

Erkennung dynamischer Gesten zur Kommandierung mobiler Roboter

Markus Ehrenmann, Tobias Lütticke und Rüdiger Dillmann

Universität Karlsruhe, Geb. 40.28, Kaiserstraße 12, 76128 Karlsruhe, Deutschland
{ehrenman,tobias.luetticke,dillmann}@ira.uka.de,
WWW Home Page: <http://wwwipr.ira.uka.de/~{ehrenman,dillmann}>

1 Einleitung

Mobile Robotersysteme werden heute meist über graphische Oberflächen in Standard-PCs, PDAs oder Teachpanel kommandiert. Auditive und gestenbasierte Kommandierungen sind seit wenigen Jahren hochaktuelle Forschungsthemen mit dem Ziel, die Schnittstellen zwischen Menschen und Maschinen direkter und intuitiver zu gestalten. Entsprechend den Bewegungen, die Menschen beim Einweisen von Fahrzeugen machen, wird am Institut für Prozessrechentechik, Automation und Robotik der Einsatz dynamischer Gesten zur Anweisung einer mobilen Plattform untersucht. „Dynamisch“ soll hierbei bedeuten, dass bei der Interpretation der Benutzerhandlungen hier ausschliesslich die Verfahrbahn einer der beiden Hände bedeutungstragend ist. Die Gelenkstellungen von Fingern und Handgelenken fließen nicht in die Gestenklassifikation ein.

Die Hauptprobleme bei dieser Spielart der Gestenerkennung lassen sich in drei Punkten zusammenfassen:

- Schnelle und robuste Verfolgung der Benutzeraktionen resp. der Handtrajektorie.
- Training von Modellparametern mit mehreren oder nur einem Benutzer.
- Klassifikation der aufgezeichneten Trajektorie im Hinblick auf die trainierten Modellgesten.

Ein wesentlicher Aspekt des letzten Punktes ist es, die Zuordnung bedeutungsloser Trajektorien zu einem Referenzmodell zu verhindern. Außerdem sollte die Klassifikation unter Echtzeitbedingungen stattfinden.

1.1 Stand der Technik

Die gegenwärtigen Methoden zur Lösung dieser Schwierigkeiten sollen kurz vorgestellt werden. Als spezielles und herausforderndes Problem gilt das bildbasierte Verfolgen der menschlichen Hand. Durch die Vielzahl von Gelenken und die komplexe Struktur ist nicht nur ein hoher Grad an Artikulationsmöglichkeiten gegeben, sondern auch von Verdeckungen. Dadurch wird eine genaue Zuordnung der einzelnen Finger erschwert. Bei der Erkennung von Gesten wird die Fingerstellung jedoch oft nicht betrachtet.

Handverfolgung: Ansätze zum Handtracking basieren auf Projektionen der Shilhouette [16, 4], Bewegungsinformationen [22] oder Farbmarkern. In manchen Applikationen wird eine vollständige 3D Rekonstruktion mit aufwändigen Handmodellen versucht, dies ist jedoch nicht immer in Echtzeit möglich [18]. In den meisten Fällen ist keine Information über die Gelenkstellungen der Finger notwendig, man stützt sich dann auf Berechnungen des optischen Flusses [12], der Korrelation von Stereofarbbildern [1] oder Farbsegmentierungen [19].

Gestenerkennung: Gesten können als Oberbegriff für durch Handstellungen ausgedrückte Symbole zur Kommandierung, Instruierung oder Dialogführung angesehen werden. Bevorzugt werden zur Erkennung sichtgestützte Systeme oder Datenhandschuhe eingesetzt. Einen Überblick über die technische Realisierung verschiedener gestenerkennender Systeme hat Kohler zusammengestellt [8]. Hier werden jedoch nur bildverarbeitende Systeme berücksichtigt. Bei Handzeichen lässt sich eine Unterscheidung in statische und dynamische Gesten treffen, bei denen sich der bedeutungstragende Teil aus der Fingerstellung bzw. aus der Handbewegung ergibt.

Statische Gesten: Kestler [6] benutzt Kamerabilder, um statische Handgesten zu erkennen und klassifizieren. Die Klassifikation geschieht nach einer Vorverarbeitung durch Vergleiche von Tensoren der Gestenmuster mit dem aktuellen Bild wie bei einer Eigenraummethode. Wird die Handverfolgung durch Konturmodelle realisiert, lassen sich direkt aus der aktuellen Merkmalsverteilung Parameter extrahieren, mit denen eine Geste beschreibbar wird [2, 5]. Dieser Ansatz entspricht im Wesentlichen dem Anpassen elastischer Graphen an Bildmerkmale in [21].

Dynamische Gesten: Ein System zur Klassifizierung der koreanischen Zeichensprache, bei der Fingerstellungen und Handbewegungen beider Hände Bedeutungsträger sind, ist mit Hilfe von Datenhandschuhen und magnetfeldbasierten Positionssensoren unter Verwendung von fuzzy min-max neuronalen Netzwerken realisiert worden [7]. Es ist jedoch zu beobachten, dass aufgrund der mit neuronalen Netzen schwierig zu erreichenden Erkennung bedeutungsloser Verfahrbahnen zunehmend Hidden-Markov-Modelle zur Klassifikation eingesetzt werden [20]. Meist werden sie mit Standard-Methoden trainiert und getestet (Einführungen in diese Modelle und Methoden finden sich in [3, 15]). Lee erkennt unter Verwendung von Hidden Markov Modellen mit sehr guten Resultaten dynamische Gesten zur Steuerung einer Vortragspräsentation [10]. Um Fehlerkennungen zu verhindern, wird ein Schwellwertmodell eingeführt. Da das Problem der Gestenerkennung bei dynamischen Handzeichen der Erkennung von Handschrift ähnlich ist, werden Methoden auf beiden Seiten in gleicher Weise verwendet [14]. Neuerdings werden Kameras auch am Benutzer befestigt, um Gesten zu erkennen [13].

Im Folgenden soll nach der Vorstellung der an unserem Institut befindlichen Experimentierumgebung mit entsprechenden Bildaufnahme- und Bildverarbeitungs-komponenten der von uns verfolgte Ansatz auf Basis von Hidden-Markov-Modellen hinsichtlich dieser Fragestellungen diskutiert werden.

2 Experimentierumgebung

Bevor die Verfahrbahn dem Klassifikator vorgelegt werden kann, muss die Beobachtung der Hand eines Benutzers erfolgen. Dies soll visuell und ohne besondere Marker geschehen. Dazu kommen die folgenden technischen und algorithmischen Lösungen zum Einsatz.

2.1 Systemkomponenten

Die zur Handverfolgung eingesetzte Sensorik ist noch nicht auf einem mobilen Roboter installiert. Sie besteht bislang aus einem dreh- und neigbaren Handgelenk der Firma Amtec, auf dem zwei Sony 777AP Farbkameras befestigt sind (siehe Abb. 1). Zur Digitalisierung kommen Matrox MeteorII Framegrabber zum Einsatz.

2.2 Bildverarbeitung

Die Detektion der Benutzerhand und die Verfolgung der Handbewegung findet auf der Basis von Hautfarbsegmentierung (siehe z.B. [23]) statt. Die eingehenden *RGB*-Bilder werden nach der Konversion in die *HSI*-Darstellung binarisiert. Als Schwellen hierfür haben sich bei Kunstlichteinstrahlung die Werte 3 bis 31 auf dem Farbtonwert *H* als geeignet erwiesen. Danach wird ein Close-Filter verwendet, um Löcher in dem resultierenden Bild zu schließen.



Abbildung1. Kamerakopf des Verarbeitungssystems

Bedingung vor der Vorführung einer Geste ist, daß nur ein Benutzer vor der Kamera steht. Dann können leicht den drei resultierenden Hautfarbregionen der Kopf des Benutzers und seine Hände zugeordnet werden (siehe Abb. 2). Aus letzteren wird die linke oder rechte zur Verfolgung ausgewählt.



Abbildung2. Schritte der Bildverarbeitung (Ausgangsbild, Farbsegmentierung, Glättung).

Durch die Verwendung lokaler Fenster bei der Bildverarbeitung und der Nebenläufigkeit mehrerer Prozesse konnte die Verarbeitungsrate dieses Schrittes auf durchschnittlich 15Hz auf einem Doppel PentiumIII-System mit 500MHz Taktung gesteigert werden.

2.3 Trajektorienfilterung

Die ermittelte Folge des Schwerpunkts des Handsegments wird einem mehrstufigen Filter zur Glättung und Datenreduktion übergeben. Genauer werden die Punktfolgen zunächst an einen Nachbarschaftsfilter weitergereicht. Dieser verwirft naheliegende, aufeinanderfolgende Positionen.

Da als Eingabe für die Hidden-Markov-Modelle keine zweidimensionalen Koordinaten dienen können, werden die Richtungsvektoren der Segmente zunächst auf ein 16elementiges Eingabealphabet abgebildet (siehe Abb. 3). Diese Anzahl sollte einerseits mächtig genug sein, um die vollzogenen Bewegungen zu repräsentieren, andererseits klein genug, um den Rechenaufwand zu beschränken.

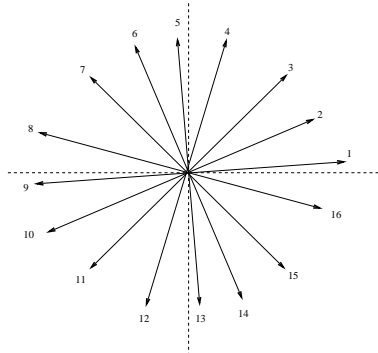


Abbildung3. Das zur Vektorquantisierung verwendete Codebuch mit 16 Wörtern.

Auch diese Folge von Richtungsindizes wird durch einen Identitätsfilter weiter verkleinert. Da jeder Richtungsindex aus einem Richtungsvektor entsteht und dessen Orientierung repräsentiert, reicht ein einziger Index aus, um eine Bewegungsrichtung zu verdeutlichen. Mehrere aufeinander folgende gleiche Indizes sind also redundant und können auf einen einzigen abgebildet werden.

Insgesamt lässt sich so eine Reduktion der Eingabefolgen je nach Gesten zwischen 14% bis 96% erzielen. Dies ist in Abb. 4 verdeutlicht.



Abbildung4. Aufgezeichnete und gefilterte Trajektorie.

Eine für die Handlungserkennung grundlegende Erkenntnis ist die, dass für Menschen Handlungen als Sequenz klar voneinander geteilter Einzelhandlungen wahrgenommen wird und dass die brauchbarste Information zur Interpretation einer Handlung im Zustand des Wechsels zwischen zwei solchen Einzelaktionen vorliegt [11]. Zur besseren Trennung von Anfang und Ende einer Geste ist vor beide Filter deshalb ein Start/Stop-Erkennen geschaltet. Er überprüft, ob die Hand kurz an einer Stelle verharret oder nicht. Dieses Ereignis dient als Auslöser für die Erkennung und legt

Start- und Endpunkt der Trajektorie fest. Eine kontinuierliche Klassifikation wie in [10] wurde wegen höherer Verarbeitungsgeschwindigkeit und Stabilität verworfen. Im Umgang mit dem System stellt diese Forderung keine nennenswerte Einschränkung dar.

3 Gestenerkennung

Das Training der Hidden-Markov-Modelle erfolgt mit dem Baum-Welch Algorithmus als etablierter Methode. Dabei kamen als Modelle Links-Rechts-Modelle mit einer Sprunbegrenzung von $\Delta = 2$ zum Einsatz. Zum Test wurden fünf Referenzgesten ausgewählt und mit je 10 Beispielen trainiert (siehe Abb. 5).

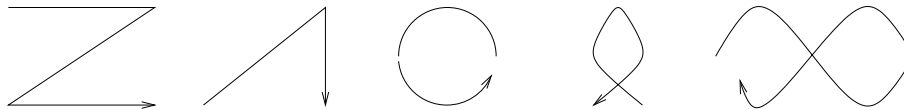


Abbildung5. Referenzgesten.

Mit den in den vorangegangenen Abschnitten eingeführten Techniken lässt sich nun das Erkennungssystem als Ganzes beschreiben. Dabei sieht der Gesamtprozess aus, wie in Abbildung 6 dargestellt.

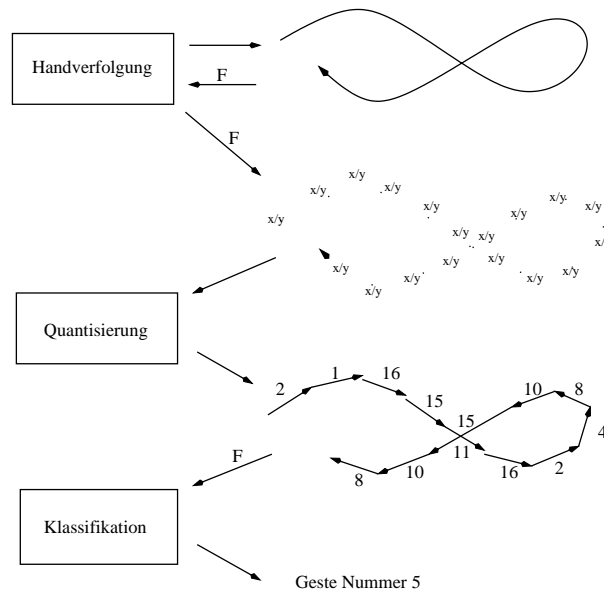


Abbildung6. Der Gesten-Erkennungsprozess mit seinen Einzelschritten im Überblick. Der Einsatz von Filtern ist durch ein „F“ gekennzeichnet.

Bei einer aufgezeichneten Trajektorie kann nun die Klassifikation nach vier Varianten erfolgen:
Maximale Wahrscheinlichkeit (MW): Beim Test auf maximale Wahrscheinlichkeit wird die gefilterte Trajektorie jedem Referenzmodell zum Test vorgelegt und dasjenige ausgewählt, das

hierauf den höchsten Wahrscheinlichkeitswert emittiert hat. Zur Ermittlung dieses Wertes wird der Viterbi-Algorithmus verwendet [17].

Schwellwertmodell (SM): Vor der Klassifikation wird durch Fusionierung sämtlicher Referenzmodelle ein Schwellwertmodell konstruiert. Die Trajektorie wird nun diesem und allen Referenzmodellen vorgelegt. Ist der höchste Wahrscheinlichkeitswert höher als derjenige des Schwellwertmodells, wird die Trajektorie akzeptiert, andernfalls als bedeutungslos verworfen. Dieses Verfahren wurde erstmals verwendet in [10].

Hierarchische Klassifikation mit maximaler Wahrscheinlichkeit (HM): Die Gesten werden hinsichtlich ihrer Länge in verschiedene Komplexitätsklassen unterteilt. Die Klassifikation nach maximaler Wahrscheinlichkeit erfolgt nur innerhalb der Klasse.

Hierarchische Klassifikation mit Schwellwertmodell (HS): Hier erfolgt ebenso eine Klasseneinteilung. Jeder Klasse wird ein Schwellwertmodell zugeordnet, dessen Wahrscheinlichkeitsausgabe auf die ausgeführte Geste hin nicht unterschritten werden darf.

Neu an diesem Ansatz ist die Ergänzung des Schwellwertmodells durch hierarchische Klassen, welche die Klassifikation erleichtern. Dies ist in Abb. 7 noch einmal verdeutlicht.

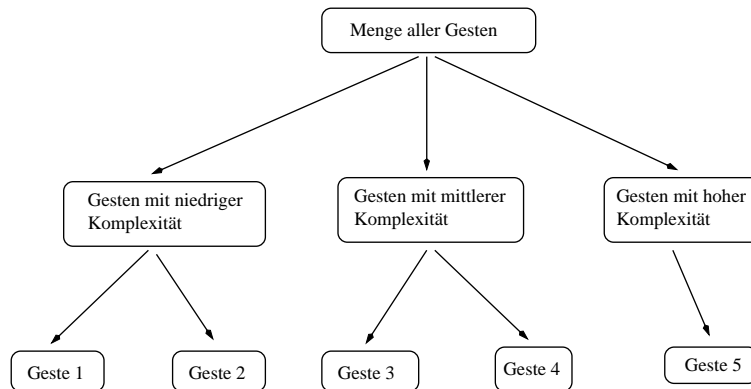


Abbildung 7. Klassifizierung der Referenzgesten aufgrund ihrer Komplexität.

4 Gütebewertung

Tests mit über 200 ausgeführten Vorführungen erbrachten die in Tabelle 1 gezeigten Resultate.

Geste	MW	SM	HM	HS
1	100%	77%	100%	68%
2	93%	79%	100%	83%
3	100%	76%	80%	84%
4	100%	85%	88%	73%
5	100%	100%	82%	73%
Gesamt	99%	83%	89%	76%

Tabelle 1. Testresultate mit den jeweiligen Klassifikationsvarianten.

Hierzu ist anzumerken: Die Klassifikation nach der maximalen Wahrscheinlichkeit zeigt zwar die besten Resultate, ist aber wegen der Unfähigkeit, bedeutungslose Trajektorien zu filtern, nicht

praxistauglich. Das Schwellwertmodell stellt hierfür so hohe Anforderungen, dass diesem Anspruch Genüge geleistet wird. Die Klassifikation mit Hilfe eines hierarchischen Ansatzes drittelt nahezu den Rechenaufwand bei drei Klassen. Wir haben uns daher dafür entschieden, diese Methode weiter zu entwickeln.

5 Zusammenfassung und Ausblick

Im vorliegenden Artikel wurden Ansätze vorgestellt, die die Verfolgung von Handbewegungen stabil bewerkstelligen und die Klassifikation von dynamischen Gesten ermöglichen. Die Klassifikation selbst erfolgt auf Basis von Hidden-Markov Modellen. Zum Ausschluss bedeutungsloser Bewegungen wurde der Klassifikator um das Schwellwertmodell erweitert. Ein neuer Ansatz besteht in der Aufteilung der Gesten in Komplexitätsklassen. Dies trägt zu einer schnelleren Erkennung von Handzeichen bei. Einen ebenso einfachen wie effektiven Ansatz zur Auslösung und Beendigung der Demonstration einer Geste stellt dazu außerdem der implementierte Start/Stop-Filter dar.

Erweiterungen dieses Ansatzes werden weniger die Klassifikation als vielmehr die Bildverarbeitung betreffen. Zunächst ist dabei an die Verfolgung von dreidimensionalen Trajektorien zu denken. Diese ist durch Sakkadenbewegungen zu ergänzen. Die Klassifikation der bisher schon ausgewählten Gesten kann dann innerhalb der Hyperebene mit der grössten Varianz geschehen.

Förderung: Die ausgeführten Forschungsarbeiten wurden im Rahmen des BMBF-Leitprogramms „Mensch-Technik-Interaktion in der Wissensgesellschaft“ im Projekt „Intelligente anthropomorphe Assistenzsysteme - Morpha“ durchgeführt.

Literatur

1. A. Arsenio and J. Santos-Victor. Robust visual tracking by an active observer. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 1342–1347, 1997.
2. A. Blake and M. Isard. *Active Contours*. Springer, 1998.
3. G. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, März 1973.
4. D. Gavrilu and L. Davis. Towards 3d model-based tracking and recognition of human movement: a multi-view approach. In *International Workshop on Face and Gesture Recognition, Zürich*, 1995.
5. T. Heap and F. Samaria. Real-time hand tracking and gesture recognition using smart snakes. Technical report, Olivetti Research Limited, 20. Juni 1995.
6. H. Kestler, M. Borst, and H. Neumann. Einfache Handgestikerkennung mit einem zweistufigen Nearest-Neighbour Klassifikator. Technical report, Universität Ulm, SFB 527, 96/6, 1996.
7. J. Kim, W. Jang, and Z. Bien. A dynamic gesture recognition system for the korean sign language (KSL). *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, 26(2):354–359, April 1996.
8. M. Kohler. *Übersicht Handgestenerkennung*. <http://ls7-www.cs.uni-dortmund.de/research/gesture/>, 2000.
9. C. Lee and Y. Xu. Online, interactive learning of gestures for human/robot interfaces, Minneapolis, Minnesota. In *Proceedings of the IEEE International Conference on Robotics and Automation*, April 1996.
10. H. Lee and J. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):961–973, Oktober 1999.
11. D. Newtson and et al. The objective basis of behaviour units. *Journal of Personality and Social Psychology*, 35, 1977.
12. P. Nordlund and T. Uhlin. Closing the loop: Detection and pursuit of a moving object with sensors onboard a moving robot. *Image Vision and Computing*, 14(4):265–275, 1996.
13. A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):107–119, Januar 2000.
14. R. Plamondon and S. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84, Januar 2000.

15. L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, Februar 1989.
16. J. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *ECCV*, pages 35–46, 1994.
17. M. S. Ryan and G. R. Nudd. The viterbi algorithm. Warwick Research Report RR 238, Department of Computer Science, University of Warwick, Coventry, Februar 1993.
18. N. Shimada and Y. Shirai. 3d hand pose estimation and shape model refinement from a monocular image sequence. In *Proceedings of the VSMM, Gifu*, pages 423–428, 1996.
19. H. Sidenbladh, D. Kragic, and H. Christensen. A person following behaviour for a mobile robot. In *Proceedings of the IEEE International Conference on Robotics and Automation, Detroit, MI, USA*, pages 670–675, April 1999.
20. T. Starner. Real-time american sign language recognition from video using hidden Markov models. In *Proceedings of the IEEE International Symposium on Computer Vision*, pages 265–270, 1995.
21. J. Triesch and Chr. von der Malsburg. Robotic gesture recognition. In *Proceedings of the Bielefeld Gesture Workshop*. Springer, 17.-19. September 1997.
22. M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *CVPR*, pages 664–665, 1991.
23. J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. In *Proceedings of ACCV, Hong Kong*, volume 2, pages 687–694, 1998.