aus Abb. 3 links gezeigt. Sowohl die Gesichtsstrukturdetektion als auch die Kopf-Schulter-Partie (Abb.4) kennzeichnet, unter anderem, den Kopf der Person. Das zur Nutzerlokalisation verwendete grobe Hautfarbmodell klassifiziert neben der Hautfarbe auch einige Hintergrundregionen (vgl. Abb. 2 rechts). Aus der Verknüpfung der 3 Cues kann die Position der Person gekennzeichnet und das personenspezifische Hautfarbmodell adaptiert werden, mit Hilfe dessen die Handdetektion und Gestenerkennung des Nutzers durchgeführt wird.
Das Ergebnis zeigt eine deutlichere Trennung der Person zum Hintergrund (Abb.5). In der Regel ist die Nutzerlokalisation der einzelnen Merkmale allerdings nicht so eindeutig wie hier dargestellt.

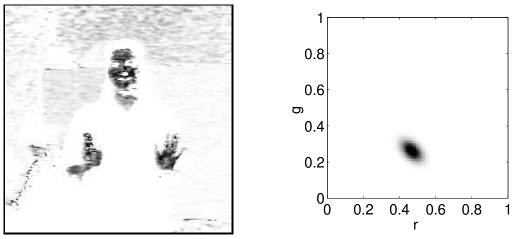


Abb. 5: Hautfarbklassifikation (l.) des auf die Person adaptierten Modells (r.): der Mittelwert des roten Farbanteils hat sich im Vergleich zum groben Modell erhöht und die Ausrichtung der Gaußfunktion hat sich geändert (vgl. Abb. 2)

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. Beispiele der für den Roboter relevanten Posen, sind im folgenden dargestellt (Abb.6). Diese Bilder stellen die Hautklassifikationsergebnisse des adaptierten Modells dar und sind auch ohne das entsprechende Originalbild leicht zu verstehen.

Abb. 6: Beispielposen (Hautfarbklassifikation): 'halt', 'fahre nach links', 'hallo (Aufmerksamkeit)'; weitere relevante Posen sind: 'fahre nach rechts', 'komme links/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 vorerst rein hautfarbbasiert auf der Basis des nutzerspezifischen Farbmodells.
Für jedes farbklassifizierte Gestenbild (vgl. Abb.6), werden 9 Momente $m_{p,q}$ (gewichtet mit einer 2-dim. Sigmoidfunktion) berechnet, die die Anordnung der Hände und des Kopfes beschreiben (vgl. [5]; $f(x,y)\widehat{=}$Pixelwerte): $m_{p,q} = \sum_x \sum_y x^p\, y^q\, f(x,y)$
Diese Momente werden mit Hilfe eines neuronalen Netzes (supervised GNG) als eine bestimmte Geste interpretiert. Dieses relativ einfache Verfahren ist möglich, da die Kopfposition bekannt und die Größe der Person im Bild einheitlich ist. Es wird jedoch vorausgesetzt, daß in einem gewissen Bereich im Bild nur noch Gesicht und Hände des Nutzers *deutlich* klassifiziert werden.
Eine Erweiterung auf die Bestimmung der Kopf- und Handorientierung als weiteres gestenrelevantes Merkmal wird angestrebt. Ein weiterer Schritt ist die Erkennung *dynamischer* Gesten, deren Ansatz bereits in Kapitel 3 kurz umrissen wurde (vgl. [11]).

5 ZUSAMMENFASSUNG

Es wurde ein System vorgestellt, welches auf der Basis von mehreren sich gegenseitig unterstützenden Merkmalen eine robuste Personenlokalisation (Gesicht) in Real-World-Umgebungen realisiert. Die Hautfarbklassifikation findet auf dem farbadaptierten Bild mit Hilfe eines intensitätsnormierten gaußförmigen 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, die durch ein räumlich verteiltes Arrangement steuerbarer Filter selektiert wird.
Durch die Verknüpfung über dynamische neuronale Felder soll die Position einer Person eindeutig bestimmt werden. Die sich anschliessende Handdetektion und Gestenerkennung wird rein hautfarbbasiert durchgeführt, indem das Hautfarbmodell auf den aktuellen Nutzer adaptiert wird. Die Gestenerkennung bezieht sich vorerst auf statische Posen, die durch die Relativpositionen von Kopf und Händen definiert sind.

Literatur

[1] Boehme, H.-J., Brakensiek, A., Braumann, U.-D., Krabbes, M., Gross, H.-M.: Neural Networks for Gesture-based Remote Control of a Mobile Robot. In: Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN'98), Anchorage, 1998

[2] Brakensiek, A., Braumann, U.-D., Corradini, A., Böhme, H.-J., Groß, H.-M.: Neuronale Verfahren zur Lokalisation und Bewertung von Handgesten in Real-World-Szenen. In: Proc. Neural Networks in Applications (NN'98), Magdeburg, pp. 211-218, 1998

[3] Braumann, U.-D., Corradini, A., Böhme, H.-J., Groß, H.-M.: Konturbasierte Personenlokalisation mittels dreidimensionaler neuronaler Felder und steuerbarer Filter. In: DAGM'98, Stuttgart, 1998

[4] Darrell, T., Basu, S., Wren, C., Pentland, A.: Perceptually-driven Avatars and Interfaces: active methods for direct control. MIT Media Lab, Technical Report No.416, USA, 1997

[5] Davis, J.W.: Appearance-Based Motion Recognition of Human Actions. MIT Media Lab, Technical Report No. 387, USA, 1996

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

[7] Kortenkamp, D., Huber, E., Bonasso, R.P.: Recognizing and interpreting gestures on a mobile robot. In: AAAI'96, 1996

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

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

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

[11] Rigoll, G., Kosmala, A., Eickeler, S.: High Performance Real-Time Gesture Recognition Using Hidden Markov Models. In: Proc. Gesture Workshop 97, Bielefeld, 1997

[12] Rowley, H. A., Baluja, S. Kanade, T.: Neural Network-Based Face Detection. In: IEEE PAMI'98, 1998

[13] Yang, J., Waibel, A.: A Real-Time Face Tracker. In: Proc. IEEE Workshop on Applications of Computer Vision (WACV'96), Sarasota, USA, 1996

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

Autorenangaben:
Dipl.-Ing. Brakensiek, Anja
Dipl.-Ing. Braumann, Ulf-Dietrich
Dr.-Ing. Böhme, Hans-Joachim
Prof. Dr.-Ing. Groß, Horst-Michael
Technische Universität Ilmenau, Fachgebiet Neuroinformatik
Postfach 100565, D-98684 Ilmenau
Tel.: +49 (0)3677/69-1307, Fax: +49 (0)3677/69-1665
(anja,ulf,hans,homi)@informatik.tu-ilmenau.de