Sprachgesteuerte Fovealisierung und Vergenz

U. Ahlrichs, J. Denzler, R. Kompe, H. Niemann
Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg
Martensstr. 3, D–91058 Erlangen, Germany
email: ahlrichs@informatik.uni-erlangen.de

Zusammenfassung

Von der Ausrüstung von Robotern mit visuellen Fähigkeiten verspricht man sich, diese in die Lage zu versetzen, komplexe Aufgaben zu verrichten. Damit wächst gleichzeitig auch die Forderung nach einer flexiblen Bedienungsschnittstelle. Natürliche Sprache kann dabei eine wichtige Hilfestellung bieten. In diesem Beitrag stellen wir ein System vor, in dem gesprochen–sprachliche Anweisungen oder Anfragen an ein Stereokamerasystem gerichtet werden können. Durch diese sprachlichen Äußerungen werden sowohl die vom Kamerasystem auszuführenden Aufgaben festgelegt als auch die betroffenen Objekte beschrieben. Um die für das Kamerasystem wesentliche Information aus der Äußerung zu erschließen, wird diese nach Abbildung auf die optimale Wortkette durch einen Worterkenner mit Hilfe von semantischen Klassifikationsbäumen interpretiert. Die vom Kamerasystem zu lösenden Aufgaben umfassen die Lokalisation oder formatfüllende Darstellung (Fovealisierung) der beschriebenen Objekte sowie die Interpretation von Lageverhältnissen zwischen jeweils zwei Objekten. Für die Objektlokalisation wird ein auf Histogrammen basierender Ansatz verwendet. Die Fovealisierung eines Objekts wird über Zoombewegungen erreicht, während die Interpretation von Lageverhältnissen auf der Beurteilung des Vergenzwinkels beruht.

1 Motivation

Schon heute werden Serviceroboter zum Beispiel zur Unterstützung behinderter und alter Menschen eingesetzt (Kawamura & Iskarous, 1994). Der Gebrauch dieser Roboter würde für den Menschen wesentlich erleichtert, wenn man ihm eine flexible Schnittstelle zur Bedienung böte. Ein effizientes Mittel hierfür kann gesprochene Sprache sein. Zentral ist dabei die Beschreibung von Aufgaben, die vom Roboter auszuführen sind, beziehungsweise der beteiligten Objekte. Diese Beschreibungen sollten über eine bloße Kommandosprache hinausgehen und idealerweise Spontansprache zulassen.

Als typische Anwendung in dem oben genannten Gebiet ist die Aufforderung eines Benutzers an den Roboter vorstellbar, ihm ein bestimmtes Buch zu bringen. Die Lösung dieser Aufgabe erfordert, daß der Roboter mit visuellen Fähigkeiten ausgerüstet ist, die es ihm erlauben, ein Buch zu lokalisieren, dieses gegebenenfalls in einer höheren Auflösung abzutasten, um Aufschriften zu entziffern, oder mit Kenntnis von Lageverhältnissen das richtige Buch zu greifen.

Als ersten Schritt in diese Richtung stellen wir ein prototypisches Gesamtsystem vor, in dem einem Kamerasystem in gesprochener Form verschiedene Aufgaben gestellt werden können. In der aktuellen Realisierung wird zunächst ein eingeschränktes Vokabular verwendet, und auf vier rechteckige Objekte zurückgegriffen, die durch ihre Farbe eindeutig charakterisiert sind. Die Äußerungen zur Kamerasteuerung sind ganze Sätze; indirekte Umschreibungen der Objekte oder Aktionen und eine weitgehend flexible Wortstellung sind erlaubt. Die Analyse erfolgt im Gegensatz zu (Socher, Fink, Kummert, & Sagerer, 1996) nicht wissensbasiert, sondern datengetrieben im Sinne des Aktiven Sehens.

Im folgenden wird zunächst das Gesamtsystem näher erläutert (Abschnitt 2). Weiterhin gehen wir genauer auf die Sprachanalyse (Abschnitt 3) sowie die Bildanalyse (Abschnitt 4) ein und stellen Ergebnisse zum Gesamtsystem vor (Abschnitt 5).

2 Das Gesamtsystem

Das in Bild 1 dargestellte Gesamtsystem gliedert sich in fünf unterschiedliche Module. Dabei handelt es sich um die akustische und die linguistische Analyse, die Objektlokalisation, die Fovealisierung sowie das Vergenzmodul. Als Eingabe erhält das System das Sprachsignal einer Benutzeräußerung. Abhängig von der in der Äußerung gestellten Aufgabe generiert das System entweder ein Sprachsignal, zum Beispiel die Verneinung einer Anfrage, oder ein Bild, das zum Beispiel das Objekt formatfüllend zeigt, als Ausgabe. Während der akustischen Analyse wird die vom Benutzer geäußerte Anfrage von einem Worterkenner auf die beste Wortkette abgebildet, die dann in der linguistischen Analyse

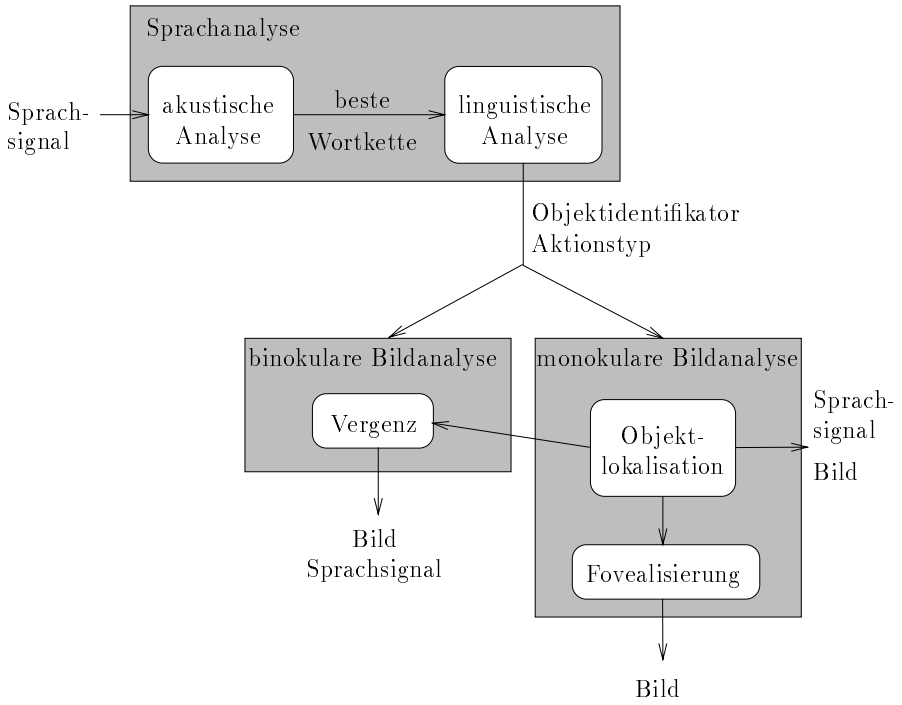

Bild 1: Architektur des Gesamtsystems

interpretiert wird. Dabei erfolgt die Bestimmung der in der Wortkette beschriebenen Objekte und der Aufgabe für das Kamerasystem. Diese Information wird an die Bildverarbeitungsmodule in Form eines Objektidentifikators und eines "Aktionstypen" weitergegeben. Die Objektlokalisation, bei der das gesuchte Objekt in einer statischen Szene durch Schwenk–/Neigebewegungen der Kamera in die Bildmitte zu holen ist, und die Fovealisierung sind zwei mögliche Aktionstypen. Die restlichen sechs Aktionstypen ergeben sich aus der Überprüfung der Relationen "rechts neben, links neben, vor, hinter, über und unter". Das Zutreffen dieser Relationen wird mit Hilfe des Vergenzmoduls untersucht.

3 Sprachanalyse

3.1 Die Stichprobe

Die Sprachanalyse stellt den ersten Teilschritt auf dem Weg zur Erfüllung der dem System durch den Benutzer gestellten Aufgabe dar. Für das Training der für die Äußerungsinterpretation verwendeten semantischen Klassifikationsbäume ist eine große Stichprobe mit typischen Äußerungen notwendig (vgl. Abschnitt 3.2). Eine manuelle Erstellung oder das Verwenden eines Wizard–of–Oz–Verfahrens (Corazza, Federico, Gretter, & Lazzari, 1993) ist relativ mühsam. Deshalb werden die Sätze hier mit einem stochastischen Satzgenerator erzeugt, der auf einer kontextfreien Grammatik basiert (Rieck, 1995). Ein weiterer Vorteil dieser automatischen Generierung der Stichprobe ist, daß die für das Trainieren der Klassifikationsbäume (vgl.Abschnitt 3.2) benötigten Referenzetiketten für Objekt– und Aktionstypen nicht manuell erstellt werden müssen.

Bei der Spezifikation der Satztypen wurde darauf Wert gelegt, daß indirekte Umschreibungen der Ojekte und Aktionen möglich sind. So ist für die Beschreibung der Objekte nicht nur ihre Farbe ein Kriterium. Vielmehr können auch Objekteigenschaften wie Form oder Materialeigenschaften verwendet werden. So kann zum Beispiel die *gelbe Eisenbahn* auch als *der aus den kleinen Legosteinen zusammengesetzte Klotz* beschrieben werden.

Die Beschreibungen der Aktionstypen werden in Anlehnung an (Stopp & Laengle, 1995) in einfache und komplexe Anweisungen bzw. Anfragen unterteilt. Einfache Anweisungen bestehen aus einer Nominalgruppe, der möglicherweise ein oder mehrere Adjektive zugeordnet sind, und einem Verb. Beispiele hierfür sind die Sätze *Suche die große rechteckige Pappschachtel* und *Zeige mir die Kiste, und zwar die rote*. Die Umschreibungen der Objektlokalisation, charakterisiert durch die Verben *finde, zeige mir, suche*, und der Fovealisierung gehören zu diesen einfachen Anweisungen. Zusätze in den Sätzen wie *möglichst groß, so groß wie möglich* oder *so groß es geht* deuten auf eine Fovealisierung hin.

Die komplexen Anweisungen umfassen die Beschreibung der Lageverhältnisse, bei denen zusätzliche Ortsangaben

auftreten, die die Lage des Objekts genauer charakterisieren sollen. Ein Beispiel hierzu ist der Satz *Finde die gelbe Eisenbahn links neben der großen rechteckigen Pappschachtel.*

3.2 Akustische und linguistische Analyse

Nach Äußerung einer Anweisung durch den Benutzer wird dieses Sprachsignal von einem Spracherkenner auf die optimale Wortkette abgebildet. Der Aufbau dieses Erkenners ist in (Schukat-Talamazzini, Kuhn, & Niemann, 1994) beschrieben. Für die Bestimmung einer Wortkette greift der Erkenner auf ein akustisches Modell, repräsentiert durch Hidden Markov Modelle, und ein linguistisches Modell zurück. Dabei wurden die Parameter des akustischen Modells vom VERBMOBIL–Worterkenner (Kuhn, 1995) übernommen und nur das Vokabular angepaßt. Mit dem linguistischen Modell wird die Wahrscheinlichkeit für das Auftreten eines Wortes in Abhängigkeit der Vorgängerwörter bewertet. Auf einer Teststichprobe mit 100 Sätzen ergab sich bei einem linguistischen Modell, in dem bis zu drei Vorgänger betrachtet werden, eine Wortfehlerrate von 3.8 %, bei bis zu vier Vorgängern eine Wortfehlerrate von 3.3 %. Die Sätze wurden von fünf Sprechern gelesen und stammen aus der in Abschnitt 3.1 beschriebenen Stichprobe.

Die linguistische Analyse basiert auf speziellen binären Bäumen, sogenannten semantischen Klassifikationsbäumen (SKBs) (Kuhn & De Mori, 1995). Diese lernen anhand einer etikettierten Stichprobe semantische und syntaktische Regeln. Die Interpretation einer Wortkette wird als Klassifikationsaufgabe gelöst, das heißt die Wortkette wird auf einen Objektidentifikator und einen Aktionstypen abgebildet. Die SKBs besitzen den Vorteil, daß sie im Gegensatz zu herkömmlich verwendeten, regelbasierten Systemen (Mast, Kummert, Ehrlich, Fink, Kuhn, Niemann, & Sagerer, 1994) anhand einer etikettierten Stichprobe automatisch trainiert werden können. Für eine neue Stichprobe ist somit nur eine erneute Trainingsphase erforderlich, während bei regelbasierten Verfahren eine neuerliche manuelle Festlegung der Regeln notwendig ist.

Wie alle binären Bäume bestehen auch die SKBs aus internen Knoten, den Nichtterminalknoten, und aus Blättern, die als Terminalknoten bezeichnet werden. Den Nichtterminalknoten sind Aufspaltungsregeln zugeordnet, die im allgemeinen die Form regulärer Ausdrücke besitzen. Mit diesen wird das Vorkommen bestimmter Schlüsselwörter in einer Wortkette sowie die Struktur der Wortkette überprüft. Jeder Nichtterminalknoten besitzt außerdem einen JA– und einen NEIN–Unterbaum, die je nach Übereinstimmung der Wortkette mit dem regulären Ausdruck betreten werden. An den Terminalknoten des Baumes wird der Wortkette die zum Knoten gehörige Klasse zugewiesen. Ein möglicher regulärer Ausdruck ist $+w_i+$ oder als Frage formuliert: "Hat die Wortkette $\boldsymbol{w}$ die Struktur $+w_i+$", wobei + eine unbekannte, nichtleere Folge von Wörtern und w_i ein Wort des Vokabulars symbolisiert.

Während des Trainings der SKBs wechseln sich Expansions– und Pruningphasen auf zwei disjunkten Trainingsmengen ab (Kuhn & De Mori, 1995). Diese Iterationen werden abgebrochen, wenn sich nach zwei aufeinanderfolgenden Pruningphasen der gleiche Baum ergibt. Die Festlegung der für die Expansionsphase benötigten Menge von Aufspaltungsregeln und der Regel zur Auswahl der besten Frage in einem Knoten oder zur Entscheidung, ob ein Knoten Terminalknoten werden soll, orientiert sich an (Breiman, 1984).

Für die Klassifikation der in diesem Problemkreis vorliegenden Sätze werden insgesamt fünf verschiedene Bäume verwendet. Die einzelnen Klassifikationsergebnisse werden miteinander zu einem Gesamtergebnis kombiniert, das die Interpretation der Wortkette darstellt. Dabei wird zunächst mit dem ersten Baum entschieden, ob in dem Satz ein oder zwei Objekte beschrieben sind. Handelt es sich um einen Satz mit nur einer Objektbeschreibung wird mit dem zweiten Baum untersucht, um welches Objekt es sich handelt. Bei zwei beschriebenen Objekten in einem Satz muß die Entscheidung getroffen werden, welches das zu lokalisierende Objekt und welches das Referenzobjekt innerhalb des Satzes ist (Stopp & Laengle, 1995). In dem Satz *Steht die gelbe Eisenbahn links neben der großen rechteckigen Pappschachtel* ist zum Beispiel das zu lokalisierende Objekt *die gelbe Eisenbahn* und das Referenzobjekt *die rechteckige Pappschachtel.* Mit dem vierten Baum wird entsprechend das zu lokalisierende Objekt und mit dem fünften das Referenzobjekt klassifiziert. Der dritte Baum wird zur Bestimmung des Aktionstypen herangezogen.

Betrachtet man die Erkennungsraten für alle Bäume unabhängig voneinander, so ergab sich bei 10000 Trainingssätzen jeweils eine Erkennungsrate von 99 %, bei 1000 Trainingssätzen von 85 %. Die Teststichprobe umfaßte dabei 2000 Sätze, die aus der in Abschnitt 3.1 beschriebenen Stichprobe entnommen wurden und in der Trainingsstichprobe nicht enthalten waren.

4 Aktive Bildanalyse

An die Sprachanalyse der Benutzeräußerung schließt sich die Reaktion des Stereokamerasystems an. Dabei kann es sich um eine Objektlokalisation, eine Fovealisierung oder die Ausführung von Vergenzbewegungen handeln.

4.1 Objektlokalisation

Wünscht der Benutzer die Suche eines Objekts, so muß sich das Stereokamerasystem zunächst einen Überblick über den Aufbau der Szene verschaffen. Durch Bewegung einer der beiden Vergenzachsen wird hierzu ein Szenenüberblick gebildet. Dieser entsteht durch die Aufnahme eines Bildes an jeder angefahrenen Achsenposition und Kopie der mittleren Bildspalte in ein Gesamtbild. Dadurch erhält man eine eindeutige Zuordnung zwischen einer Bildspalte im Szenenüberblick und der Achsenposition, die angefahren werden muß, damit sich die entsprechende Spalte im Bildmittelpunkt befindet. Somit kann eine Sakkade durchgeführt werden, die das gesuchte Objekt in die Bildmitte bewegt.

Für die Suche des Objektmittelpunktes wird die Farbinformation des Objekts ausgenutzt. Diese wird in einer Vorabphase automatisch gelernt und in Form eines Farbhistogramms repräsentiert. Als Farbraum wird der rgb–Farbraum verwendet, der aus dem RGB–Farbraum durch Normierung der Farbwerte mit der Intensität entsteht. Durch die Normierung erreicht man eine gewisse Farbkonstanz. Außerdem sind dann zur Repräsentation der Objekte nur zweidimensionale Histogramme erforderlich. Zusätzlich wird der Farbraum in sogenannte Bins unterteilt, die bei uniformer Einteilung der drei Farbachsen zum Beispiel in 64 Abschnitte die Form eines Würfels besitzen. Jedem Bin werden die Farbwerte zugeordnet, die in den durch dieses Bin repräsentierten Bereich des Farbraums fallen.

Vor der eigentlichen Bestimmung des Objektmittelpunktes wird in einem Zwischenschritt das gelernte Histogramm des Objekts in den Szenenüberblick "rückprojiziert" (Swain & Ballard, 1991). Ziel ist es dabei, Farben abzuschwächen, die auch in anderen Objekten auftreten, um den anschließenden Suchprozeß zu vereinfachen. Bei der Rückprojektion wird zunächst ein Verhältnishistogramm aus gelerntem Histogramm und Histogramm des Szenenüberblicks gebildet. Dazu werden die relativen Häufigkeiten der Bins in beiden Histogrammen dividiert und das Minimum von dem Divisionsergebnis und Eins gebildet. Für jedes Pixel im Szenenüberblick wird der Wert des Verhältnishistogramms desjenigen Bins in das rückprojizierte Bild eingetragen, dem das Pixel im Farbhistogramm des Szenenüberblicks zugeordnet worden ist.

Die Lokalisation eines Objekts beruht auf einer Suche des Objektmittelpunktes in dem bei der Rückprojektion entstehenden Grauwertbild. Hierzu wird eine vertikale und eine horizontale Projektion herangezogen. Bei der vertikalen Projektion werden die Grauwerte in jeder Bildspalte aufaddiert, so daß linke und rechte Objektgrenze in dem dabei entstehenden eindimensionalen Feld mittels Heuristiken bestimmt werden können. Die horizontale Projektion, bei der analog die Grauwerte der Bildzeilen aufaddiert werden, beschränkt sich auf den Bereich zwischen linker und rechter Objektgrenze. Sie dient der Berechnung der oberen und unteren Objektgrenze, aus denen sich mit den beiden anderen Objektgrenzen der Objektmittelpunkt ergibt. Durch entsprechende Bewegung der Vergenz– und Tiltachse des Stereokamerasystems befindet sich das gesuchte Objekt nach der Objektlokalisation im Mittelpunkt des Bildes.

Im Gegensatz zu (Mahlmeister, Pahl, & Sommer, 1996), die eine Histogramm–Schnittbildung zur Klassifikation von Objekten mit bekannter Position im Bild realisierten, besteht die zu lösende Aufgabe bei der Objektlokalisation in der Rückprojektion des Histogramms des zu suchenden Objekts. Eine interessante Erweiterung für zukünftige Arbeiten stellt das in (Mahlmeister et al., 1996) vorgestellte Farb–Richtungshistogramm zur Steigerung der Robustheit des Lokalisationsergebnisses dar.

4.2 Fovealisierung und Fokussierung

Ziel der Fovealisierung ist die Fixation eines Objekts und eine Erhöhung der Auflösung, mit der das Objekt dargestellt wird. Dies erlaubt eine genauere Betrachtung des Objekts und liefert mehr Information über strukturelle Details in einem Objekt. Da das Auflösungsvermögen von CCD–Chips abgesehen von Spezialhardware in allen Bereichen gleich ist, kann das biologische Prinzip auf Kameras nicht direkt übertragen werden. Eine Erhöhung der Auflösung kann auch durch eine Erhöhung der Brennweite erreicht werden. Eine solche Veränderung der Brennweite entspricht dem Heranzoomen an ein Objekt.

Im allgemeinen stimmt jedoch bei einer Kamera die mechanische Achse, entlang der sich die Linse bewegt, nicht mit der optischen Achse überein. Als Folge ergibt sich während der Zoomsteuerung eine Verschiebung eines anfänglich im Bildmittelpunkt liegenden Objekts in Richtung Bildrand. Deshalb müssen Abweichungen zwischen Objektschwerpunkt und Bildmittelpunkt innerhalb eines geschlossenen Rückkopplungssystems entgegengesteuert werden. Dazu muß nach jeder Zoombewegung der Objektschwerpunkt bestimmt werden. Da vor der Fovealisierung zunächst immer eine Objektlokalisation durchgeführt wird, kann man davon ausgehen, daß sich das betroffene Objekt ungefähr im Bildmittelpunkt befindet. Aus den Farbwerten innerhalb einer 8–Nachbarschaft wird pro Farbkanal ein Mittelwert gebildet, der für eine Schwellwertoperation herangezogen wird. In dem dabei entstehenden Binärbild wird heuristisch der Objektschwerpunkt bestimmt, wobei jeweils nur in den mittleren drei Bildspalten und in den mittleren drei Bildzeilen nach den Objektgrenzen gesucht wird. Aus den Grenzen läßt sich dann der Objektschwerpunkt berechnen. Abweichungen zwischen Objektschwerpunkt und Bildmittelpunkt wird durch Bewegung der entsprechenden Achse entgegengesteuert. Das Ergebnis eines solchen Zoomvorgangs zeigt Bild 2.

Während der Fovealisierung wird im allgemeinen der Tiefenbereich maximaler Schärfe verändert und somit die Abbildung des interessierenden Objekts verschlechtert. Zwar verfügen moderne Kameras über sogenannte Autofokus–

Bild 2: Verschiedene Stadien bei der Fovealisierung

Systeme, die den Bildinhalt automatisch scharf stellen. Im Bereich des Rechnersehens, gerade innerhalb der neuen Verarbeitungsstrategie des aktiven Sehens, sollen jedoch gezielt anhand von Fokussierung und Defokussierung Tiefeninformationen über die Umwelt erlangt werden (Krotkov & Bajcsy, 1993). Deshalb müssen Autofokus–Systeme bei der Bildaufnahme deaktiviert werden, woraus sich die Notwendigkeit ergibt, diese Funktion über Software zur Verfügung zu stellen.

In (Krotkov, 1989) werden einige Verfahren zur Bestimmung der Fokus–Position vorgeschlagen, die einen bestimmten Tiefenbereich in die Bildebene scharf abbildet. Das generelle Vorgehen besteht in der Bestimmung einer Gütefunktion für die Schärfe im Bild und deren im praktischen Einsatz durchgeführten Optimierung. Zu den Gütefunktionen zählen Funktionen, die hohe Frequenzen im Bild bestimmen, die Grauwerthistogrammentropie, das Histogramm zur lokalen Grauwertveränderung, die Grauwertvarianz und die Summen–Modulus–Differenz.

In dieser Arbeit wurden zwei frequenzbasierte Verfahren verwendet: die Fouriertransformation und das Tenengrad–Verfahren, das direkt Kanten im Bild mißt und die Kantenstärke maximiert. Die Optimierung der Gütefunktionen erfolgt über die in (Krotkov, 1989) vorgeschlagene Fibonacci–Suche. In Bild 3 sind drei Bilder einer Szene zu sehen, die bei unterschiedlichen Fokus–Motorpositionen aufgenommen wurden. Die Fokus–Motorpositionen wurden über das Tenengrad–Verfahren ermittelt, das drei Maxima in der Bewertungsfunktion liefert (siehe Bild 4). Diese drei Maxima entsprechen drei Objekten in drei unterschiedlichen Tiefenbereichen, die sich innerhalb des Bildausschnitts befanden, in dem die Gütefunktion ausgewertet wurde. Beide Verfahren liefern vergleichbare Ergebnisse, wobei die Maxima des Tenengrad–Verfahrens in den durchgeführten Versuchen ausgeprägter und somit leichter zu bestimmen waren. Wählt man die Position des Bildausschnittes derart, daß dieser nur ein Objekt enthält, erhält man bei beiden Gütefunktionen eine unimodale Funktion.

4.3 Vergenz

Um die relative Lage zweier Objekte zueinander zu bestimmen, ist Information über die relative Tiefe der Objekte notwendig. Dazu kann bei der Verwendung eines Stereokamerasystems ausgenutzt werden, daß der von den beiden optischen Achsen eingeschlossene Winkel (Vergenzwinkel) ein Maß für die Tiefe von Objekten ist. Dies gilt jedoch nur, wenn sich die optischen Achsen in einem Objektpunkt schneiden. In diesem Fall ergibt sich dort keinerlei Verschiebung (Disparität) innerhalb der Stereobilder. Ziel der Vergenzsteuerung ist es somit, die Disparität durch Rotation der beiden Kameras um die vertikale Achse so weit wie möglich zu reduzieren. Da sich im Anschluß an die Objektlokalisation der

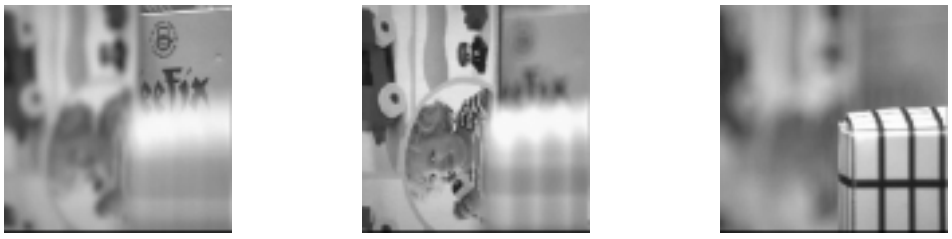

Bild 3: Bilder der komplexen Szene, bei der der zur Fokussierung verwendete Bildausschnitt drei Objekte in unterschiedlichen Tiefenbereichen enthält. In den drei Bildern sind Fokusmotorpositionen angefahren, die jeweils ein Objekt scharf abbilden.

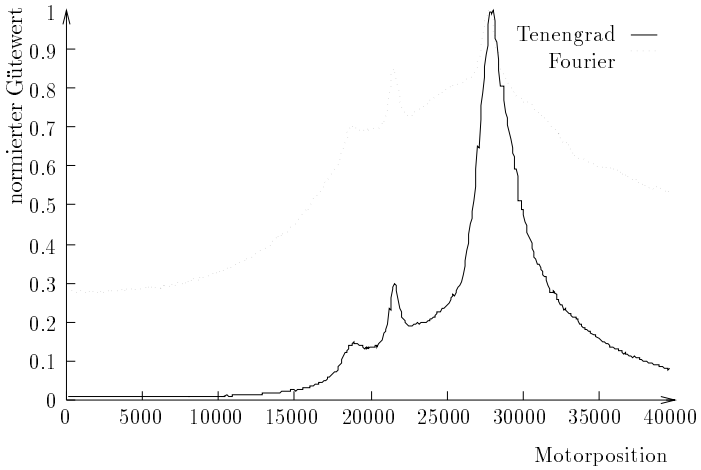

Bild 4: Gütefunktion für die Fokussierung in einem Bildausschnitt mit drei Objekten

Schwerpunkt des zu untersuchenden Objekts im Bildmittelpunkt befindet, können anhand des nach der Anwendung der Vergenzsteuerung bestimmten Vergenzwinkels Aussagen über die relative Tiefe der Objekte getroffen werden.

Für die Bestimmung von Disparitäten zwischen zwei Stereobildern hat sich in letzter Zeit neben den beiden klassischen Ansätzen, den korrelations– und den merkmalsbasierten Verfahren, ein neuer Ansatz etabliert, bei dem die Bestimmung der Disparität auf der Berechnung von Phasendifferenzen beruht. Diese Verfahren liefern dichte Disparitätskarten, jedoch ohne den bei den korrelationsbasierten Verfahren notwendigen Suchprozeß, und sind deshalb gerade für den Echtzeitansatz geeignet (Hansen & Sommer, 1996). Das Prinzip der phasenbasierten Verfahren beruht auf der Verschiebungseigenschaft der Fouriertransformation. Diese besagt, daß sich eine globale Verschiebung eines Signals im Ortsbereich in einer Phasenverschiebung im Frequenzbereich widerspiegelt. Die Disparitäten zwischen den Stereobildern sind jedoch nicht notwendigerweise konstant für alle Bildpunkte. Deshalb wird eine Möglichkeit benötigt, lokale Phasendifferenzen zu bestimmen. Diese Anforderung wird zum Beispiel von Gaborfiltern erfüllt. Wir verwenden die in (Theimer & Mallot, 1995) angegebene Definition von Gaborfiltern, bei der sich die in Bild 5 dargestellte Aufteilung der Frequenzebene ergibt. Durch die Anpassung der Größe der Gaußschen Einhüllenden des Gaborfilters an dessen Position im Frequenzraum erreicht man, daß für hohe Frequenzen eine hohe Auflösung im Ortsbereich und für niedrige Frequenzen eine geringe Auflösung bereitgestellt wird.

Faltet man beide Stereobilder mit einem Gaborfilter kann aus der Phasendifferenz zwischen den sich bei der Faltung ergebenden Bildern die Disparität in jedem Bildpunkt berechnet werden. Dabei ist jedoch zu beachten, daß eine Verschiebung senkrecht zur Orientierung des Filters nicht detektierbar ist (stereoskopisches Aperturproblem) und der

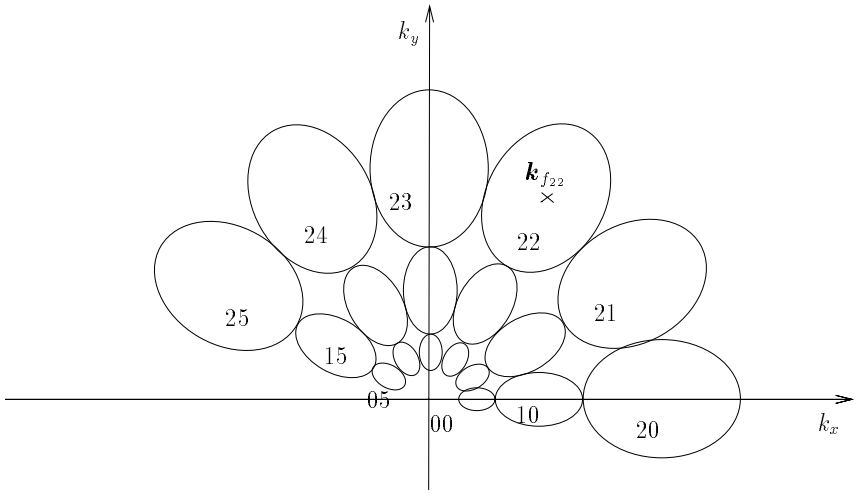

Bild 5: Filterhierarchie für eine Menge von komplexen Gaborfiltern nach (Theimer & Mallot, 1995)

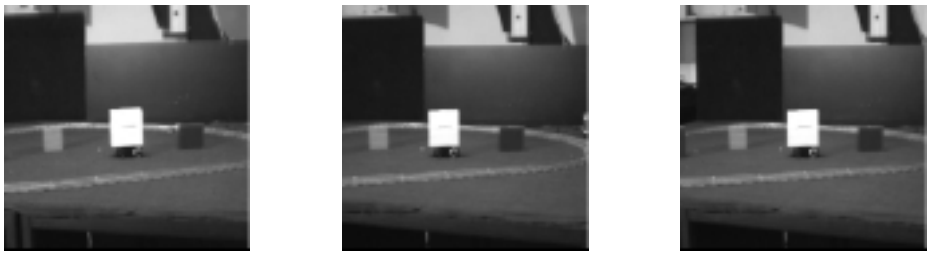

Bild 6: Ergebnisse für die Vergenzbewegungen. Links und in der Mitte die Bilder der linken und rechten Kamera, rechts das Bild der rechten Kamera, in dem sich nach der Vergenzsteuerung das Objekt im Mittelpunkt des Bildes befindet.

Filter gegenüber Verschiebungen in Richtung der Filterorientierung am empfindlichsten ist. Ist nun das eine Stereobild nicht nur in eine Richtung gegenüber dem anderen verschoben, läßt sich mit nur einem Filter die Verschiebung nicht exakt rekonstruieren. Dann benötigt man mehrere Filter aus einem Filterring (vgl. Bild 5). Anhand der so berechneten Phasenverschiebungen läßt sich über eine Fehlerminimierung die Disparität schätzen.

Weitere Verbesserungen des Ergebnisses lassen sich durch die Kombination der Filterantworten von Filtern aus verschiedenen Filterstufen erzielen. Wie schon erwähnt wurde, besitzen Filter mit höheren Mittenfrequenzen auch eine größere räumliche Ausdehnung im Frequenzbereich. Aufgrund der Unschärferelation nimmt die räumliche Ausdehnung dieser Filter im Ortsbereich mit einer Vergrößerung der Bandbreite im Frequenzbereich ab. Deshalb können mit diesen Filtern nur kleinere Disparitäten detektiert werden. Beginnt man nun mit einem Filter niedriger Mittenfrequenz, der zwar einen großen, dafür aber unpräzisen Wertebereich für die Disparitäten besitzt, kann damit eine grobe Schätzung der Disparität vorgenommen werden. Verschiebt man dann das Bild um den dabei ermittelten Disparitätswert, kann mit einem Filter nächster Stufe die Schätzung präzisiert werden.

Um die Beeinflussung des Ergebnisses durch falsche Schätzungen zu reduzieren, werden die einzelnen Disparitätsschätzungen mit Hilfe von Konfidenzwerten bewertet. Dazu wird zum Beispiel untersucht, ob die Amplitudenantworten annähernd übereinstimmen oder diese überhaupt oberhalb eines Schwellwertes liegen.

Anhand der Disparitätsschätzung läßt sich dann anhand von trigonometrischen Beziehungen der Winkel α berechnen, um den die Vergenzachse bewegt werden muß, damit sich die Disparität zwischen den beiden Stereobildern reduziert.

Bild 6 zeigt die Ergebnisse für die Reduktion der Disparität mit Hilfe der Vergenzsteuerung, wobei hier wie bei (Theimer & Mallot, 1995) die Filtergröße mit der Bildgröße übereinstimmte und die räumlichen Ausdehnung des Filters in x–Richtung 25 Pixel betrug.

5 Experimente und Bewertung

Für die Beurteilung der Leistungsfähigkeit des Gesamtsystems wurde die Stichprobe mit den 100 gelesenen Sätzen ausgewertet (vgl. Abschnitt 3.2). Die Experimente des Bildverarbeitungsteils wurden mit Hilfe der Kameraplattform BiSight der Firma TRC durchgeführt. Dieser Stereokopf besitzt vier mechanische und für jede Kamera zwei optische Freiheitsgrade, die innerhalb eines geschlossenen Rückkopplungssystem regelbar sind. Zusätzlich läßt sich die Einstellung der Blende innerhalb eines offenen Rückkopplungssystems verändern.

Kombiniert man die Klassifikationsergebnisse der einzelnen SKBs für die vom Worterkenner auf die optimale Wortkette abgebildeten Sätze miteinander, so ergibt sich eine Erkennungsrate von 86 %. Bemerkenswert ist, daß diese über der Satzerkennungsrate von 76 % liegt, was die Robustheit der SKBs gegenüber Worterkennungsfehlern verdeutlicht.

Bei der Berücksichtigung der Sätze, die von den SKBs korrekt klassifiziert wurden, löste das Stereokamerasystem die Objektlokalisation und die Fovealisierung in allen Fällen korrekt. Bei der Vergenzsteuerung ergab sich in 80 % der Fälle eine korrekte Lösung. Dies ist darauf zurückzuführen, daß der Kontrast zwischen dem dunklen Hintergrund und blauem oder rotem Objekt sehr gering war. Dadurch erfolgte oftmals keine ausreichende Reduktion der Disparität, woraus sich fehlerhafte Vergenzwinkel ergaben.

Bei Rechenzeitmessungen auf einer HP735 ergab für die Objektlokalisation eine Rechenzeit von 0.6 s, für die Fokussierung von 25 s, für das Zoomen von 166 s und eine Iteration zur Adaption des Vergenzwinkel von 0.8 s. Dabei wird beim Zoomen ein Großteil der Zeit für die Fokussierung nach jedem Zoomschritt benötigt.

6 Zusammenfassung und Ausblick

In diesem Beitrag haben wir ein prototypisches System vorgestellt, in dem einem Stereokamerasystem bestimmte Aufgaben gestellt werden können. Hierbei handelt es sich um die Lokalisation von Objekten, die Fovealsierung und die Beurteilung von Lageverhältnissen zwischen zwei Objekten. Die Auswahl der Aufgabe für das Kamerasystem sowie der betroffenen Objekte erfolgt über gesprochen–sprachliche Äußerungen.

Sowohl Sprach– als auch Bildanalyse des gesamten Systems sind so flexibel, daß sie ohne großen Aufwand auf neue Domänen oder andere Objekte oder Aktionen adaptiert werden können. Es kann auch statt Objektidentifikatoren eine Hierarchie von Attributen verwendet werden, die die Objekte beschreiben. Was die Sprachanalyse angeht, braucht man lediglich eine entsprechend etikettierte Stichprobe; die SKBs können dann vollautomatisch trainiert werden. Ein System wie wir es hier beschrieben haben, sollte letztendlich spontane Sprache verarbeiten können. Das Training anhand automatisch generierter Sätze erlaubt in kurzer Zeit einen Prototypen zu realisieren, der dann in on–line Tests dazu verwendet werden kann, weitere Sprachdaten zu sammeln.

Eine ausführlichere Beschreibung der Experimente und Ergebnisse sowie der zugrundeliegenden mathematischen Verfahren sind in (Ahlrichs, 1996) nachzulesen.

Literatur

Ahlrichs, U. (1996). Sprachgesteuerte Fovealisierung und Vergenz . Tech. rep., Diplomarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, Erlangen.

Breiman, L. (1984). *Classification and Regression Trees*. Wadsworth, Belmont CA.

Corazza, A., Federico, M., Gretter, R., & Lazzari, G. (1993). Design and acquisition of a task–oriented spontaneous–speech data base. In Roberto, V. (Ed.), *Intelligent Perceptual Systems, Lecture Notes in Artificial Intelligence*, pp. 196–210 Heidelberg. Springer Verlag.

Hansen, M., & Sommer, G. (1996). Real-time vergence control using local phase differences. *Machine Graphics and Vision*, *5*(1/2), 51–63.

Kawamura, K., & Iskarous, M. (1994). Trends in Service Robots for the Disabled and the Elderly. In *Intelligent Robots and Systems*, pp. 1647–1654 München.

Krotkov, E., & Bajcsy, R. (1993). Active vision for reliable ranging: Cooperating focus, stereo and vergence. *International Journal of Computer Vision*, *11*(2), 187–203.

Krotkov, E. (1989). *Active Computer Vision by Cooperative Focus and Stereo*. Springer Verlag.

Kuhn, R., & De Mori, R. (1995). The Application of Semantic Classification Trees to Natural Language Understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *17*, 449–460.

Kuhn, T. (1995). *Die Erkennungsphase in einem Dialogsystem*, Vol. 80 of *Dissertationen zur künstlichen Intelligenz*. infix, St. Augustin.

Mahlmeister, U., Pahl, H., & Sommer, G. (1996). Color–orientation indexing. In *DAGM 1996, Heidelberg*, pp. 3–10 Heidelberg.

Mast, M., Kummert, F., Ehrlich, U., Fink, G., Kuhn, T., Niemann, H., & Sagerer, G. (1994). A Speech Understanding and Dialog System with a Homogeneous Linguistic Knowledge Base. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *16*(2), 179–194.

Polifroni, J., Seneff, S., & Zue, V. (1991). Collection of Spontaneous Speech for the ATIS Domain and Comparative Analyses of Data Collected at MIT and TI . In *Proc. Speech and Natural Language Workshop* San Mateo, California. Morgan Kaufman.

Rieck, S. (1995). *Parametrisierung und Klassifikation gesprochener Sprache*, Vol. 10: Informatik/Kommunikationstechnik no. 353 of *Fortschrittberichte*. VDI Verlag, Düsseldorf.

Schukat-Talamazzini, E., Kuhn, T., & Niemann, H. (1994). Speech Recognition for Spoken Dialogue Systems. In Niemann, H., De Mori, R., & Hanrieder, G. (Eds.), *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM / FORWISS Workshop*, pp. 110–120. Infix.

Socher, G., Fink, G., Kummert, F., & Sagerer, G. (1996). Talking about 3D Scenes: Integration of Image and Speech Understanding in a Hybrid Distributed System. In *Proc. Int. Conf. on Image Processing*, pp. 809–812 Lausanne.

Stopp, E., & Laengle, T. (1995). Natürlichsprachliche Instruktionen an einen autonomen Serviceroboter. In *Autonome Mobile Systeme*, pp. 299–307 Karlsruhe.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Theimer, W., & Mallot, H. (1995). Phase–based binocular vergence control and depth reconstruction using active vision. *Computer Vision, Graphics, and Image Processing*, *60*(3), 343–358.