Eingereicht zum 14. Fachgespräch Autonome Mobile Systeme '98
30. November - 1. Dezember 1998, Karlsruhe
AMS'98 - 14. Fachgespräch Autonome Mobile Systeme, Karlsruhe, 1998, pp. 85-92, Springer-Verlag 1998

Monokulare visuelle Hindernisdetektion auf Basis merkmalsbasierter Bildsegmentierung

M. Krabbes[2], S. Weber[1], V. Stephan[1], H.-J. Böhme[1], H.-M. Groß[1]

[1] Technische Universität Ilmenau
Fachgebiet Neuroinformatik*
stefan.weber@rz.tu-ilmenau.de
{vstephan,hans,homi}@informatik.tu-ilmenau.de

[2] Otto-von-Guericke-Universität Magdeburg
Institut für Automatisierungstechnik
krabbes@infaut.et.uni-magdeburg.de

Zusammenfassung Dieser Beitrag stellt eine Architektur vor, die kollisionsvermeidende Roboternavigation auf Basis monokularer visueller Information in Indoor-Umgebungen ermöglicht. Dies gelingt durch neuronale Verknüpfung von Bildmerkmalen mit Tiefeninformation, die sich ebenfalls aus dem visuellen Datenstrom extrahieren läßt. Basierend auf einem ersten Ansatz, der eine Segmentierung von Kamerabildern in *Hindernis* und *befahrbaren Untergrund* ausschließlich mit Hilfe geeignet trainierter Neuronaler Gase durchführt, wird ein weiterführender Ansatz vorgestellt, der verschiedene Segmentationsverfahren zu einem stabilen Gesamtergebnis verarbeitet. Die Leistungsfähigkeit des Verfahrens wird dargestellt, Probleme im realen Einsatz werden beschrieben und entsprechende Lösungsansätze aufgezeigt.

1 Einleitung

Bei der visuellen lokalen Navigation mobiler Roboter konnten durch Einsatz neuronaler Verarbeitungsstrukturen bereits vielversprechende Erfolge erzielt werden. Das Grundprinzip besteht dabei häufig in einer direkten Feed-Forward-Verarbeitung bestimmter Bildmerkmale zu einem Fahrzeugsteuerbefehl. Da aus einem Einzelbild jedoch keine Tiefeninformation extrahierbar ist, sind diese Verfahren auf stabile Umwelteigenschaften angewiesen, die auch für den visuellen Sensor die Unterscheidung zwischen *befahrbarem Untergrund* und *Hindernis* ermöglicht. Während diese Bedingung für Straßenfahrzeuge noch weitgehend erfüllt wird [7], sind in Indoor-Umgebungen Zusammenhänge zwischen visuellen Umweltmerkmalen und nutzbaren Fahrwegen kaum noch mit der notwendigen Stabilität zu gewährleisten [2, 4]. Um den visuellen Sensor weiter für die lokale Roboternavigation zu erschließen, ist es deshalb notwendig, die nutzbaren Bildmerkmale adaptiv mit der zur Verfügung stehenden Entfernungsinformation zu verknüpfen.

* Die Arbeiten sind Teil des vom Thüringer Ministerium für Wissenschaft, Forschung und Kultur (TMWFK) geförderten Projektes GESTIK.

Mit der hier vorgestellten Hybridarchitektur (Abb. 1) wird versucht, die über die Fahrzeugbewegung aus Bildfolgen zu gewinnende Tiefeninformation mittels einer lernfähigen Struktur auf andere Bildmerkmale (wie z.B. Farbe und Struktur) abzubilden. Auf diese Weise soll sich aus jeder Bildregion zu jedem Zeitpunkt eine zur kollisionsfreien Navigation ausreichende Tiefeninformation extrahieren lassen. Dafür müssen Modellannahmen über visuelle Umwelteigenschaften herausgebildet werden, die *Hindernisse* vom *Untergrund* unterscheidbar machen.
Die Extraktion der benötigten Tiefeninformation wird durch die Anwendung der Methode des optischen Flusses auf invers-perspektivisch kartierte Bilder erreicht [1], mit der auf Basis der lokalen Bildverschiebungen für einen Bildpunkt die (binäre) Unterscheidung gewonnen werden kann, ob dieser auf der Bewegungsebene liegt ($\widehat{=}$ *Untergrund*) oder ob er zu dieser Ebene einen Höhenoffset besitzt ($\widehat{=}$ *Hindernis*).

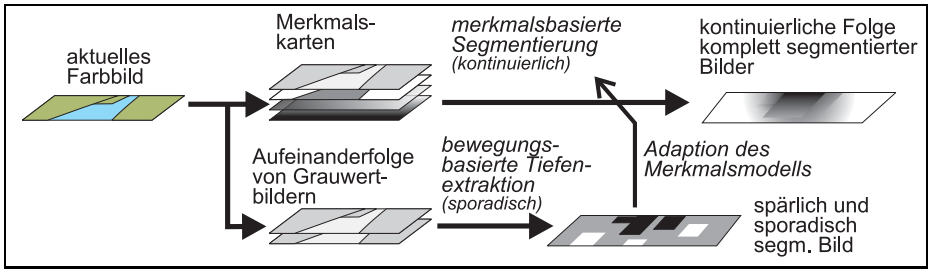

Abbildung 1. *Prinzipdarstellung der hier vorgestellten Hybridarchitektur: Eine kontinuierliche merkmalsbasierte Untergrundsegmentierung wird sporadisch an sich ändernde Umgebungseigenschaften adaptiert.*

2 Architektur

2.1 Invers-perspektivische Kartierung

Die monokulare visuelle Navigation eines autonomen mobilen Roboters basiert auf der Abbildung des Fahrweges durch eine am Roboter (geneigt) montierte Kamera. Durch die sogenannte invers-perspektivische Kartierung werden die perspektivischen Verzerrungen in der Bildebene der Kamera bezüglich einer zweidimensionalen Bezugsebene (in diesem Fall der Bewegungsebene des Roboters) eliminiert, wodurch sich eine virtuelle Aufsicht auf diese Ebene ergibt (Abb. 2). Die Korrektheitsbedingung für die nachfolgend erläuterte Transformationsvorschrift entspricht genau der Voraussetzung von Hindernisfreiheit für ein fahrendes mobiles System. Da ein aus der Bewegungsebene herausragendes Hindernis dagegen verzerrt abgebildet wird, erzeugt es in einer Aufeinanderfolge mehrerer invers-perspektivisch transformierter Bilder gegenüber dem Untergrund einen höheren Geschwindigkeitsbetrag und kann auf diese Weise segmentiert werden [5]. Die Abbildung der Punkte der Bildebene p_H zurück auf die Bewegungsebene p_B läßt sich unter der Randbedingung, daß alle Bildpunkte von Punkten auf dieser Bewegungsebene stammen, in homogenen Koordinaten mathematisch korrekt durch eine umkehrbare Transformationsvorschrift beschreiben (Kamera nur geneigt):

$$p_H \;=\; p_B \, \mathbf{R}_u(\omega) \, \mathbf{P}_{ZP} \, \mathbf{T}_{OA} \qquad (1)$$

wobei die Matrix $\mathbf{R}_u(\omega)$ die Drehung des Koordinatensystems zur Kompensation der Kameraneigung ω, die Matrix $\mathbf{P}_{ZP}$ die Zentralprojektion des Bildes zurück auf die Bewegungsebene und die Matrix $\mathbf{T}_{OA}$ die Verschiebung des Koordinatensystems entlang der optischen Achse beinhaltet [1] (Abb. 2). Zusätzlich zu den benötigten Erweiterungen zur Skalierung und Rücktransformation [1] wurde diese Berechnungsvorschrift mit einem verzeichnungskompensierenden Term ergänzt [3].

$$r' = \frac{r}{1 + k|r|^2} \tag{2}$$

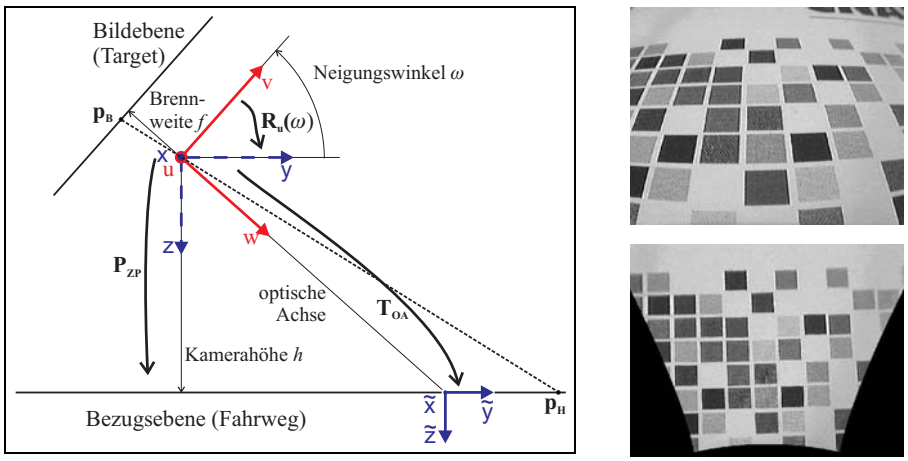

Abbildung 2. *Prinzip und Wirkung der invers-perspektivischen Kartierung. Links: Projektionsanordnung. Durchstoßpunkt eines Projektionsstrahls mit der Bildebene: p_B; mit der Bezugsebene: p_H; r. oben: Originalbild, Kamera 45° geneigt; r. unten: nach Kompensation der Kameraverzeichnung und Elimination der perspektivischen Verzerrungen ergibt sich eine virtuelle Aufsicht.*

2.2 Zur Auswertung des optischen Flusses

Für den Vergleich des aktuellen Bildes mit dem bzgl. der relativen Eigenbewegung des Roboters kompensierten vorherigen Bild wird ein korrelatives Verschiebungsdetektionsverfahren (optischer Fluß) verwendet. Die Ermittlung der lokalen Bildverschiebungen ist insbesondere in schwach strukturierten Bildregionen stark fehlerbehaftet und eine kontinuierliche Berechnung aufwendig. Bei einer monokularen Konfiguration sind außerdem ständige Roboterbewegungen notwendig. Deshalb soll in dem hier vorgestellten Verfahren der sporadische Vergleich aufeinanderfolgender Bilder (lediglich) der Extraktion von Trainingsbeispielen für ein darauf aufbauendes merkmalsbasiertes Segmentierungsverfahren dienen. Für dieses anschließende Training ist aber eine Selektion der ermittelten Verschiebungen nach deren Vertrauenswürdigkeit notwendig. Es werden nur diejenigen lokalen Verschiebungsvektoren weiterverwendet, bei denen die Differenz der minimalen *Summe der Absoluten Differenzen (SAD)* zwischen korrespondierenden Pixeln zur nächsthöheren SAD einen Mindestwert übersteigt [8]. Untersuchungen zur Realisierung einer solchen Hybridarchitektur für die visuelle Roboternavigation fanden auf zwei verschiedenen Systemen statt:

– Zunächst wurde der Miniaturroboter KHEPERA (Abb. 3 links) in einer speziellen Umgebung mit strukturiert bedruckten Oberflächen eingesetzt. Da dessen Bewegung zwischen zwei Bildaufnahmen nur begrenzt genau steuerbar ist (Antriebsregelung, Moment der Bildaufnahme usw.), verwendet das Verfahren des optischen Flusses hier das aktuelle Bild und ein in der Relativbewegung des Roboters kompensiertes vorheriges Bild.
– Bei dem zweiten eingesetzten Robotersystem MILVA (Abb. 3 rechts), das in realen Indoor-Umgebungen operieren soll, läßt sich die Relativbewegung zwischen zwei aufeinanderfolgenden Bildaufnahmen so genau steuern, daß sie durch eine entsprechende Bildverschiebung immer hinreichend genau kompensierbar ist und das jeweilige Differenzbild verwendet werden kann. Jedoch ist auch hier die gewonnene Information nur in strukturierten Bildregionen nutzbar, weshalb ebenfalls eine SAD-basierte Selektion mit dem oben beschriebenen Verfahren erfolgt. In diesem Falle wird jedoch in einem kleinen Suchfenster eines der beiden Bilder mit sich selbst verglichen und das Ergebnis mit dem Differenzbild für eine Segmentation zusammengeführt.

Die anschließende Zuordnung der Verschiebungsvektoren bzw. Pixeldifferenzen zu einer der beiden Klassen *„Untergrund“* oder *„Hindernis“* basiert auf plausiblen Annahmen: alle geringfügigen Veränderungen zwischen bewegungskompensierten Aufnahmen werden als *„Untergrund“* klassifiziert, Vektoren mit erheblicher vertikaler Verschiebung bzw. alle erheblichen Pixeldifferenzen lassen sich der Klasse *„Hindernis“* zuordnen, da hier der oben beschriebene Effekt der größeren Verschiebung vorliegt. Diese Segmentation durch ein Verfahren des optischen Flusses wird mit den Segmentationsergebnissen anderer Verfahren zusammengeführt.

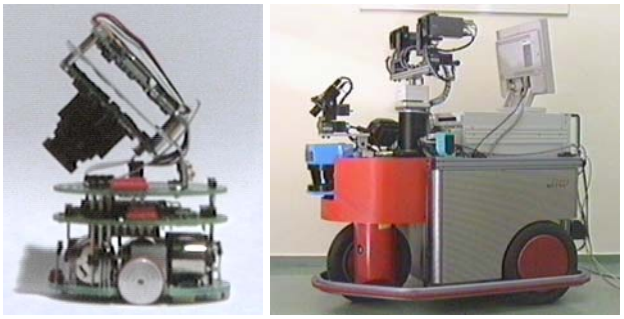

Abbildung 3. *Die verwendeten Roboterplattformen. Links: Miniaturroboter* KHEPERA *mit geneigt montiertem Kameramodul; rechts: mobile Roboterplattform* MILVA *des Fachgebiets Neuroinformatik (Navigationskamera: Mitte vorn).*

2.3 Adaptives Merkmalsmodell

Die Merkmale der beiden zu unterscheidenden Klassen werden durch je ein Neuronales Gas (NG) [6] repräsentiert. Die Anwendung eines separaten, unüberwachten Vektorquantisierers je Klasse erwies sich leistungsfähiger als eine einzelne, überwacht lernende Architektur mit Ausgabeschicht. Es wird dadurch eine Entkopplung von der Statistik der Trainingsmengen in den beiden Klassen erreicht. Dies wirkt sich hier vorteilhaft aus, da die Anzahl der aus einem einzelnen Bild extrahierbaren Trainingsbeispiele je

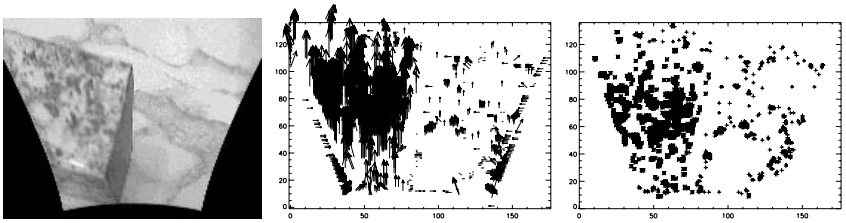

Abbildung 4. *Ergebnisse der Verarbeitungsschritte auf* KHEPERA. *Links: eines der beiden Bilder zur Verschiebungsberechnung: deutlich ist der Hindernisquader mit feinstrukturierter Oberfläche auf dem marmorierten Untergrund zu erkennen; Mitte: selektiertes Vektorfeld als Ergebnis der Verschiebungsberechnung; rechts: Ergebnis der Vektorklassifikation (+ Untergrund, * Hindernis).*

Klasse aufgrund unterschiedlicher Flächenanteile und auftretender Störungen (z.B. Helligkeitssprünge) extrem schwankt. Durch die mehrfachen Eingangsmodalitäten wird die notwendige Trennbarkeit der beiden Klassen gewährleistet.

Das Modell des KHEPERA Der abzubildende Merkmalsraum besteht aus den folgenden lokalen Bildmerkmalen, die jeder Position im farbigen Eingangsbild zugeordnet werden können: intensitätsnormierte Farbe (Farbinformation), normierte Bildkoordinaten (Ortsinformation) und Antworten von drei isotropen Bandpaßfiltern in aufeinanderfolgenden Frequenzbereichen (Strukturinformation). Für das *sporadisch stattfindende* Training der beiden NGs wird das ortsentsprechende Trainingspattern jedes selektierten und klassifizierten Verschiebungs- bzw. Differenzwertes dem jeweiligen Netz als Trainingsmuster präsentiert. Zur Repräsentation des 7-dimensionalen Merkmalsraumes erwiesen sich 6 bis 8 Neuronen in jedem Netz als ausreichend.

Das Modell der MILVA Der abzubildende Merkmalsraum besteht hier nur aus der intensitätsnormierten Farbe (Farbinformation). Für das *sporadisch stattfindende* Training der beiden NGs werden die Trainingsbeispiele jedoch aufgrund des Ergebnisses der Verschmelzung der verschiedenen Segmentationen aquiriert, somit fließen in das Merkmalsmodell über das Training auch die Informationen der anderen Ansätze ein.

Die *kontinuierliche* netzbasierte Bildsegmentierung erfolgt, indem jedem Pixel eines Eingangsbildes diejenige Klasse zugeordnet wird, durch deren NG die oben beschriebenen lokalen Bildmerkmale am besten reproduziert werden. Ein Punkt des Bildes gehört also zur Klasse des NGs mit dem geringeren Abstand zum jeweiligen Best-Matching-Neuron. Zeigen beide Netze zu große Reproduktionsfehler, ist von einem (noch) unbekannten Objekt auszugehen. (Abb. 5)

3 Ergebnisse und Ausblick

Das in Abb. 5 gezeigte Segmentierungsergebnis in der KHEPERA-Anwendung nach einem bzw. sechs Trainingszyklen verdeutlicht, daß das Hindernis visuell vom Untergrund unterschieden werden kann und die detektierten lokalen Bildverschiebungen ausreichend Trainingsinformation für diese Segmentierung liefern.

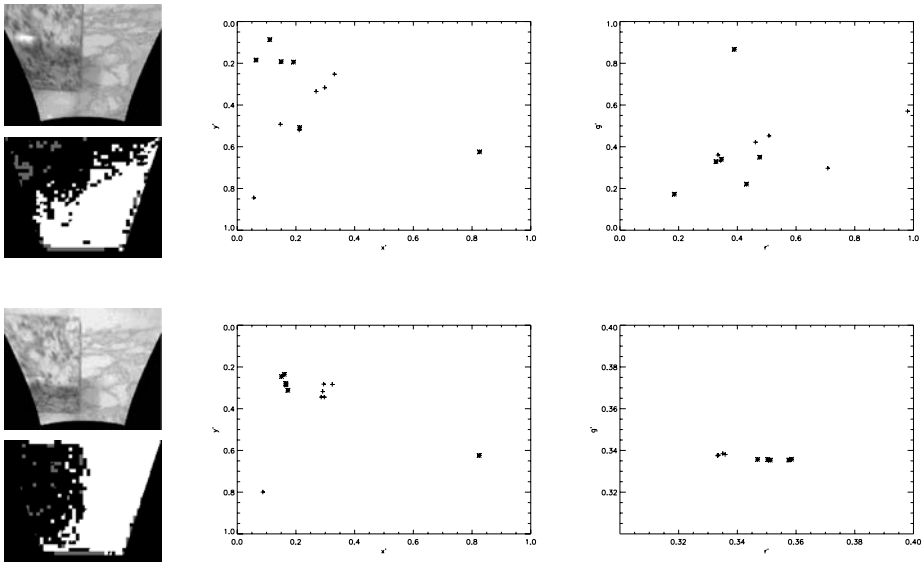

Abbildung 5. *Die Leistungsfähigkeit der Segmentierung nach dem ersten (obere Zeile) und nach fünf weiteren (untere Zeile) Trainingsschritten. Jeweils: Inputbild (links oben); Segmentierungsergebnis (links unten, weiß: Untergrund, schwarz: Hindernis, grau: unbekannt); Neuronenpositionen im NG der Hindernis- (*) bzw. der Untergrund-Klasse (+) über den beiden Positionskoeffizienten abgetragen (Mitte); über den beiden helligkeitsnormierten Farbkoeffizienten abgetragen (rechts).*

Der Einsatz der hier vorgestellten Architektur auf dem MILVA-Roboter in realer Indoor-Umgebung gestaltete sich aus zwei Gründen problematisch:

1. Auf dem Untergrund erscheinende Spiegelungen wirken sich in zweifacher Hinsicht negativ aus. Einerseits sind darin häufig die visuellen Merkmale von Hindernissen sichtbar, andererseits unterscheidet sich auch die Relativbewegung der Reflexionen von der des Untergrundes, was in beiden Fällen zu einer Hinderniswahrnehmung auf dem eigentlich befahrbaren Untergrund führt. Eine Lösung dieser Problematik bietet der Einsatz eines geeignet orientierten Polarisationsfilters vor dem Kameraobjektiv (Abb. 6).

2. Die meisten Indoor-Oberflächen sind derart fein strukturiert, daß Bildaufnahme und -vergleich erheblich gesteigerte Auflösungen erfordern, um aus den Strukturinformationen Bewegungshypothesen zu extrahieren. Die auf der MILVA verwendeten Bilder reproduzieren einen $1m$ breiten und ca. $1,5m$ tiefen Ausschnitt vor dem Roboter auf 205×160 Pixel. Damit lassen sich lediglich Strukturen größer als $3mm$ sowie Bewegungen der Objektkanten mit vertretbarem Aufwand erfassen. Diesen Objektkanten kann das korrekte visuelle Merkmal nicht genau zugeordnet werden, weil nicht sicher ist, auf welcher Seite der Kante das zugehörige Objekt liegt.

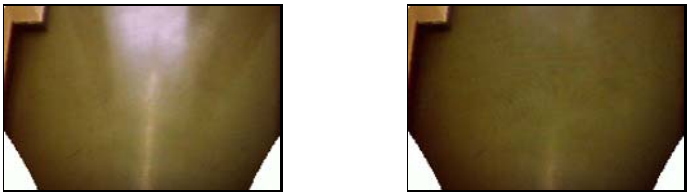

Abbildung 6. *Wirkung eines Polarisationsfilters. Störende Spiegelungen auf dem Untergrund sowohl von natürlichen (oberer Bildbereich) als auch künstlichen (unterer Bildbereich) Lichtquellen werden wirksam unterdrückt (links ohne, rechts mit Polarisationsfilter; beide invers-perspektivisch).*

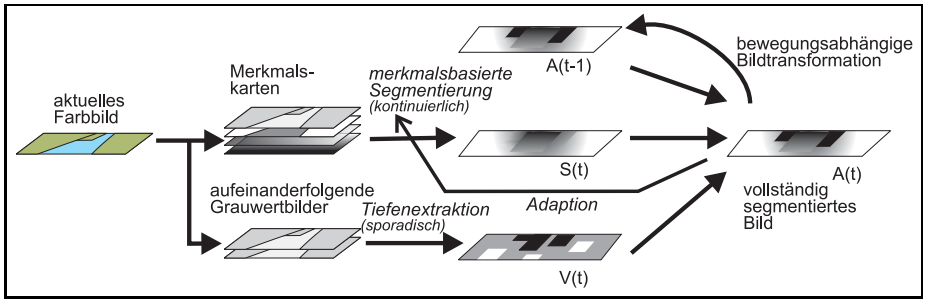

Abbildung 7. *Prinzipdarstellung der vorgeschlagenen erweiterten Architektur. In das Endergebnis fließen neben der Bildsegmentierung auch die unmittelbare Verschiebungsinformation sowie das Ergebnis des vorhergehenden Berechnungsschrittes ein.*

Deshalb wird eine Erweiterung der bisherigen Hybridarchitektur vorgeschlagen, deren Realisierung und Untersuchung zur Zeit erfolgt. Dabei soll zum Modelltraining nicht mehr ausschließlich die Bewegungsinformation und zur Bildsegmentierung ausschließlich das Merkmalsmodell verwendet werden. In einer rückgekoppelten Struktur fließen alle verfügbaren Informationen in ein gemeinsames Ergebnis ein, welches wiederum die Grundlage der Adaption des Merkmalsmodells bildet. Die merkmalsbasierte Segmentierung wird sowohl mit der unmittelbar vorhandenen Bewegungsinformation als auch mit dem Endresultat des vorherigen Schrittes fusioniert, das entsprechend der Relativbewegung verschoben ist. (Abb. 7). Auf diese Weise kann die gesamte nutzbare Information zur Herausbildung sicherer Hypothesen verwendet werden, um daraus korrekte Modellannahmen abzuleiten.

Das Gesamtverhalten der Fusionierung läßt sich in einer Differenzengleichung beschreiben,

$$A(t) = \alpha_1 \cdot P(A(t-1), \Delta x) + \alpha_2 \cdot V(t) + \alpha_3 \cdot S(t); \quad \sum_i \alpha_i = 1 \tag{3}$$

in der der Eintrag des um Δx (vom Roboter zurückgelegter Weg) verschobenen letzten Fusionsergebnisses $A(t-1)$, der verfügbaren Verschiebungsinformation $V(t)$ und des

Segmentierungsergebnisses der NGs $S(t)$ in das neue Gesamtresultat $A(t)$ über die Koeffizienten α_i überführt wird. Die Grenzwerte 0 und 1 repräsentieren die beiden Klassen *Hindernis* und *Untergrund*. Erste Experimente mit dieser Architektur lieferten vielversprechende Resultate.

Der vorgestellte Ansatz ist neben einem beherrschbaren Berechnungsaufwand und dem Vorteil der trotz monokularer Konfiguration möglichen Einzelbildauswertung vor allem in der Lage, die Aussage *Untergund/Hindernis* auf Bildregionen **ohne Struktur** und somit ohne erfaßbare Verschiebungsinformation zu extrapolieren. Damit sind Voraussetzungen gegeben, um auch visuell in realen Indoor-Umgebungen hindernisvermeidend zu navigieren, wobei das Verfahren auch auf binokulare Kamerasysteme übertragbar ist.
Das visuelle Merkmalsmodell liegt in der abstrahierten Repräsentation Neuronaler Gase vor und kann so als lokale Umgebungsbeschreibung zur Verkopplung mit einer globalen Navigationsebene genutzt werden. Da das Segmentierungsergebnis eine virtuelle Aufsicht darstellt, lassen sich Abstände zu und zwischen Hindernissen direkt extrahieren oder auch sogenannte Occupancy-Grids [9] erstellen. In den sich anschließenden Anwendungen des Fachgebiets Neuroinformatik soll das Ergebnisbild jedoch als unmittelbare Inputaktivierung für neuronale Netze zur lokalen Roboternavigation nach der ALVINN-Architektur dienen [7, 4].

Danksagung Die Autoren danken Ulf-Dietrich Braumann vom FG Neuroinformatik der TU Ilmenau für Anregungen, die in diese Arbeit eingeflossen sind.

Literatur

1. Bohrer, S.: Visuelle Hinderniserkennung durch Auswertung des optischen Flusses in inversperspektivischen Szenen. VDI-Verlag (1994)
2. Fäustle, P., Daxwanger, W. und Schmidt, G.: Steuerung lokaler Fahrmanöver durch direkte Kopplung abbildender Sensorik an ein künstliches neuronales Netz. Tagungsband zum 10. Fachgespräch Autonome Mobile Systeme, Springer-Verlag (1994) 214–225
3. Jähne, B.: Digitale Bildverarbeitung. Springer-Verlag, 3. Auflage (1993)
4. Krabbes, M., Böhme, H.-J., Stephan, V. und Groß, H.-M.: Extension of the ALVINN-Architecture for Robust Visual Guidance of a Miniature Robot. EUROBOT'97 – Proceedings of the 2nd EUROMICRO Workshop on Advanced Mobile Robots, IEEE Computer Society Press (1997) 8–14
5. Mallot, H. A., Bülthoff, H. H., Little, J. J. und Bohrer, S. Inverse Perspective Mapping Simplifies Optical Flow Computation and Obstacle Detection. Biological Cybernetics **64** (1991) 177–185
6. Martinetz, T. und Schulten, K.: A "Neural Gas" Network Learns Topologies. Artificial Neural Networks – Proceedings of the ICANN'91, Elsevier Science (1991) 397–402
7. Pomerleau, D. A.: Neural Network Perception for Mobile Robot Guidance. Kluwer Academic Publishers (1993)
8. Stöffler, N. O. und Färber, G.: An Image Processing Board with an MPEG Processor and Additional Confidence Calculation for Fast and Robust Optic Flow Generation in Real Environments. Proceedings of the ICAR'97 (1997) 845–850
9. Thrun, S., Fox, D. und Burgard, W.: Probabilistic Methods for State Estimation in Robotics. Tagungsband zum Workshop SOAVE'97, VDI-Verlag (1997) 195–202