Automatische farbbasierte Extraktion natürlicher Landmarken und 3D–Positionsbestimmung auf Basis visueller Information in Indoor–Umgebungen

Joachim Denzler*, Matthias Zobel†
Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg
{denzler,zobel}@informatik.uni-erlangen.de

Zusammenfassung

In diesem Beitrag wird ein Verfahren zur automatischen Identifikation von Landmarken in Indoor–Umgebungen auf Basis von Farbhistogrammen vorgestellt. Es wird gezeigt, daß die automatisch definierten Landmarken mittels stochastischer Simulation robust wiedergefunden werden können und somit eine Selbstlokalisation eines autonomen mobilen Systems möglich ist. Mittels Tilt/Vergenz–Bewegungen wird die relative 3D–Position eines Binokular–Systems zur Landmarke rekonstruiert, was einen ersten Schritt zur automatischen Landkartenerzeugung darstellt.

1 Einleitung

Eine der wichtigsten Anforderungen, die an autonome mobile Systeme üblicherweise gestellt wird, ist die Navigation in einer dynamischen, sich verändernden Umgebung. Unter dem Begriff Navigation faßt man die Einzelaufgaben Selbstlokalisation, Pfadplanung und Lokomotion zusammen [1]. Im Bereich der Robotik gibt es zahlreiche Ansätze, die auf Grundlage von Ultraschall–, Infrarot– und Lasersystemen eine kollisionsfreie Lokomotion realisiert haben [1]. Die Selbstlokalisation geschieht in vielen Fällen über Odometrie, die in regelmäßigen Abständen über eine fest vorgegebene Karte der Umgebung und künstlich angebrachte, bzw. manuell definierte natürliche Landmarken korrigiert wird. Diese Landmarken sind in vielen Fällen so gewählt, daß sie mit den verwendeten Sensoren einfach und robust erfaßt und bei Bedarf auch unterschieden werden können.

Ein im Bereich der autonomen mobilen Systeme bis heute ungelöstes Problem stellt die automatische Auswahl und Extraktion natürlicher Landmarken und deren Verarbeitung zur automatischen Erstellung einer 3D–Karte der Umgebung dar. In sogenannten Indoor–Umgebungen, die man typischerweise bei Servicerobotern, z. B. in Krankenhäusern, vorfindet, bietet sich eine visuelle Exploration der Umgebung auf Basis von Farbstrukturen an.

Farbhistogramme sind ein Standardansatz zur Erfassung von Farbeindrücken innerhalb einer Szene. Anhand der Einsatzgebiete Objektlokalisation [3] und Objektverfolgung mit einer Kamera [2] wurde deren Anwendbarkeit mehrfach demonstriert. Der in diesem Beitrag

*Diese Arbeit wurde teilweise durchgeführt im Rahmen des Projektes DIROKOL, gefördert von der Bayerischen Forschungsstiftung

†Diese Arbeit wurde durchgeführt im Rahmen des Sonderforschungsbereichs SFB 603, gefördert von der DFG

vorgeschlagene Ansatz zur automatischen Extraktion von natürlichen Landmarken und deren 3D–Vermessung bedient sich dieses klassischen Vorgehens. Dabei wird das Farbhistogramm einer Arbeit von Heisele folgend [2] erweitert. Als Landmarken werden solche Bins definiert, die eine gewisse Anzahl von Einträgen besitzen und zusätzlich einem kompakten Bereich im Bild entsprechen, d. h. eine kleine Varianz in der Position aufweisen.

Die so automatisch detektierten Landmarken werden anschließend mittels eines sich auf der autonomen mobilen Plattform montierten binokularen Pan/Tilt/Vergenz–Systems vermessen. Dazu erfolgt eine entsprechende Tilt/Vergenz–Bewegung, um in beiden Kameras die Landmarke in die Bildmitte zu bewegen. Über die dann vorliegenden Vergenzwinkel kann die 3D–Position relativ zum autonomen mobilen System bestimmt werden. Die 3D–Position wird im folgenden bei der automatischen 3D–Umgebungskartenerstellung benötigt. Neben der eigentlichen Landmarke wird deren Position im 3D relativ zum mobilen System eingetragen. Somit lassen sich während der Navigation Mehrdeutigkeiten in den Landmarken, z. B. verschiedene Türen oder Türschilder auflösen, indem nach der Position im 3D gesucht wird, die einer beobachteten Konfiguration von Landmarken am besten entspricht.

2 Farbbasierte Extraktion natürlicher Landmarken

Bei der Entwicklung eines Verfahrens zur automatischen Landmarkenidentifikation müssen zwei Punkte beachtet werden: es müssen genügend Landmarken gefunden werden, um eine Selbstlokalisation zu gewährleisten und eine spezielle Landmarke muß effizient und robust wiedergefunden werden können.

Der hier vorgestellte Ansatz baut auf sogenannten Farbhistogrammen auf [3]. Die Motivation dafür kann in den verstärkten Bemühungen in der Bildverarbeitungsgemeinschaft gesehen werden, Farbnormierungsalgorithmen zu entwickeln und somit einen der bisher entscheidensten Nachteile — die Varianz in den gemessenen Farbwerten — zu beheben. Außerdem existieren Verfahren zur Objekterkennung mittels Histogrammrückprojektion, die in einer späteren Phase der Arbeit zur Identifikation einer bestimmten Landmarke durch Vergleich von Histogrammen eingesetzt werden kann. Neben diesen eher theoretischen Motivationsgründen spielen praktische Überlegungen zur effizienten, hierarchischen Berechnung unter Echtzeitbedingungen eine wichtige Rolle, die von Histogrammen im allgemeinen erfüllt werden.

Im folgenden wird davon ausgegangen, daß die Umgebung, in der sich das autonome mobile System bewegen soll, relativ einfach strukturiert ist und eine genügend große Variation in der vorhandenen Farbe beobachtet werden kann. Beide Anforderungen werden sowohl für Szenen aus dem Krankenhausbereich als auch für die Büroflurumgebung erfüllt, die in Abschnitt 5 als Testumgebung dient. Als Landmarken wird das häufige Auftreten eines Farbwerts innerhalb eines räumlich begrenzten Gebiets im Farbbild definiert. Dies kann anhand von Farbhistogrammen erkannt werden, wenn neben der Häufigkeit des Treffer innerhalb eines Histogramm–Bins auch der Mittelwert der Positionen der beitragenden Farbpixel und deren Positionsvarianz berechnet wird. Ein vergleichbares Vorgehen findet sich in [2] zur Verfolgung von Personen mittels Farbhistogrammen.

Mittels eines solchen Farbhistogramms mit räumlicher Information werden solche Histogramm–Bins als Landmarken definiert, deren Anzahl Treffer eine gewisse Schwelle überschreitet und deren Positionsvarianz unter einer gewissen weiteren Schwelle liegt. Der Mittelwert der Positionen definiert dann die Position der Landmarke im Farbbild. Unterschiedliche Histogramm–Bins, deren Positionsmittelwerte sich nur wenig unterscheiden, können zusammengefaßt und somit neben einfarbigen Landmarken auch bunte Objekte identifiziert werden.

3 Visuelle Selbstlokalisation mittels stochastischer Simulation

Die identifizierten Landmarken aus dem letzten Abschnitt können zur Selbstlokalisation verwendet werden. Dafür wird in Abhängigkeit von der Position $\boldsymbol{X}$ und der Blickrichtung $\boldsymbol{R}$ des mobilen Systems eine Wahrscheinlichkeitsverteilung über die Histogramm–Einträge definiert. Jedes geeignete Histogramm–Bin b_{ijk}, d. h. eine identifizierte Landmarke, repräsentiert somit eine Verteilung über die Positionen $\boldsymbol{x}$ im Bild, so daß $f(\boldsymbol{x}) = b_{ijk}$. Die Indizes i, j, k kennzeichnen die Komponenten des Farbvektors, d. h. im Falle von RGB–Farbwerten den Rot–, Grün– und Blaukanal. Natürlich ist dieses Vorgehen auch auf andere Farbräume übertragbar. Für die Verteilung der Positionen der Farbpixel eines Histogramm–Bins b_{ijk} nehmen wir eine Normalverteilung an. Die Parameter der Normalverteilung, Mittelwert und Varianz, werden aus den Informationen des Histogramm–Bins direkt übernommen. Das Problem der Selbstlokalisation kann jetzt über eine maximum a posteriori Schätzung

$$(\boldsymbol{X}^*, \boldsymbol{R}^*) = \underset{\boldsymbol{X}, \boldsymbol{R}}{\operatorname{argmax}}\, p(\boldsymbol{X}, \boldsymbol{R}|B) = \underset{\boldsymbol{X}, \boldsymbol{R}}{\operatorname{argmax}}\, \frac{p(B|\boldsymbol{X}, \boldsymbol{R})p(\boldsymbol{X}, \boldsymbol{R})}{p(B)}$$

wahrscheinlichkeitstheoretisch beschrieben werden. Zu beachten ist hierbei, daß keinerlei zeitliches oder räumliches Kontextwissen in dieser Modellierung einfließt. Dies ist Gegenstand der aktuellen Untersuchungen.

Ausgehend von den definierten Landmarken wird zur Selbstlokalisation des Systems folgendes Vorgehen auf Basis einer stochastischen Simulation eingeschlagen. Unter der Erwartungshaltung sich an einer Position $\boldsymbol{X}$ mit Blick in Richtung $\boldsymbol{R}$ zu befinden, erwartet man, die korrespondierenden Landmarken in Form von signifikanten Histogramm–Bins zu beobachten. Aus der bekannten Dichte $p_{b_{ijk}}(\boldsymbol{x}|\boldsymbol{X}, \boldsymbol{R})$ der Verteilung der Farbpixel im Bild werden Positionen im Bild generiert, an denen die Farbe dem Farbbereich des Bins b_{ijk} entsprechen müßte. Die Anzahl der Treffer ergibt damit die Wahrscheinlichkeit für das Vorhandensein der Landmarke b_{ijk}.

Da normalerweise mehr als eine Landmarke beobachtet werden kann, muß die Kombination mehrere Landmarken geeignet modelliert werden. Für den ersten Schritt des Ansatzes wird Unabhängigkeit des Auftretens von Landmarken angenommen, d. h.

$$p(B|\boldsymbol{X}, \boldsymbol{R}) = \prod_{b_{ijk}} p(b_{ijk}|\boldsymbol{X}, \boldsymbol{R})$$

In Zukunft wird diese Unabhängigkeitsannahme fallen gelassen.

4 3D–Positionsbestimmung mittels Vergenz

Zur Erstellung einer Umgebungskarte, die bei der Navigation eines autonomen Systems eine wichtige Rolle spielt, muß Information über die räumliche Lage der einzutragenden Elemente vorhanden sein. Diese Information kann sowohl absolut, bezüglich eines Weltkoordinatensystems, als auch relativ, bezüglich des autonomen Systems, eingetragen werden. In diesem Abschnitt wird gezeigt, wie die Lage der detektierten Landmarken relativ zum verwendeten Binokular–System mittels geeigneter Tilt/Vergenz Bewegungen ermittelt werden kann. Im Abschnitt 5 wird auf die Genauigkeit der Positionsbestimmung eingegangen.

Voraussetzung für die Positionsbestimmung auf die hier vorgestellt Weise ist, daß die internen und externen Parameter der beiden Kameras bekannt sind. Dies kann man durch

Kalibrierung der Kameras nach der Methode von Tsai sicherstellen [4], wenn man beide Kameras mit dem selben, gleich positionierten Referenzmuster kalibriert. Man erhält aus der Kalibrierung für jede Kamera eine Rotationsmatrix[1] $\mathbf{R}_{w\to i}$ und einen Translationsvektor $\mathbf{t}_{w\to i}$, $i = 1, 2$, mit denen die Transformation von Weltkoordinaten $\mathbf{x}_w$ in Kamerakoordinaten $\mathbf{x}_i$ nach

$$\mathbf{x}_i = \mathbf{R}_{w\to i}\mathbf{x}_w + \mathbf{t}_{w\to i}.$$

gegeben ist. Damit läßt sich die Basis des Binokular–Systems, der Vektor $\mathbf{b}_w$ zwischen den beiden Koordinatenursprüngen der Kameras, berechnen als

$$\mathbf{b}_w = \mathbf{t}_{w\to 2} - \mathbf{t}_{w\to 1}.$$

Die Mitte der Basis stellt den für die Positionsbestimmung verwendeten Referenzpunkt dar.

Zur Bestimmung der Position eines Punktes im Raum werden die optischen Achsen, also die z–Achsen der Koordinatensysteme der beiden Kameras auf diesen Punkt ausgerichtet, d. h. die optischen Achsen schneiden sich in diesem Punkt. Der Vektor zwischen der Basismitte und dem Schnittpunkt ist der gesuchte Positionsvektor, aus dem z. B. der Abstand des Punkts zur Basis berechnet werden kann.

Die oben genannte Justierung der optischen Achsen erfolgt durch entsprechende Drehungen der beiden Kameras in vertikaler und horizontaler Richtung. Als Maß für die durchzuführende Änderung der aktuellen Kameralagen wird die Abweichung des auf die beiden Bildebenen projizierten Raumpunkts von den Hauptpunkten der Kameras benutzt. Der Hauptpunkt einer Kamera ist der Durchtrittspunkt der optischen Achse durch die Bildebene. Die Pixelkoordinaten der Hauptpunkte sowie die Brennweiten der Kameras werden ebenfalls durch die Kamerakalibrierung bestimmt.

Aus den ermittelten Abweichungen und den internen Kameraparametern kann man durch geeignete Transformationen die jeweils erforderlichen Rotationswinkel bestimmen, welche wiederum in entsprechende Achsenbewegungen des Binokular–Systems überführt werden können. Beschreibt man die optischen Achsen beider Kameras und die Basis im Koordinatensystem der linken Kamera, so läßt sich der Schnittpunkt der optischen Achsen $\mathbf{s}_1$ durch Gleichsetzen ihrer Geradengleichungen und Lösen des entsprechenden Gleichungssystems ermitteln. Damit ergibt sich der gesuchte Positionsvektor $\mathbf{p}$ einfach aus

$$\mathbf{p} = \mathbf{s}_1 - \frac{1}{2}\mathbf{b}_1.$$

Die vorgestellte Methode konnte für die Experimente nicht ohne weiteres auf das verwendete Binokular–System übertragen werden. Es sind Anpassung und Annahmen notwendig, die in Abschnitt 5 skizziert sind.

5 Experimente und Ergebnisse

Die Experimente gliedern sich in folgende drei Teile: automatische Landmarkenidentifikation auf Basis von Farbclustern, Selbstlokalisation durch stochastische Simulation und 3D–Positionsermittlung durch Vergenz.

Landmarkenidentifikation Untersucht wurden 300 Bilder aus einer Büroflurumgebung. Die Abtastrate betrug 1 Hz. Für jedes Bild wurde ein Satz von Landmarken automatisch definiert. Die durchschnittliche Anzahl betrug 4,2 Landmarken pro Bild, maximal traten 17

[1] Der Index bezeichnet das jeweils gültige Koordinatensystem: 1 = Kamera links, 2 = Kamera rechts, w Weltkoordinaten

Landmarken pro Bild auf. Bei einer großen Anzahl von Landmarken, trat Überadaption ein, so daß nur dieses Bild wiedererkannt werden konnte, nicht aber ähnliche Bilder. Bei insgesamt 28 Bildern wurden keine Landmarken gefunden. Bei 22 dieser 28 Bilder führte die mobile Plattform eine Drehung durch, so daß die Kamera nur auf eine Wand blickte, die keine signifikanten Merkmale enthielt.

Die durchschnittliche Zeit für die Landmarkenidentifikation betrug 230 msec, für jeweils 100 Vergleiche mit stochastischer Simulation werden 10 ms benötigt (gemessen auf einer SGI O^2 mit R10000 Prozessor). In Abbildung 1 ist ein Beispiel einer erfolgreichen Extraktion von signifikanten Landmarken zu sehen. Im Abbildung 1, links, wurde die Tafel und drei Poster an der Tür extrahiert. Nach ca. 4 Minuten bewegte sich das System an der selben Position im Flur und konnte eines der Poster sowie die Tafel erneut als signifikante Landmarke identifizieren. Zu sehen ist, daß die Tafel in mehrere Teillandmarken zerfällt, da die Farbwerte in mehrere unterschiedliche benachbarte Histogramm-Bins fallen.

Selbstlokalisation Mittels den in der Landmarkenidentifikationsphase gewonnen Landmarken wurde eine Selbstlokalisation durchgeführt, indem für alle Bilder eine Bewertung bezogen auf die Landmarken mittels stochastischer Simulation vorgenommen wurde (vergleiche Abschnitt 3). Erwartet wird, daß jeweils das Originalbild als beste Übereinstimmung wiedergefunden wird, und ähnliche Bilder in der Rangfolge vorne stehen werden.

Zur Auswertung wurde eine Verwechslungsmatrix (s. Abbildung 2) erstellt. Die Zeile entspricht dem Originalbild, die Spalte dem Vergleichsbild. Die Farbfelder kennzeichnen die unterschiedlichen Bereiche innerhalb des Flurbereichs. In der Verwechslungsmatrix bedeuten dunkle Grauwerte hohe, helle Grauwert schlechte Übereinstimmung. Man erkennt in Abbildung 2 den erwünschten Effekt, daß die Hauptdiagonale stark besetzt ist und auch Nebendiagonalelemente an ähnlichen Positionen (kodiert durch die Farbwerte) Übereinstimmung kennzeichnen. Anhand der Farbverläufe in einigen Zeilen erkennt man diejenigen Bilder, für die sehr viele Landmarken definiert wurden und somit eine Wahrscheinlichkeit für Übereinstimmung ungleich Null nur für das Originalbild berechnet wurde. Aufgrund der Sortierung der Übereinstimmung ergibt sich dann der zu beobachtende signifikante Grauwertverlauf.

In Abbildung 4 findet man zwei Beispiele für die Selbslokalisation: das Originalbild mit dem Rang in der Übereinstimmung sowie eines der ähnlichsten Bilder.

3D–Positionsbestimmung Wie am Ende von Abschnitt 4 angedeutet, mußte das Verfahren zur 3D–Positionsbestimmung an die vorhandene Experimentierumgebung angepaßt werden.

Auf Grund des verwendeten Kalibriermusters kann es bei der Bestimmung der externen Kameraparameter zu Ergebnissen kommen, die für eine einzelne Kamera durchaus anwendbar sind, für die Kopplung von mehreren Kameras miteinander jedoch nicht. Damit verbunden ist, daß die Bestimmung der Basis mit einem Fehler behaftet ist, der direkt in die Berechnung der 3D–Position eingeht. Daher wurden für die Experimente die Annahmen gemacht, daß die beiden Kameras einen Basisabstand[2] von 253 mm haben und die x–Achse der linken Kamera mit der Basis zusammenfällt. Die optische Achse der linken Kamera muß dazu während der Kalibrierung möglichst senkrecht auf der physikalischen Basis stehen.

Das verwendete Binokular–System weist die Tatsache auf, daß die beiden Kameras keine voneinander unabhängigen Tilt–Bewegungen vollführen können. Vielmehr können beide Kameras, wegen ihrer mechanischen Befestigung, nur um den selben Winkel vertikal geneigt werden. Damit ergibt sich das Problem, daß die optischen Achsen nicht zwangsläufig zum Schnitt gebracht werden können. Als Lösung bietet sich an, den Schnittpunkt der beiden optischen Achsen als den Mittelpunkt der Verbindungsstrecke anzusehen, die senkrecht auf

[2]Basisabstand laut Herstellerangaben

den beiden Geraden steht. Weiterhin wird anstelle der einzelnen vertikalen Pixeldifferenzen die gemittelte Differenz sowie die gemittelte Brennweite zur Festlegung der erforderlichen Tilt–Bewegung verwendet.

Als Maß für die Genauigkeit der Positionsbestimmung wurde die Entfernung von Objekten zur Basismitte des Binokular–Systems gewählt, da die Referenzwerte manuell einfach zu ermitteln sind. Das System vollführte einen horizontalen Schwenk über eine Laborszene, wobei an neun Schwenkpositionen automatisch Landmarken extrahiert wurden. In der Szene befanden sich drei Objekte in bekannter Entfernung. Abbildung 3 zeigt die Laborszene mit den bekannten Objekten und den ermittelten Referenzentfernungen. Die Untersuchung der Szene erfolgte mehrmals, wobei unterschiedliche Zoom–Einstellungen und Neigewinkel der Kameras eingestellt wurden. In Tabelle 1 sind die Ergebnisse der Entfernungsmessung für die drei Referenzobjekte aufgeführt. Es zeigte sich, daß die Entfernung eines Objekts im Mittel mit einer Genauigkeit von 2,2 % gemessen werden konnte.

Landmarke	Detektionen	$\min x$	$\max x$	$\overline{x}$	σ_x	x_0	$\frac{x_0-\overline{x}}{x_0}$
Gelb	40	174 cm	223 cm	202 cm	7 cm	207 cm	2,4 %
Grün	21	150 cm	210 cm	189 cm	13 cm	195 cm	3,1 %
Blau	20	151 cm	296 cm	204 cm	30 cm	206 cm	1,0 %

Tabelle 1: Ergebnisse der Positionsbestimmung der drei Referenzobjekte in der Laborszene. x_0 bezeichnet die Referenzentfernung, x die experimentell ermittelte Entfernung. Die letzte Spalte gibt die mittlere relative Genauigkeit an.

6 Zusammenfassung und Ausblick

Dieser Beitrag stellte einen Ansatz zur automatischen, unüberwachten Identifikation und Extraktion von Landmarken in einer Indoor–Umgebung vor. Es wurde gezeigt, daß sich Farbhistogramme, die um räumliche Verteilungsinformation erweitert werden, geeignet sind, um Landmarken zu definieren, die effizient mittels stochastischer Simulation wiedergefunden werden können. Zusätzlich wird die 3D–Position der Landmarke relativ zum mobilen System mit einer Genauigkeit von bis zu 1% durch eine Tilt/Vergenz–Bewegung ermittelt.

Diese Ergebnisse stellen die Grundlage für eine Weiterentwicklung des Systems in Richtung einer automatischen 3D–Landkartenerstellung dar. Die 3D–Information fließt dann auch direkt in die maximum a posteriori Schätzung aus Abschnitt 3 ein und ermöglicht die Unterscheidung bei visuellen Mehrdeutigkeiten.

Literatur

[1] J. Borenstein, H.R. Everett, and L. Feng. *Navigating Mobile Robots.* A K Peters, Wellesley, Massachusetts, 1996.

[2] B. Heisele, U. Kressel, and W. Ritter. Tracking non-rigid moving objects based on color cluster flow. In *IEEE Computer Vision and Pattern Recognition*, pages 257–260, 1997.

[3] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, November 1991.

[4] R.Y. Tsai. A versatile camera calibration technique for high–accuracy 3d machine vision metrology using off–the–shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, 1987.

Abbildung 1: Automatische Landmarkenextraktion im Bild 17 und Bild 245 (Abtastrate: 1 Hz)

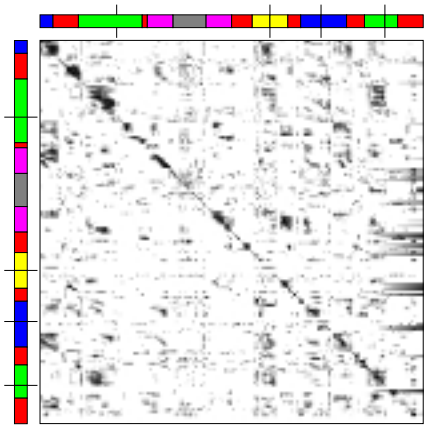

Abbildung 2: Verwechslungsmatrix der Positionen bei der Selbstlokalisation

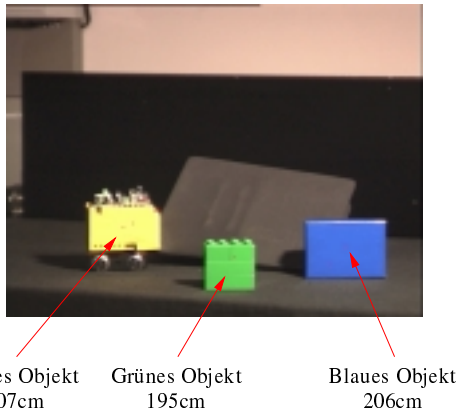

Abbildung 3: Referenzentfernungen der Objekte zur Basis der Binokular–Systems

Abbildung 4: Ähnliches Bild zum Originalbild zusammen mit den Plazierungen bei der Bewertung