Active-Vision zur Selbstorganisation von Verhalten in senso-motorischen Systemen

H.-J. Böhme, H.-M. Groß, U.-D. Braumann, D. Heinke, T. Pomierski,
Anja Brakensiek
Fachgebiet Neuroinformatik
Technische Universität Ilmenau
98684 Ilmenau (Thüringen)
e-mail: hans@informatik.tu-ilmenau.de
http://www.prakinf.tu-ilmenau.de/NeuroInf/ni.html

1 Einleitung und Motivation

Für ein in unbekannter Umgebung explorierendes System erscheint die Detektion von stabil auftretenden, sensorischen Situationen notwendig, da auf deren Basis bereits einfache Verhaltensweisen generiert werden können. Unter diesem Blickwinkel stellt sich der Problemkreis der Objekterkennung im Vergleich zum technischen Standpunkt für ein biologisches System völlig anders dar. Geht es bei der technischen Objekterkennung um das Auffinden meist bereits a priori bekannter und mehr oder weniger komplex kodierter Strukturen in Bildern ohne Verhaltensbezug, so besteht für das biologische System das Ziel vielmehr in der Extraktion verhaltensrelevanter Information (siehe auch [5]). Beispiele für solche Verhaltensweisen in Bezug auf Objekte könnten die Vermeidung von Hindernissen, die Bestimmung der eigenen Position, ein zielgerichteter Zugriff oder die Interpretation eines Objekts im Sinne einer Handlungsanweisung sein.
Vor diesem Hintergrund möchte der vorliegende Beitrag ein bereits realisiertes Modellkonzept (siehe auch [3], [4]) zur ***Selbstorganisation eines verhaltensorientierten Objektverständnisses***[1] vorstellen und die konzeptionelle Weiterentwicklung dieses Ansatzes zur ***Selbstorganisation eines verhaltensorientierten Posen-Verständnisses bei der natürlichen Mensch-Maschine-Kommunikation*** diskutieren.

2 Selbstorganisation eines verhaltensorientierten Objektverständnisses

Als "Umwelt" für das System dienten 100 stationäre Farbszenen, in denen verschiedene Objekte in verschiedenen Arrangements vor komplex strukturiertem Hintergrund auftraten (siehe beispielhaft Abbildung 4). Rotations- und Skalierungsaspekte wurden explizit nicht berücksichtigt. Die möglichen Aktionen des Systems in seiner Umwelt wurden an den Prozeß des internen Scannings angelehnt (siehe auch [6]). Die Selbstorganisation des emergenten Objektverständnisses soll in einer ***beobachtbaren Verhaltensänderung*** im Sinne einer charakteristischen zeitlichen Abtastung der Objekte (stabile visuelle Strukturen der Umwelt) resultieren. Dies bedeutet eine Umrangierung innerhalb des Abtastpfades, so daß zu einem Objekt gehörende Bildgebiete bevorzugt zeitlich unmittelbar aufeinanderfolgend abgetastet werden, wie dies in Abbildung 1 rechts in schematischer Form zu sehen ist. Die Umrangierung der zu einem Objekt gehörenden Bestandteile in die unmittelbare zeitliche Nachbarschaft beim Analyseprozeß stellt nach unserer Auffassung eine elementare Verarbeitungsleistung dar, die einen wesentlichen Beitrag zur Lösung des Binding-Problems liefern kann. Weiterhin ermöglicht eine Umgruppierung elementarer Szenenbestandteile zu unterschiedlichen Bedeutungsinhalten eine kontinuierliche Modifikation des Systemwissens. Dies garantiert, daß eine Interpretation des aktuellen Inputs unter Nutzung des bereits vorhandenen Systemwissens überhaupt erst möglich wird.
Der durch die selektive Aufmerksamkeit realisierbare wahlweise Zugriff auf elementare Szenenbestandteile sichert dem System die Flexibilität, die notwendig ist, um beim Interpretationsprozeß eine kombinatorische Explosion möglicher Entscheidungsvarianten zu vermeiden.

Abbildung 2 rechts zeigt die neuronale Architektur, die in Anlehnung an die links dargestellten cortikalen und subcortikalen Verarbeitungspfade entstand. Das ***Attention Control Network (ACN)*** erstellt in Verbindung mit dem

[1] gefördert durch das Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF), Förderkennzeichen: 413-5839-01 IN 101D - NAMOS-Projekt

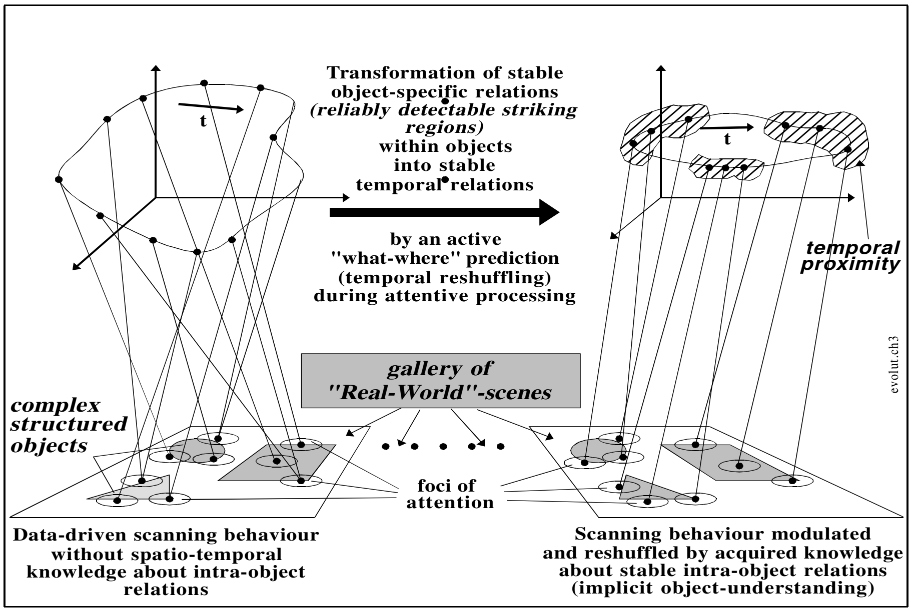

Abbildung1: *Umrangierung bedeutungsmäßig zusammengehörender Szenenbestandteile in eine unmittelbare zeitliche Nachbarschaft beim Analyseprozeß*

Saliency System (SS) eine Auffälligkeitskarte der aktuellen Szene. Diese Auffälligkeitskarte bildet zunächst die alleinige Basis für den Abtastprozeß, da der Aufmerksamkeitsfocus (Searchlight-Metapher, [1]) ohne Wissen nacheinander nur auf auffällige Bildregionen mit absteigender Auffälligkeit gerichtet wird. Dem *ACN* obliegt weiterhin die Aufgabe, die attentive und die präattentive Steuerung des Abtastverlaufs zu koordinieren.
Um einen schnellen Übergang von der frühen visuellen Verarbeitung zu einer an den *IT* angelehnten lokalen Repräsentation eines endlichen Satzes komplexer Merkmale [2] zu vollziehen, wurde im Modell ein *Attentional Focus Identifier (AFI)* eingeführt. Der *AFI* generiert eine Beschreibung der gerade fokussierten Bildregion in Form eines komplexen Merkmals, welches die Information aus den verschiedenen Merkmalskarten (Farbe, Textur, usw.) bezüglich dieser Region zusammenfaßt. An den *AFI* schließt sich unmittelbar der funktionell sowohl an den *IT* als auch an den PP angelehnte *Feature Transition Memory (FTM)* an, welcher die eigentliche Basis einer raum-zeitlichen Objektrepräsentation innerhalb der Modellarchitektur darstellt.

Der *Feature Transition Memory* bildet die erste attentive Ebene der Modellarchitektur, in der ausschließlich lokales Wissen über stabil auftretende Übergänge zwischen unmittelbar aufeinanderfolgend abgetasteten Focusregionen sowie deren räumlichen Relationen (Was?-Wo?) gebildet wird. Die Grundlage dafür bildet ein permanentes *Reinforcement-Lernen*, das korrekt vorhergesagte Abtastbewegungen - bezüglich Inhalt und Position der erwarteten nachfolgenden Focusregion - belohnt. Da sich Übergänge auf Objekten (Intraobjekt-Beziehungen) gegenüber Übergängen zwischen Objekten (Interobjekt-Beziehungen) oder zwischen Objekten und Hintergrund durch Reproduzierbarkeit auszeichnen, gelingt es im Verlauf des Lernprozesses mehr und mehr, Übergänge auf Objekten bei der Abtastung zu bevorzugen.
Der *Episodic Object Memory (EOM)*, der sich am *Präfrontalen Assoziationscortex* orientiert, bildet die zweite attentive Ebene zur Repräsentation von Wissen innerhalb der Modellarchitektur. Ihre Notwendigkeit ergibt sich aus der Tatsache, daß im *FTM* ausschließlich stabile Übergänge zwischen unmittelbar aufeinanderfolgenden Fokusregionen gelernt und zur Steuerung des Abtastverlaufs verwendet werden. Dadurch fehlt dieser Ebene noch das globale Wissen über größere Bildzusammenhänge. Der *EOM* versucht nun, gerade dieses objektspezifische Wissen zu extrahieren, indem nur unmittelbar aufeinanderfolgend korrekt prädizierte und damit mit großer Sicherheit objekttypische Übergänge zwischen den Fokusregionen gelernt werden. Da der *EOM* die Verarbeitung im *FTM* moduliert, fließen jetzt das unspezifische Wissen der ersten attentiven Ebene (lokale Übergangshypothesen) und das objektspezifische Wissen der zweiten attentiven Ebene (globale Hypothesen) im *FTM* zusammen. Schematisiert ist der Prozeß der Einflußnahme des EOM auf den FTM in der Abbildung 3 dargestellt. Im Ergebnis dieser Rückwirkung soll sich die Auswahl der nächsten korrekten Sprunghypothese im FTM deutlich beschleunigen, da zur Erzeugung einer gemeinsamen Hypothese beider attentiven Ebenen jetzt lokales und globales Wissen zusammenfließt. Unter Beschleunigung der Hypothesenselektion im FTM soll dabei verstanden werden, daß sich die Anzahl der nicht verifizierbaren Sprunghypothesen im Vergleich zur ausschließlichen Nutzung von lokalen Hypothesen der ersten attentiven Ebene deutlich reduziert, was die Simulationsbeispiele in Abbildung 4 auch belegen.

Anhand der exemplarischen Simulationsbeispiele in Abbildung 4 wird deutlich, wie sich das Abtastverhalten im

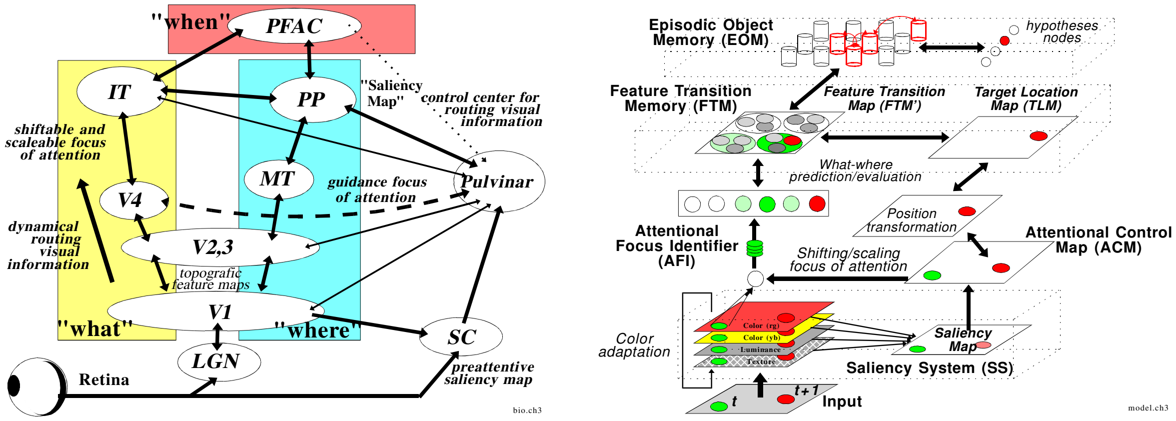

Abbildung2: **(Links)** *Die beim Architekturentwurf berücksichtigten biologischen Verarbeitungsstrukturen: LGN-Lateral Geniculate Nucleus, SC-Superior Colliculus, visuelle Areale V1-V4, IT-Inferotemporaler Cortex, MT-Medial temporaler Cortex, PP Postparietaler Cortex, PFAC-Präfrontaler Assoziationscortex* **(Rechts)** *Schematische Darstellung der wesentlichen Subsysteme der Modellarchitektur*

Verlaufe des Selbstorganisationsprozesses von einer anfänglich rein auffälligkeitsbasierten, vergleichsweise uneffektiven Abtastung hin zu einer effektiven Abtastung der in den Szenen enthaltenen Objekte entwickelt. Dazu zeigt die ***obere Reihe von links nach rechts*** den Abtastverlauf für ein und dieselbe Szene in unterschiedlichen Stadien des Lernprozesses (zu Beginn, nach Präsentation von 20, 30, 40 und 90 weiteren Szenen aus der Galerie). Die schwarz dargestellten Verschiebungen des Aufmerksamkeitsfocus verdeutlichen den Verlauf der datengetriebenen Suche (Sequenz auffälliger lokaler Bildstrukturen), wissensbasierte Abtastbewegungen sind weiß markiert. Ohne Wissen ist das System nicht in der Lage, Zusammengehörigkeitshypothesen über den Pinguin oder die elliptische Ringstruktur (links oben) aufzustellen und damit einen effektiven Abtastverlauf zu antizipieren ("Was-Wo-Wann"). Im Verlauf des Selbstorganisationsprozesses werden allmählich stabile lokale und globale Zusammenhänge durch FTM und EOM erfaßt und erfolgreich prädiziert. Als Resultat zeigt die Abbildung ***oben rechts*** einen erfolgreich prädizierten wissensbasierten Abtastverlauf, nachdem alle Szenen der Galerie ein einziges mal präsentiert wurden. Die beiden Szenen ***(unten links)*** zeigen einen Vergleich des Abtastverhaltens über einer weiteren Szene aus der Galerie vor und nach dem Selbstorganisationsprozeß. Die Abbildungen ***unten rechts*** verdeutlichen den Einfluß lokaler und globaler Hypothesen auf die raum-zeitliche Abtastdynamik: (links) bei alleiniger Nutzung lokaler Hypothesen aus dem FTM und (rechts) mit top-down Modulation der lokalen Hypothesenbildung durch globale Hypothesen aus dem EOM. Die Anzahl der dünn weiß markierten, nicht erfolgreichen Suchsprünge, die ausschließlich aufgrund lokaler Hypothesen vorgeschlagen, aber nicht verifiziert werden konnten, wurde durch den Einfluß der globalen Hypothesen deutlich reduziert.

3 Selbstorganisation eines verhaltensorientierten Posen-Verständnisses bei der visuellen Mensch-Maschine-Kommunikation

Gab es bei dem bisher beschriebenen Modell neben der Verschiebung des Aufmerksamkeitsfocus keine weitere Interaktion mit der Umwelt (effiziente Abtastung der Objekte selbst besitzt noch keine weiterreichende Handlungsrelevanz), so soll das verfolgte Konzept in derzeit laufenden Arbeiten auf ein komplexeres Verhaltensszenario, die natürliche Interaktion unserer mobilen Roboterplattform mit einem potentiellen Nutzer, übertragen werden.
Dabei ergeben sich einige konzeptionelle Veränderungen, die im folgenden kurz diskutiert und teilweise mit ersten Simulationsergebnissen unterlegt werden.
Im Rahmen dieses Beitrages ist eine umfassende Darstellung aller relevanten Eckpunkte unmöglich, so daß nur ausgewählte Aspekte kurz diskutiert werden können.

Abtastverhalten Die linken drei Abbildungen in Abbildung 5 zeigen für das aufmerksamkeitsbasierte Sehsystem des Roboters die angestrebten, typischen auffälligkeitsbasierten Abtastverläufe (schwarz markiert).
Da das Sehsystem noch kein Wissen über objektspezifische Zusammenhänge besitzt, ist es auch nicht in der Lage, die z.B. zu einer unbekannten Positur (als verhaltensrelevantes "Objekt Fahrtrichtungsänderung") gehörenden auffälligen Bildstrukturen (Hände, Gesichter, evt. Armwinkel usw.) mittels seines Aufmerksamkeitsfocus unmittelbar nacheinan-

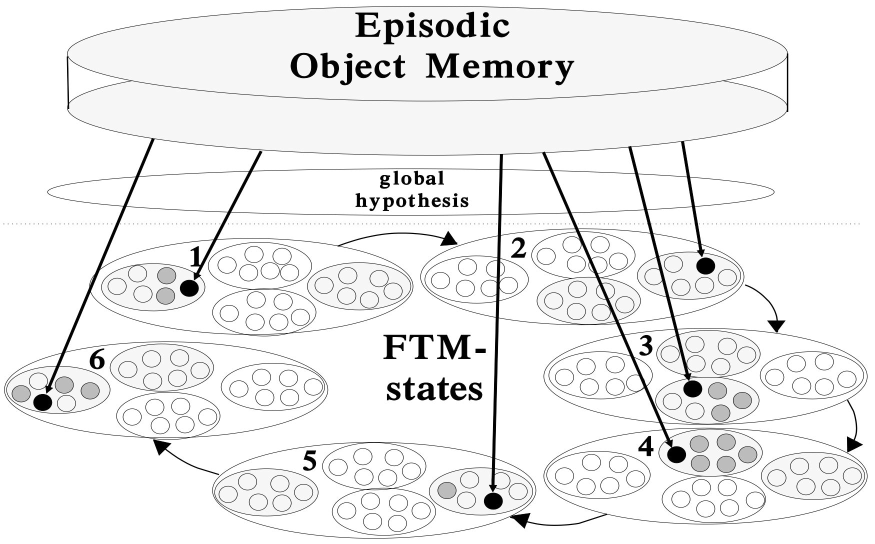

Abbildung3: *Darstellung von sechs aufeinanderfolgenden Etappen der Einflußnahme des Episodic Object Memory auf die Selektion der Sprungkandidaten im FTM der ersten attentiven Ebene; grau hervorgehoben sind die gerade aktiven FTM-Cluster, die im Ergebnis der Analyse der Fokusregionen durch AFI und TFI aktiviert werden; mittelgraue Knoten bedeuten lokale Hypothesen, die ausschließlich aufgrund des innerhalb der ersten attentiven Ebene akkumulierten Umweltwissens selektiert worden wären; schwarz sind die Knoten markiert, an denen durch Zusammenwirken von lokalen und globalen Hypothesen der ensprechende Sprungkandidat ermittelt wurde*

der abzutasten. Daher ergeben sich zahlreiche datengetriebene Sprünge zwischen den Objekten (und u.U. auf auffällige Bildstrukturen im Hintergrund). Ziel des Selbstorganisationsprozesses ist es, nur ***die stabilen und gleichzeitig verhaltensrelevanten (und damit erfolgreich prädizierbaren) "Intraobjekt-Beziehungen"*** zu extrahieren und intern so zu repräsentieren, daß der Abtastprozeß wissensbasiert so umrangiert werden kann, daß die objektspezifisch zusammengehörigen Bildstrukturen zeitlich unmittelbar nacheinander und damit besonders effektiv abgetastet werden. Eine solche Veränderung des Abtastverlaufs wäre dann der beobachtbare Ausdruck der erfolgreichen Selbstorganisation eines verhaltensbasierten Objektverständnisses, wie dies in Abbildung 5 in der rechten Szene als angestrebtes Ergebnis verdeutlicht wird.

Auffälligkeitsberechnung Da die natürliche Interaktion unserer mobilen Roboterplattform mit einem potentiellen Nutzer ein sehr komplexes Verhaltensszenario darstellt, erscheint es uns notwendig, möglichst schnell die Ebene der eigentlichen Selbstorganisation eines verhaltensorientierten Posenverständnisses zu erreichen. Vor diesem Hintergrund seien die für die Posenrepräsentation und -interpretation als wesentlich angenommenen Bildstrukturen – Gesichter und Hände – a priori auffällig. Dazu wird das aktuelle Eingangsfarbbild auf allen Ebenen einer Auflösungspyramide hinsichtlich ***Hautfarbe*** und ***Gesichtsstruktur*** analysiert. Die Analyseergebnisse sollen Auffälligkeitskarten bilden, in denen hautfarbene Regionen und insbesondere Gesichter stark hervortreten.

Erste Ergebnisse zur farbbasierten Detektion von Hautregionen zeigt Abbildung 6, wobei im gezeigten Beispiel ein vortrainiertes GNG-Netzwerk [7] zum Einsatz kam. Zur grauwertbasierten Gesichtsdetektion wird ein einfacher Matched-Filter-Ansatz angestrebt. Farbe und Struktur sollen dabei als sich ergänzende Informationsqualitäten angesehen werden, um keine vorrangige Abhängigkeit des Systemverhaltens von einer der beiden Qualitäten zu erzeugen. In welcher Art und Weise beide Informationen verknüpft werden, ist ein Gegenstand aktuell laufender Arbeiten. Auf weitere Details muß an dieser Stelle verzichtet werden, da die Untersuchungen zu beiden Verfahren noch nicht abgeschlossen sind und sich somit permanent Änderungen ergeben.

Repräsentation von Posen Die Art und Weise der Wissensrepräsentation wird sich stark an den bereits beschriebenen Ansatz im Sinne einer Interaktion von Was-, Wo- und Wann-System anlehnen. Dabei soll die Trennung der

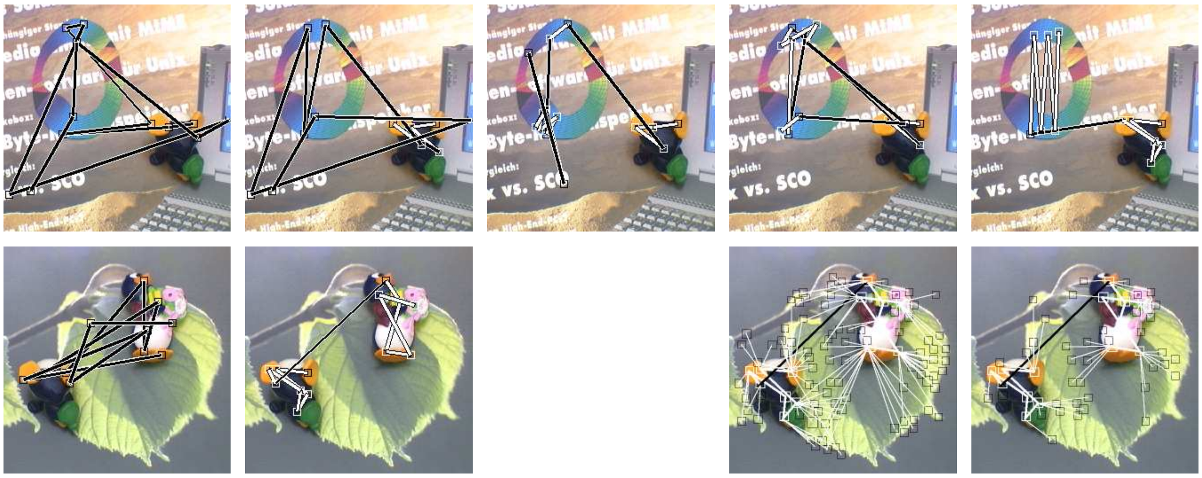

Abbildung4: *Simulationsergebnisse (Erläuterung siehe Text)*

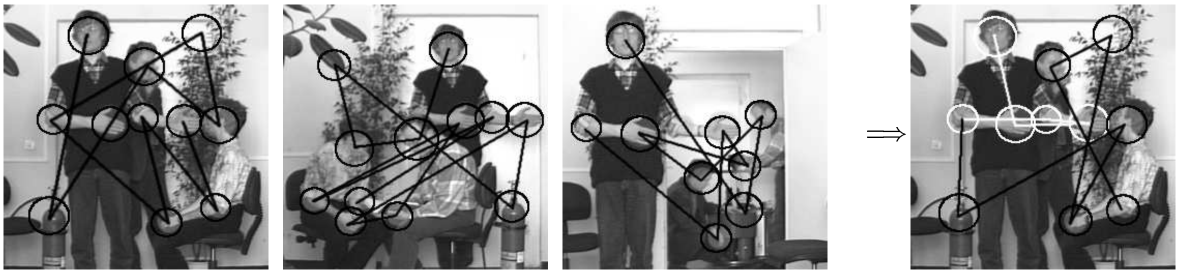

Abbildung5: *Szenen aus dem Verhaltensszenario zur natürlichen Nutzer-Roboter-Interaktion*

beiden attentiven Ebenen zugunsten einer durchgehenden Modellarchitektur aufgegeben werden. Die zur Anwendung kommenden Lernverfahren werden sich wie bei dem in Abschnitt 2 beschriebenen Modell (siehe auch [9]) stark an Reinforcement-Ansätzen orientieren, wobei derzeit zusätzlich ein Verfahren implementiert und untersucht wird, welches ein entropiebasiertes Reinforcement-Signal zur Steuerung des Abtastverhaltens anwendet [8]. Ziel ist dabei, die nächstfolgende Abtastbewegung zu initiieren, die zu einem Szenenmerkmal führt, welches einen maximalen Informationszuwachs hinsichtlich der angestrebten Unterscheidung von Objekten liefert.

Weiterhin ist vorgesehen, kontinuierlich aufeinanderfolgende Szenen in Feldern aus dynamischen (T1-)Neuronen zu verarbeiten, um so in mehreren unmittelbar aufeinanderfolgenden Szenen auftretende, stabile Konstellationen von Gesicht und Händen eines potentiellen Nutzers als Auffälligkeitskriterium detektieren zu können.

Aus dem angestrebten ***Erlernen der Verhaltensrelevanz von Posen*** ergeben sich jedoch im Vergleich zu dem in Abschnitt 2 beschriebenen Modell zusätzliche Anforderungen an die räumlich-zeitliche Repräsentation einer Pose, da diese zumindest so lange stabilisiert und im System gehalten werden muß, bis nach der Ausführung der vom System mit dieser Pose assoziierten Aktion entschieden werden kann, ob diese Pose

a) überhaupt Handlungsrelevanz besitzt, also mit einer bestimmten Aktion zu assoziieren ist, und ob

b) die assoziierte Aktion richtig oder falsch war.

Verhaltensrelevanz von Posen Wurde im NAMOS-Projekt die Stabilität von Merkmalsbeziehungen im Sinne der Statistik ihres Auftretens als vorrangiges Kriterium für den Selbstorganisationsprozeß genutzt, so ist im Kontext einer Interaktion zwischen System und Nutzer die Stabilität der auftretenden Konstellationen von Gesicht und Händen zwar Voraussetzung für die Verhaltensrelevanz einer solchen möglichen Pose, Stabilität dieser Konstellationen kann jedoch nicht das ***alleinige*** Kriterium für eine wirkliche Handlungsrelevanz einer Pose darstellen. Vielmehr erachten wir es als unbedingt notwendig, daß das System die Handlungsrelevanz einer möglichen Pose durch die unmittelbare

Abbildung6: *An zwei Beispielszenen soll der aktuelle Stand der farbbasierten Detektion von Hautregionen veranschaulicht werden. Links sind jeweils die Originalszenen dargestellt, rechts die dazu korrespondierenden Hauthypothesen.*

Interaktion mit dem Nutzer und der Umwelt erfährt (siehe auch [10]), was wiederum die unmittelbare Bewertung der durch das System aufgrund einer möglichen Pose ausgeführten Handlung voraussetzt. Diese Bewertung soll, um die Komplexität des gesamten Ansatzes nicht weiter explodieren zu lassen, vorab definiert werden. Denkbar wäre z.B. , bestimmte Posen, die im System als Vorwissen abgelegt sind, fest mit einem positiven bzw. negativen systeminternen Reinforcementsignal zu verknüpfen. Weiterhin wird in diesem Zusammenhang die Nutzung taktiler und akustischer Information untersucht werden.

Als ein geeignetes Szenario für das ***Erlernen der Verhaltensrelevanz von Posen*** könnte beispielsweise die ausschließlich posenbasierte Navigation des Systems durch einen (zunächst rein simulativ vorgegebenen) Hindernisparcours dienen. Damit wäre das Systen gezwungen, jede "erkannte" Pose unmittelbar auf ihre Handlungsrelevanz hin zu überprüfen . Die Bewertung der aktuell ausgeführten Handlung ergäbe sich entweder aus der Information bezüglich Kollision/Nichtkollision (als Reinforcement-Signal aus der Umwelt) bzw. über ein adäquates, vordefiniertes Reinforcement-Signal, welches vom Nutzer, mit dem das System gerade interagiert, selbst bereitgestellt werden muß.
Aus diesen Überlegungen lassen sich aus unserer Sicht derzeit zwei mögliche, grobe Strategien zur Selbstorganisation eines verhaltensorientierten Posenverständnisses ableiten:

1. Zunächst erfolgt in einem überwiegend statistikbasierten Lernprozeß die Repräsentation möglicher ***Posenkandidaten*** (räumlich stabile, häufig wiederkehrende Konstellationen von Gesicht und Händen). Daran schließt sich in einem zweiten Schritt die Überprüfung der bislang ermittelten Posenkandidaten bezüglich deren Handlungsrelevanz im Rahmen der Interaktion zwischen System und Nutzer an.

2. Die Repräsentation handlungsrelevanter Posen erfolgt sofort ausschließlich auf der Basis der unmittelbaren System-Nutzer-Interaktion, ohne daß das System vorher wie in Strategie 1 eine passive, beobachtende Phase durchläuft. Auch hier wird jedoch die Statistik des mehrmaligen Auftretens einer mit einer bestimmten Handlung

zu verknüpfenden Pose für die letztendliche Repräsentation mitentscheidend sein.

Welcher der beiden Strategien der Vorzug zu geben ist, müssen weiterführende Untersuchungen und konzeptionelle Überlegungen erbringen.

Literatur

[1] **Treisman, A. (1988).** Features and Objects: The Fourteenth Bartlett Memorial Lecture. In: *The Quarterly Journal of Experimental Psychology*, pp. 201-237

[2] **Fujita, I. et al. (1992).** Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, Vol. 360, pp. 343-346

[3] **Gross, H.-M. et al. (1995).** Elementare kognitive Mechanismen in neuronalen Architekturen - Selbstorganisation eines verhaltensbasierten Objektverständnisses. *Abschlußbericht zum BMFT-Verbundprojekt NAMOS*, Schriftenreihe FG Neuroinformatik, ISSN 0945-7518, TU Ilmenau

[4] **Gross, H.-M.; Heinke, D.; Boehme, H.-J.; Braumann, U.-D.; Pomierski, T. (1995).** A Behaviour-oriented Approach to an Implicit "Object-understanding" in Visual Attention. *Proc. of ICNN'95, IEEE-Intern. Conference on Neural Networks 1995, Perth,*, pp. 657-662, IEEE Press

[5] **Mallot, H.A., Kopecz, J. und von Seelen, W. (1992).** Neuroinformatik als empirische Wissenschaft. *Kognitionswissenschaft*, Vol. 3, pp. 12-23

[6] **Olshausen, B.; Andersen, C. & van Essen, D. (1993).** A Neural Biological Model of Visual Attention and Invariant Pattern Recognition. *Journal of Neuroscience.* Vol. 13, pp. 4700-4719

[7] **Fritzke, B. (1995).** A Growing Neural Gas Network Learns Topologies. *Advances in Neural Information Processing Systems 7*, MIT Press

[8] **Bandera, C., Vico, F.J., Bravo, J.M., Harmon, M.E., Baird, L.C. (1996).** Residual Q-Learning Applied to Visual Attention. *Proc. of the Thirteenth Int. Conf. on Machine Learning, Bari, Italy*

[9] **Heinke, D., Gross, H.-M. (1993).** A Simple Selforganizing Neural Network for Selective Visual Attention. *Proc. of ICANN-93, Amsterdam, pp. 63-66*

[10] **Hamker, F.H., Gross, H.-M. (1996).** Intentionale Aufmerksamkeit: Ein alternatives Konzept für technische visuo-motorische Systeme. *this volume*