Personentracking auf einer mobilen Roboterplattform unter Verwendung eines multimodalen Detektionsansatzes

Stefan Feyrer, Andreas Zell

Wilhelm-Schickard-Institut für Informatik, Abt. Rechnerarchitektur
Universität Tübingen
Köstlinstraße 6, D-72074 Tübingen
{feyrer, zell}@informatik.uni-tuebingen.de

Zusammenfassung. Dieser Beitrag beschreibt ein System, das es ermöglicht, mit einem mobilen Serviceroboter Personen visuell zu detektieren und mittels Kamera- und Fahrbewegungen zu verfolgen. Die Detektion basiert auf einem multimodalen Ansatz, der Farb-, Bewegungs- und Konturinformation kombiniert. Durch den Einsatz eines Stereoverfahrens wird eine räumliche Lokalisation der Person ermöglicht. Die Navigation erfolgt mit dem Ziel, einer erkannten Person durch Fahrbewegungen zu folgen. Stationäre und dynamische Hindernisse werden dabei durch Auswertung von Abstandssensoren und Verwendung einer lokalen Umgebungskarte umgangen.

1 Einleitung

Eine wichtige Voraussetzung für die Akzeptanz von mobilen Servicerobotern ist eine Mensch-Maschine-Schnittstelle, die es dem Benutzer erlaubt, in möglichst natürlicher Art und Weise mit dem Roboter zu kommunizieren und zu interagieren. Das Verstehen und das Formulieren gesprochener Sprache gehören dabei ebenso zu den Basisfähigkeiten wie die visuelle Erfassung der Umgebung und insbesondere des menschlichen Interaktionspartners. Mehrere dieser Einzelfähigkeiten in ein reales System zu integrieren, das sich völlig autonom in einer dynamischen natürlichen, d.h. nicht auf die Bedürfnisse des Roboters angepaßten Umgebung intelligent zurechtfindet und zudem bestimmte Serviceaufgaben erledigt, ist bisher lediglich ansatzweise und eingeschränkt auf einfache Szenarien gelungen [1, 4] und wird auch künftig eher als Fernziel einer längeren Entwicklung zu betrachten sein.

Auch dieser Beitrag greift nur einen einzelnen Aspekt eines solchen Gesamtsystems heraus: die visuelle Erkennung von Personen. Wie jede Wahrnehmungsleistung beinhaltet auch diese mehrere Beschreibungs- bzw. Interpretationsebenen. Auf der untersten Ebene erfolgt im Detektionsschritt eine Diskriminierung der Person von anderen Umgebungsobjekten. Darauf aufbauend kann ein Modell entwickelt werden, das beispielsweise eine Erkennung von Gesten oder Bewegungsformen zuläßt. Dazu ist es bereits notwendig, die Bildebene zu verlassen und den Übergang in den Szenenbereich zu vollziehen. In einer weiteren Stufe ist die Handlungsinterpretation angesiedelt, die eine Abstraktion in Richtung einer qualitativ-begrifflichen Repräsentation vornimmt und dazu Wissen über Intentionen und Handlungsabläufe einbeziehen muß. Wir beschränken uns hier auf die ersten beiden Ebenen, wobei die Implementierung der zweiten Ebene nicht auf eine Gestenerkennung abzielt, sondern die Lokalisation einer Person in der dreidimensionalen Szene und ein aktives Folgeverhalten des Roboters realisiert.

Das hier vorgestellte System bedient sich des Prinzips des aktiven Sehens [2]. Einerseits findet durch die Bewegungsnachführung der Kameras mit dem Ziel, den Kopfbereich der momentan im Blickfeld befindlichen Person möglichst im Zentrum des Bildes zu halten, eine für das aktive Sehen charakteristische Kopplung von Apperzeption und Aktorik statt. Andererseits wird die Kohärenz von Bildern in einer Bildsequenz ausgenutzt, um die Auswertung von interessanten Bildbereichen, sogenannten Regions of interest, robuster und effizienter zu gestalten.

Von der technischen Seite aus betrachtet müssen für ein System, das auf einer mobilen Plattform eingesetzt werden soll, eine Reihe von Randbedingungen berücksichtigt werden. Zum einen muß eine hohe Robustheit gewährleistet sein, d.h. es dürfen beispielsweise keine einschränkenden Annahmen über Szenenaufbau und Beleuchtungsverhältnisse gemacht werden. Ferner ist es wichtig, daß sich die Verarbeitungsgeschwindigkeit in einem Bereich bewegt, der ein natürlich wirkendes Reaktionsverhalten erlaubt. Nicht zuletzt muß dabei auch die aufgrund der limitierten Energieressourcen beschränkte Rechenleistung eines mobilen Roboters im Auge behalten werden. Diese Randbedingungen erzwingen bei der Auswahl der eingesetzten Verfahren eine Präferenz zugunsten von weniger allgemeinen, dafür aber effizienteren Algorithmen. Des weiteren müssen entsprechende Einschränkungen bezüglich der verwendeten Bildauflösung gemacht werden.

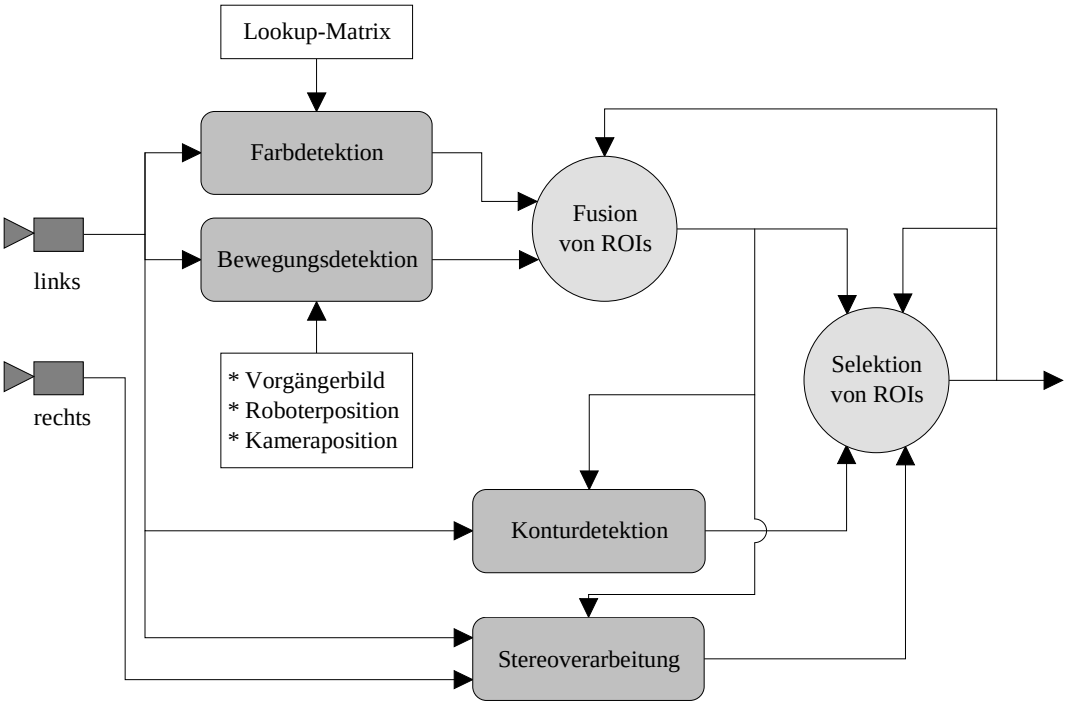


Abbildung 1: Struktureller Aufbau der Detektionskomponente

Während frühere Systeme zur visuellen Erkennung von Personen meist lediglich mit einem einzelnen Erkennungsmerkmal arbeiteten [13, 16], werden heute zunehmend Kombinationen mehrerer Merkmale verwendet, um der Forderung nach Robustheit und allgemeiner Einsetzbarkeit nachzukommen [3, 12]. Auch dem hier beschriebenen System liegt ein multimodaler Ansatz zugrunde, der neben einer farbbasierten Komponente eine Bewegungsdetektion und ein Verfahren zur Extraktion von Konturpunktketten beinhaltet. Die Auswertung von Abstandsinformation, die über ein Stereobildverfahren gewonnen wird und eine räumliche Lokalisierung der Person in der Szene zuläßt, wird für eine zusätzliche Plausibilitätsprüfung herangezogen und kann somit einen weiteren Beitrag zur Erhöhung der Stabilität des Systems leisten.

2 Visuelle Detektion von Personen

Der erste Teil des Detektionsvorgangs basiert auf der Extraktion von Hautfarbregionen sowie auf der Bestimmung von Bildbereichen, in denen signifikante Objektbewegungen in der Szene stattgefunden haben. Eine Fusionskomponente fügt dabei zusammengehörige benachbarte Bereiche zusammen. Zur anschließenden Selektion von Kopfregionen werden neben geometrischen und strukturellen Merkmalen zusätzlich Konturinformationen verwendet. Eine schematische Übersicht der Detektionskomponente ist in Abbildung 1 gegeben.

Farbbasierte Detektion

Die farbbasierte Komponente verwendet zur Detektion von Hautfarbe die in den HSI-Raum transformierten Farbwerte der einzelnen Bildpunkte, wobei der Intensitätsanteil nicht berücksichtigt wird. Im dabei entstehenden HS-Unterraum bleibt anschaulich gesprochen lediglich die reine Farbinformation erhalten, wodurch eine Invarianz bezüglich der Helligkeitswerte der einzelnen Bildpunkte erzielt wird. Diejenigen Elemente im HS-Raum, die Hautfarbwerte darstellen, sind durch eine Lookup-Matrix festgelegt. Diese Matrix basiert auf der manuellen Segmentierung von mehr als 50 Bildsequenzen, welche mit verschiedenen Personen und bei unterschiedlichen Beleuchtungsverhältnissen aufgenommen wurden. Die Verwendung einer Lookup-Matrix hat sich als notwendig erwiesen, da die ermittelten Hautfarbbereiche im HS-Raum weder einer Normalverteilung entsprachen noch durch geometrische Primitive einfach zu beschreiben waren. Um die aufwendige Online-Umrechnung der RGB-Werte in HSI-Werte zu vermeiden, wird eine in den RGB-Raum zurücktransformierte Tabelle verwendet. Einzelheiten zu den verwendeten Algorithmen sind in [6] zu finden.

Bewegungsdetektion

Die visuelle Detektion von Bewegung auf mobilen Systemen wirft stets das Problem auf, aus den Bildinformationen Fremdbewegung von durch Eigenbewegung verursachten Bildveränderungen zu unterscheiden. Bei diesem System werden zwei Fälle unterschieden. Bewegen sich lediglich die auf einer Pan-Tilt-Unit montierten Kameras, nicht aber der gesamte Roboter, so findet eine Detektion von Bewegung mittels eines Differenzbildverfahrens statt, das mit zwei bei temporär stationärer Kamera aufgenommenen Bildern arbeitet. Bewegen sich

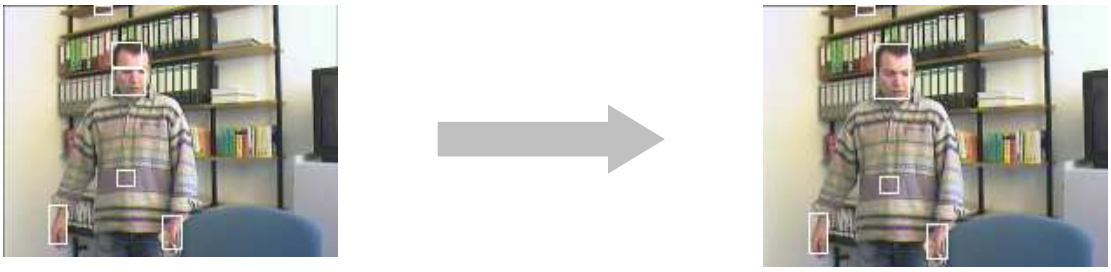


Abbildung 2: Fusion von Regionen

sowohl die Pan-Tilt-Unit als auch der Roboter, wird eine Eigenbewegungskompensation durchgeführt. Im ersten Fall ergibt sich zwar zwangsläufig eine gewisse Verzögerung der Ausführung bis zur jeweiligen Beendigung der Kamerabewegung, was eine sehr schnelle, sakkadische Kameranachführung notwendig macht. Der Ansatz hat aber zwei Vorteile. Erstens ergeben sich bei der Differenzbildanalyse wesentlich genauere Ergebnisse, da es sich bei beiden einfließenden Bildern um Originale handelt und keine Ungenauigkeiten durch fehlerhafte oder veraltete Orientierungs- und Positionswerte der Pan-Tilt-Unit oder der Roboterbasis enthalten sind, zweitens wird die aufwendige Berechnung eines Prädiktionsbildes eingespart.

Die Kompensation der Eigenbewegung erfolgt prinzipiell nach dem in [13] beschriebenen und in [6] erweiterten Verfahren. In jedem Verarbeitungszyklus wird zunächst die Positions- und Orientierungsänderung der Kamera bezüglich des zuletzt betrachteten Zeitpunkts berechnet. Dies ist einfach möglich, da das Programm selbst sowohl die Pan-Tilt- als auch die Fahrsteuerung vornimmt. Aufgrund dieser Eigenbewegungsdaten kann eine Prädiktion des Vorgängerbildes für die neue Kamerastellung berechnet werden. Dieses Bild wird durch Differenzbildung mit dem tatsächlich aufgenommenen Bild verglichen. Bewegungen in der Szene können somit extrahiert werden. Bei der Berechnung des Prädiktionsbildes ist allerdings für jeden Bildpunkt die Kenntnis des Abstandes zum Szenenobjekt notwendig. Die Abstandswerte werden über das in Abschnitt 3 beschriebene Stereoverfahren ermittelt. Dies erhöht zusätzlich die Komplexität der Berechnungen und erfordert eine Reduktion der in diesem Teilschritt verwendeten Bildauflösung um den Faktor 4.

Selektion von interessanten Regionen

Zur Clusterung der detektierten Bereiche kommt eine Kombination aus Region-growing und einer schnellen Variante des Flood-Fill-Algorithmus zum Einsatz [6]. Darin integriert ist gleichzeitig ein 3x3-Open-Operator, der ausgefranste Ränder abträgt und dadurch gleichzeitig kleinere Bereiche eliminiert.

Die entstehenden Clusterbereiche werden anschließend einem strukturbasierten Auswahlverfahren unterzogen mit dem Ziel, Gesichtsbereiche von anderen Clustern zu trennen. Als Kriterien werden dazu unter anderem das Höhen-Breiten-Verhältnis des umgebenden achsenparallelen Rechtecks sowie dessen Füllgrad verwendet. Tritt der Fall ein, daß die detektierte Region sich an einem der Bildränder befindet, geht das Höhen-Breiten-Verhältnis nicht in die Bewertung mit ein, da in diesem Fall keine Aussage über die tatsächliche Objektform gemacht werden kann. Des weiteren wird für jeden Bereich der Durchschnitt der relativen Häufigkeiten der HS-Farbkombinationen der zum Bereich gehörenden Bildpunkte bestimmt. Dieser Wert gibt Aufschluß darüber, wie stark die Farbe des Bereichs von der durch die manuelle Segmentierung ermittelten „mittleren Hautfarbe“ abweicht. Um ein Springen der selektierten Regionen zu vermeiden und ein gleichmäßiges Tracking zu ermöglichen, fließen zudem der Betrag der Positionsänderung und der Größenänderungsfaktor im Verhältnis zu den im jeweiligen Vorgängerbild selektierten Bereichen ein. Mögliche Zeitdifferenzen zwischen aufeinanderfolgenden Bildern durch unterschiedlich lange Verarbeitungszeiten werden dabei berücksichtigt.

Neben strukturellen Merkmalen fließen auch die Ergebnisse der Bewegungsdetektion in den Selektionsprozeß ein. Überschreitet der Anteil der von der Bewegungsdetektion extrahierten Bereiche einen bestimmten Prozentsatz der Fläche einer Region, wird deren Bewertung entsprechend angehoben.

Es ist geplant, in zukünftigen Versionen des Systems eine Positions- und Größenprädiktion der Bereiche mittels eines Kalman-Filters einzubauen, um somit eine bessere Kompensation der Verzögerungszeit zwischen Bildakquisition und Reaktion des Systems zu erreichen.

Fusion von Regionen

Oftmals kommt es vor, daß Gesichtsbereiche nicht als zusammenhängende Region, sondern als Menge von nahe beieinanderliegenden Einzelbereichen detektiert werden. Insbesondere entstehen bei weiter entfernten Personen und somit entsprechend niedriger Auflösung in Höhe der Augenlinie Unter-

brechungen der Segmentierungsbereiche. Ein typisches Beispiel ist in Abbildung 2 links gegeben.

Um zusammengehörige Regionen zu fusionieren, wird ein Algorithmus verwendet, der auf einer beschränkten Tiefentraversierung eines Nachbarschaftsgraphen basiert. Im ersten Schritt wird eine Adjazenzmatrix berechnet, die die Nachbarschaftsrelation der Region beschreibt. Ein Eintrag in die Adjazenzmatrix erfolgt, wenn sich zwei Rechtecke direkt überlappen oder wenn ihr Abstand unter und gleichzeitig der projektive Überlappungsgrad über einem jeweiligen Schwellwert liegen. Der projektive Überlappungsgrad ist dabei definiert als das maximale Verhältnis zwischen der Länge des Überdeckungsbereichs (projiziert in Richtung des minimalen Abstandes) und der entsprechenden Seitenlänge des Rechtecks.

Im zweiten Schritt wird eine Tiefentraversierung im Adjazenzgraphen durchgeführt. Die Auswertung der Testsequenzen hat ergeben, daß lediglich in weniger als 0,1% der Fälle ein Zusammenfügen von mehr als 4 Regionen nötig ist. Aus Effizienzgründen wurde deshalb die Traversierung auf eine Tiefe von 4 beschränkt. Jede im Laufe der Suche auftretende Kombination von benachbarten Regionen wird mittels des im vorigen Abschnitt beschriebenen Bewertungsverfahrens untersucht. Diejenige Kombination mit der höchsten Bewertung wird fusioniert und aus dem Adjazenzgraphen entfernt. Der algorithmische Aufwand von $O(2^n)$ für einen Graphen mit n Knoten kann somit auf $O(n^5)$ reduziert werden. Berücksichtigt man die Tatsache, daß ausschließlich planare Graphen vorliegen und zwischen je zwei benachbarten Rechtecken ein hoher projektiver Überlappungsgrad existiert, ergibt sich ein linearer Aufwand. In Abbildung 2 rechts ist das Ergebnis eines Fusionsvorgangs dargestellt.

Integration von Konturinformation

Neben Farbe und Bewegung bietet sich als weiteres charakteristisches Erkennungsmerkmal von Personen die Umrißlinie des Kopf-/Schulterbereichs an. Dieser Bereich ist im Vergleich zu anderen Körperpartien stärker invariant gegenüber Körperhaltungen, da keine Bewegungen der Extremitäten berücksichtigt werden müssen.

Die Konturdetektion greift auf ein Verfahren zurück, das unter dem Begriff *aktive Strahlen* bekannt ist [5]. Ausgehend von einem Referenzpunkt innerhalb der Objektkonturen werden die Grauwertverläufe auf Geraden durch diesen Punkt untersucht. Mit Hilfe einer Energieminimierung wird eine Punktkette ermittelt, die den tatsächlichen Objektkonturverlauf möglichst gut annähern. Aus Effizienzgründen wird die Minimierung der Energie hier auf ein kombinatorisches Problem zurückgeführt, das mittels einer lokalen Optimumssuche gelöst wird. Die berechneten Konturverläufe werden hinsichtlich ihrer Ähnlichkeit bezüglich einer statistisch ermittelten Durchschnittskontur bewertet [6].

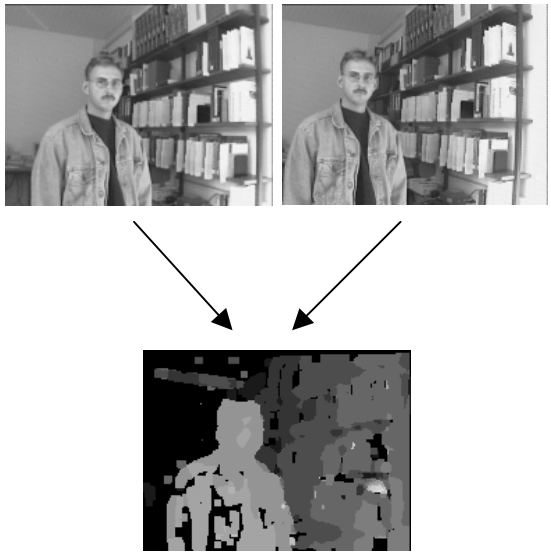
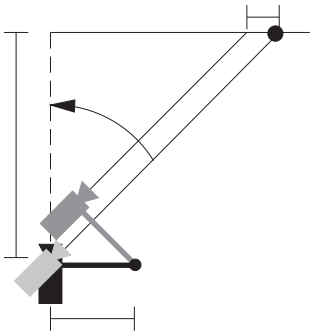


Abbildung 3: Stereobildpaar und daraus ermitteltes Tiefenbild

3 Gewinnung von Tiefeninformation

Tiefeninformationen, d.h. Informationen über den Abstand zwischen Kamera und im Bild sichtbaren Szenenobjekten, fließen an mehreren Stellen in das System ein. Zum einen werden Abstandswerte benötigt, um die Kompensation der Eigenbewegung bei bewegtem Roboter durchführen zu können. In diesem Fall ist es notwendig, daß für alle Bildpunkte Abstandsinformationen vorliegen. Um den Abstand zwischen Kamera und detektierter Person zu bestimmen, muß die Stereobildanalyse lediglich für die betreffende Kopfregion durchgeführt werden, was den Rechenaufwand deutlich verringert. Diese Abstandsinformation wird zum einen im Selektionsprozeß als weitere Plausibilitätsprüfung verwendet, wobei von der Annahme ausgegangen wird, daß sich der Abstand zwischen Person und Kamera innerhalb eines Verarbeitungszyklus nur mit sehr geringer Wahrscheinlichkeit wesentlich verändert hat. Zum anderen wird damit eine Datenstruktur aufgebaut, die den von der detektierten Person durchlaufenen Pfad in Weltkoordinaten beschreibt. Die Navigationskomponente verwendet diese Datenstruktur, um einen geeigneten Folgepfad zu berechnen.

Zur Gewinnung von Tiefeninformation wird eine Standard-Stereogeometrie [9] eingesetzt, bei der die optischen Achsen der beiden Kameras parallel verlaufen und die Zeilen der beiden Stereobilder parallel zu den Verbindungsgeraden der beiden optischen Zentren liegen. Dies hat den Vorteil, daß epipolare Linien entlang der Bildzeilen verlaufen und die Korrespondenzanalyse somit vereinfacht wird. Nachteil dieser Stereogeometrie ist, daß eine äußerst sorgfältige Justierung der Kameras gewährleistet sein muß. Bereits geringe Orientierungsdifferenzen der optischen Achsen, vor allem



$$\Psi = \Delta x \cdot \frac{\Psi_{max}}{w}$$

$$\Theta = \Delta y \cdot \frac{\Theta_{max}}{h}$$

Abbildung 4: links: Kameraplattform mit Pan-Tilt-Unit; rechts: Kameranachführung bei zentrierter (hell) und nicht zentrierter Drehachse (dunkel)

in vertikaler Richtung, führen zu inakzeptablen Ergebnissen. Aus diesem Grund ist eine automatische Kalibrierung unerläßlich. Diese wird hier über die Auswertung der Korrespondenz von signifikanten Punkten, die durch Verwendung des Moravec-Feature-Detektors [13] gewonnen werden, durchgeführt. Sowohl Translationen in vertikaler und horizontaler Richtung als auch Rotationen um die optischen Achsen werden dabei berücksichtigt.

Für die Korrespondenzanalyse, d.h. die Bestimmung von korrespondierenden Punkten im linken und rechten Bild, deren Abstand in der Bildebene schließlich auf die Szenentiefe schließen läßt, gibt es eine Reihe unterschiedlicher Möglichkeiten [9]. In diesem System erfolgt sie mittels einer Korrelation von Bildausschnitten der Größe 5x5, die zuvor einer Helligkeitsnormierung unterzogen werden. Als Korrelationsmaß wird die in [8] vorgeschlagene Funktion verwendet. Um die Anzahl von Fehlkorrespondenzen zu reduzieren, werden die Ergebnisse anschließend mit Hilfe eines Medianfilters und diverser morphologischer Operatoren nachbearbeitet. Das Ergebnis einer Tiefenberechnung ist in Abbildung 3 dargestellt.

4 Kamerasteuerung

Die verwendete binokulare Kameraplattform (Abbildung 4 links) besitzt zwei Freiheitsgrade, die durch einen Schwenkwinkel ψ und einen Nickwinkel θ beschrieben sind. Vergenzbewegungen sind nicht möglich. Geht man von dem Modell einer Lochkamera aus und setzt man zusätzlich voraus, daß sowohl die Schwenk- als auch die Nickachse durch das Projektionszentrum verlaufen, so lassen sich die nötigen Rotationswinkel für die Zentrierung eines Objekts in der Bildmitte durch einfache Verhältnisgleichungen bestimmen (Abbildung 4 rechts). Lediglich die Bildbreite *w* und die Bildhöhe *h* in Pixel sowie die beiden Öffnungswinkel des Sichtbereichs ψ_{max} und θ_{max} müssen bekannt sein.

Bei der verwendeten Kameraplattform ist sowohl die Schwenkachse als auch die Nickachse um 6 cm bezüglich des jeweiligen Projektionszentrums versetzt. Der dadurch entstehende absolute Fehler (angegeben in Weltkoordinaten) ist unabhängig vom Abstand *D* zwischen Objekt und Kamera und kann durch einfache geometrische Überlegungen hergeleitet werden. Für die Schwenkachse ergibt sich ein maximaler absoluter Fehler von 0,4 cm, für die Nickachse von 0,3 cm. Beide Werte liegen unterhalb der Meßgenauigkeit selbst bei nahen Objekten. Aus diesem Grund wird auf eine Kompensation des Achsenversatzes verzichtet.

5 Robotersteuerung und Navigation

Die zweite Systemkomponente setzt die Ergebnisse des Erkennungsmoduls in Bewegungen des Roboters um mit dem Ziel, der erkannten Person zu folgen bzw. sich ihr bis auf einen definierten Abstand zu nähern. Zur Navigation wird auf die sogenannte Potentialfeld-Methode [11] zurückgegriffen. Es handelt sich dabei um eine Echtzeit-Pfadplanungsmethode, die mittels der Auswertung der jeweils aktuellen Sensorwerte einem vorgegebenen Pfad kollisionsfrei folgt. Dazu wird ein Potentialfeld definiert, das sich aus einer anziehenden und einer abstoßenden Komponente zusammensetzt. Das abstoßende Potential baut um Hindernisse herum eine Potentialbarriere auf, die der Roboter nicht überwinden kann und die den Pfad möglichst weit an den Hindernissen vorbeiführt. Da keine globale Hinderniskarte vorliegt, fließen die von den Sensoren gelieferten Abstandswerte zu Umgebungsobjekten direkt in die Berechnung des abstoßenden Potentials ein. Das anziehende Potential wird durch eine parabolische Senkenfunktion definiert, in die der Winkel und der Abstand zum nächsten anzusteuernden Punkt einfließt. Alle Punkte des Pfades werden bezüglich ihrer direkten Erreichbarkeit überprüft. Unter den direkt erreichbaren Punkten wird der aktuellste ausgewählt. Ist kein Punkt des Pfades direkt erreichbar, wird als Zielposition diejenige Position verwendet, an der zuletzt eine Person erkannt wurde. Der zu verfolgende Pfad wird durch die in regelmäßigen zeitlichen Abständen von der Detektionskomponente gelieferten Raumpositionen der erkannten Person fest-

gelegt. Zur Bestimmung des Abstandes zwischen Roboter und Person werden die aus dem zuvor beschriebenen Stereoverfahren gewonnenen Werte herangezogen.

6 Ergebnisse und Ausblick

Das System kommt auf einem mobilen Roboter des Typs RWI B21 zum Einsatz (Abbildung 5). Der Roboter besitzt zwei Doppelprozessor-PentiumPro-PCs, von denen einer für die Bildverarbeitung, der andere für die Navigation verwendet wird. Die beiden Systemkomponenten kommunizieren über ein schnelles Interprozeß-Kommunikationsprotokoll. Die Bildakquisition erfolgt über zwei CCD-Farbkameras, die auf einer Pan-Tilt-Unit montiert sind. Für die Navigation können wahlweise die 24 ringförmig angebrachten Ultraschallsensoren oder der eingebaute 2D-Laserscanner verwendet werden.

Das System erreicht bei einer Auflösung von 192x144 Bildpunkten eine Verarbeitungsgeschwindigkeit von 8 Bildern pro Sekunde bei stationärem Roboter. Durch die bei fahrendem Roboter benötigte Kompensation der Eigenbewegung sinkt die Verarbeitungsgeschwindigkeit auf 4 Bilder pro Sekunde. Eine Sequenz von detektierten Kopfregionen ist in Abbildung 6 dargestellt. Die Einzelbilder wurden im Abstand von 5 Sekunden aufgenommen, um größere Bewegungsänderungen zu demonstrieren. Die Geschwindigkeit der Fahrsteuerung hängt stark von der Beschaffenheit der Umgebung und von der Entfernung zu Hindernisobjekten ab. Insbesondere das Verhalten im Nahbereich von Hindernissen läßt noch Spielraum für künftige Geschwindigkeitsoptimierungen.

Die Aufzeichnung einer Beispielfahrt ist in Abbildung 7 gegeben. Links ist der zurückgelegte Weg der Person dargestellt, rechts die Bahn des nachfolgenden Roboters. Bei den dunkel eingezeichneten Objekten handelt es sich um stationäre Hindernisse. Um eine Vorstellung von den zeitlichen Verhältnissen während der Beispielfahrt zu vermitteln, ist eine Momentaufnahme hervorgehoben, bei der sich der Roboter noch vor und die Person bereits hinter einem Hindernis befindet.

Abbildung 5: Die beiden mobilen Serviceroboter Robin und Colin

Ein Nachteil der momentanen Implementierung ist die Tatsache, daß das Detektionsmodul sehr wesentlich von der Farbkomponente abhängt. Dies hat zur Folge, daß sich die Personen, um noch detektiert zu werden, zwar in einem Winkelintervall von ca. 180°, jedoch stets dem Roboter zugewandt befinden müssen. In der beschriebenen Beispielfahrt ging die Person rückwärts. Um jedoch die geplante Anwendung, das Nachfolgen des Roboters bei auch in Vorwärtsrichtung vorausgehender Person realisieren zu können, muß eine Flexibilisierung der Detektion vorgenommen werden, die beispielsweise das Farbmerkmal lediglich für eine Initialisierung heranzieht und sich anschließend ausschließlich auf andere, eventuell neue Merkmale stützt.

Abbildung 6: Beispielsequenz mit ermittelten Kopfregionen und Konturverläufen

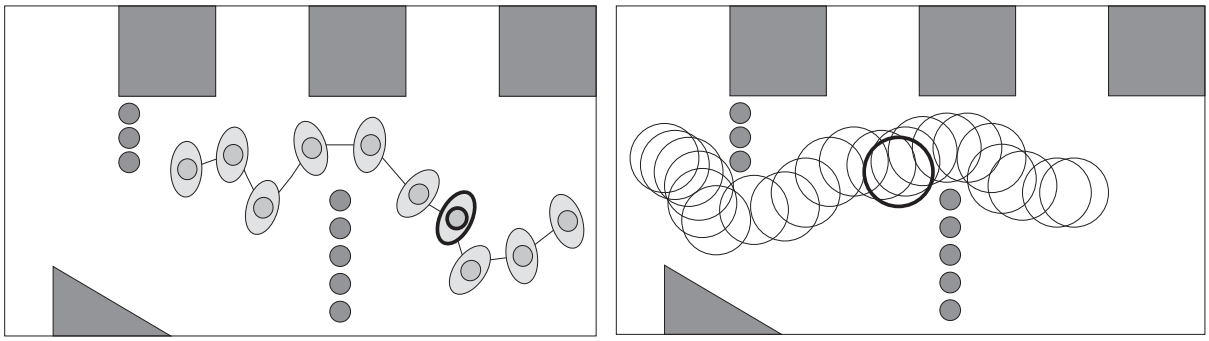


Abbildung 7: links: Pfad der Person; rechts: Pfad des nachfolgenden Roboters

Als zusätzliche Erweiterungen sind eine Verbesserung der Bildfolgenanalyse unter stärkerer Einbeziehung der Kohärenz von Bildsequenzen und eine Positionsprädiktion im dreidimensionalen Szenenraum geplant.

Literatur

1. T. Bergener, C. Bruckhoff, P. Dahm, H. Janßen, F. Joublin, R. Menzner: *Arnold: An Anthropomorphic Autonomous Robot for Human Environments*, Workshop SOAVE 97, Selbstorganisation von adaptivem Verhalten, Fortschritt-Berichte VDI, Reihe 8, Nr. 663, S. 25-34, 1997
2. A. Blake, A. Yuille: *Active Vision*, MIT Press, 1992
3. J. L. Crowley, F. Berard: *Multi-Modal Tracking of Faces for Video Communications*, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Puerto Rico, 1997
4. W. Daxwanger, E. Ettelt, C. Fischer, F. Freyberger, U. Hanebeck, G. Schmidt: *ROMAN: Ein mobiler Serviceroboter als persönlicher Assistent in belebten Innenräumen*, in G. Schmidt, F. Freyberger (Hrsg.): Autonome Mobile Systeme 1996, Informatik aktuell, Springer, S. 314-333, 1996
5. J. Denzler: *Aktives Sehen zur Echtzeitobjektverfolgung*, Dissertation, Technische Fakultät der Universität Erlangen-Nürnberg, 1997
6. M. Ebner: *Extraction of Moving Objects with a Moving Mobile Robot*, 3rd IFAC Symposium on Intelligent Autonomous Vehicles, Madrid, 1998
7. S. Feyrer, A. Zell: *Ein integrierter Ansatz zur Lokalisation von Personen in Bildfolgen*, Workshop Dynamische Perzeption, Proceedings in Artificial Intelligence, Bd. 8, infix-Verlag, S. 183-190, 1998
8. P. Fua: *A Parallel Stereo Algorithm That Produces Dense Depth Maps and Preserves Image Features*, Machine Vision and Applications, Bd. 6, Nr. 1, S. 35-49, 1993
9. X. Jiang, H. Bunke: *Dreidimensionales Computersehen – Gewinnung und Analyse von Tiefenbildern*, Springer, 1997
10. K. Konolige: *Small Vision Systems: Hardware and Implementation*, Eighth International Symposium on Robotics Research, Hayama, Japan, 1997
11. J.-C. Latombe: *Robot Motion Planning*, Kluver Academic Publishers Group, 1991
12. S. McKenna and S. Gong: *Tracking Faces*, Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition, Killington, Vermont, S. 271-277, 1996
13. H. Moravec: *Towards Automatic Visual Obstacle Avoidance*, Proceedings of the 5th International Joint Conference on Artificial Intelligence, Vision-I, S. 584, 1977
14. D. Murray, A. Basu: Motion Tracking with an Active Camera, IEEE Transactions on Pattern Analysis and Machine Intelligence, Bd. 16, Nr. 5, S. 449-459, 1994
15. K. Rohr: *Towards Model-Based Recognition of Human Movements in Image Sequences*, CVGIP: Image Understanding, Bd. 59, Nr. 1, S. 94-115, 1994
16. M. Sommerau, G. Mamier, A. Zell, M. Vogt, P. Levi: *Fast Face Localization and Tracking with Model-Based Time Synchronization*, in G. Sagerer, S. Posch, F. Kummert (Hrsg.): Mustererkennung 1995, Informatik aktuell, Springer, S. 516-523, 1995