DIRK GEHRIG, HILDEGARD KÜHNE & TANJA SCHULTZ

Erkennung von menschlichen Bewegungen mit Hidden Markov Modellen

Einleitung

Ein wichtiges und ständig wachsendes Forschungsgebiet innerhalb der Robotik sind Humanoide Roboter. Diese sollen ein menschenähnliches Aussehen haben und sich menschlich verhalten. Dazu ist es notwendig, dass sie menschliche Tätigkeiten und menschliches Verhalten erkennen können. Dabei spielt die Erkennung menschlicher Bewegungen eine wesentliche Rolle.

Ein Standardverfahren zur Modellierung von menschlichen Bewegungen für die Bewegungserkennung sind Hidden Markov Modelle (HMMs) (Rabiner, 1989). Sie eignen sich sehr gut für die Modellierung und Erkennung von Zeitreihen, wie z.B. Trajektorien menschlicher Bewegungen, oder deren Eigenschaften (Fischer, 2009). Dabei sind sie flexibel was die variierende Dauer von Bewegungsausführungen angeht. Dies ermöglicht die Erkennung menschlicher Bewegungen ohne vorherige Normierung der Bewegungsdauer. Durch die Kombination von HMMs einzelner Bewegungsprimitiven können sowohl komplexe Bewegungen als auch Bewegungssequenzen, die sich aus einer Folge von Bewegungsprimitiven zusammensetzen, erkannt werden.

Hidden Markov Modelle

Die Grundidee bei der Modellierung einer Bewegungsprimitive durch ein HMM besteht darin, dass die Bewegungsprimitive in einzelne Zustände unterteilt werden kann, die zeitlich nacheinander folgen. Dabei kann der Zustand nicht direkt beobachtet werden, sondern nur die vom Zustand emittierten Bewegungsmerkmale. Eine solche Unterteilung der Bewegung in Zustände könnte für eine Handbewegung von einem Punkt im Raum zu einem anderen, z.B. um ein Objekt von A nach B zu stellen, aus folgenden Teilen bestehen. Der erste Zustand könnte die Beschleunigungsphase der Bewegung sein, der zweite eine nahezu gleichförmige Bewegung als Mittelteil der Bewegung und am Ende ein dritter Zustand, der die Abbremsphase modelliert. Bei der Modellierung der Bewegungserkennung werden die Zustände allerdings in der Regel nicht explizit modelliert. Übergangswahrscheinlichkeiten geben die Wahrscheinlichkeit eines Zustandswechsels von einem Zeitpunkt zum nächsten an. Die optimale Anzahl an Zuständen, d.h. die Länge der Bewegungsprimitive, wird z.B. anhand der Erkennungsergebnisse einer Kreuzvalidierung bestimmt. Jeder der Zustände modelliert (emittiert) die Merkmale, z.B. Gelenkstellungen oder deren Winkelgeschwindigkeiten, die für den entsprechenden Teil der Bewegung typisch sind. Die Beobachtungswahrscheinlichkeiten, d.h. die statistische Verteilung der Merkmale innerhalb eines HMM-Zustandes, wird dabei

häufig mit Gaußmischverteilungen (GMM) modelliert. Die Anzahl der Gaußglocken für die Mischverteilung wird für die zur Verfügung stehende Datenmenge und die Bewegungskategorie optimiert. Diese werden mit Hilfe von EM-Training anhand von Trainingsdaten iterativ so geschätzt, dass sie die Trainingsdaten optimal modellieren. Um eine Erkennung einer Bewegungstrajektorie durchzuführen, wird eine Dekodierung durchgeführt. Die Folge an HMMs, welche die Bewegungstrajektorie am besten beschreiben, wird gesucht und als die erkannte Folge von Bewegungsprimitiven ausgegeben.

Scenario

Im Sonderforschungsbereich 588 „Humanoide Roboter“, werden HMMs verwendet, um menschliche Bewegungen in einer Küchenumgebung zu erkennen. Auf dem humanoiden Roboter müssen diese Bewegungen anhand von unpräzisen und oft verrauschten Videodaten erkannt werden. Für die Entwicklung des komplexen Erkennungssystems wurden außerdem Experimente auf der Basis eines markerbasierten IR-Trackings durchgeführt. Für das Erkennungssystem wurden die Oberkörperbewegungen von 100 Bewegungssequenzen aufgenommen (Gehrig, 2009).

Dazu stand ein Proband vor einem Tisch, auf dem typische Küchenutensilien wie ein Messer, eine Schüssel usw. an vordefinierten Positionen lagen. Der Proband wurde aufgefordert typische Küchenbewegungssequenzen durchzuführen. Teil jeder Bewegungssequenz war das Positionieren der Objekte vor dem Probanden, das Verwenden der Objekte und das Zurücklegen an die Ausgangsposition. Insgesamt wurden die 5 folgenden Bewegungssequenzen jeweils 20 Mal aufgenommen: Einschenken, Apfel Reiben, Rühren, Apfel Schneiden und Stampfen. Die Bewegungssequenzen bestehen aus insgesamt 24 verschiedenen Bewegungsprimitiven.

Datenerfassung

Zur Datenerfassung wurden zwei verschiedene Systeme parallel eingesetzt. Zum einen wurde die Bewegung markerbasiert mittels eines Vicon -Trackingsystems erfasst. Die aufgenommenen Markerpositionen wurden durch Berechnung der inversen Kinematik auf ein entsprechend angepasstes biomechanisches Menschmodell

abgebildet (Simonidis, 2008). Das Ergebnis dieser präzisen Bewegungserfassung ist die Konfiguration der im Modell abgebildeten Gelenkwinkel über den Aufnahmezeitraum. Aus den Gelenkwinkeln wurde durch Subtraktion aufeinanderfolgender Gelenkstellungen Winkelgeschwindigkeiten berechnet.

Um Bewegungsdaten mit Hilfe von Videobildern zu erfassen, wurden die Bewegungsrichtungen für jeden Frame berechnet. Dazu wurden zunächst sich bewegende, markante Punkte im Bild gesucht und über die Zeit verfolgt. Für die Detektion der Punkte wurde der Ansatz von Shi und Tomasi (Shi, 1994) verwendet. Markante Punkte entsprechen dabei Bildausschnitten mit starken Gradienten in zwei verschieden Richtungen wie z.B. Ecken. Für die Verfolgung wurde der Ansatz von Lucas und Kanade (Lucas, 1981) verwendet, wobei die Bewegungsrichtung jedes sich bewegenden Punktes von Bild I_t nach I_{t+1} geschätzt wurde. Über die Bewegungsrichtungen aller Punkte in einem Bild wurde ein 30-dimensionales Histogramm gebildet. Um eine Zusätzliche Aussage über die Stärke der Bewegungsrichtung treffen zu können, werden die einzelnen Histogrammeinträge mit der Länge des Bewegungsvektors gewichtet. Für den hier beschrieben Ansatz wurde ein feingranulares, globales Histogramm gewählt, da dieses auch kleinere, entsprechend feine Bewegung angemessen abbildet. Für Aufnahmen unter anderen Umweltbedingungen oder auch Aufnahmen von ausladenderen Bewegungen ist auch eine örtliche Unterteilung des Aufnahmeraums und eine weniger feine Histogrammaufteilung möglich.

Für eine gute Vergleichbarkeit wurden beide Datensätze mit 20 Hz gesampled. Bei beiden Datensätzen wurden Mittelwert und Varianz je Sequenz normiert. Um mit den Daten der Bewegungssequenzen HMMs initialisieren zu können, ist es notwendig die Trajektorien in die einzelnen Bewegungsprimitive zu zerlegen. Für die Experimente wurde hier eine manuelle Zerlegung durchgeführt.

Experimente und Ergebnisse

Die Initialisierung der HMMs wurde durchgeführt, indem alle Datensätze zu jedem Bewegungsprimitiv gleichmäßig in die Anzahl an Zuständen des jeweiligen HMM aufgeteilt wurden. Mit Hilfe des Neural Gas Algorithmus wurden die Gaußglocken der einzelnen Zustände initialisiert. Anschließend wurden die HMMs auf den unzerlegten Trajektorien von Merkmalsvektoren mit Viterbi-basiertem EM-Training trainiert. Die Suche der besten Folge von Bewegungsprimitiven für eine Trajektorie der Testdaten wurde mit zeitsynchroner Strahlsuche durchgeführt. Dabei waren alle Übergänge von einer Bewegungsprimitive zu einer anderen jeweils gleich Wahrscheinlich. Die Experimente wurden in 10-facher Kreuzvalidierung durchgeführt. Mit Hilfe der Experimente wurde untersucht, ob eine auf den Typ der Bewegungsprimitive angepasste Anzahl an Zuständen je Primitive und Gaußglocken je Zustand eine Verbesserung der Erkennungsergebnisse liefert. Es wurden Experimente

durchgeführt, bei denen die Anzahl der Gaußglocken pro Zustand sowie die Anzahl der Zustände pro Bewegungsprimitive auf Entwicklungsdaten identisch für alle Bewegungsprimitive und individuell für jede Bewegungsprimitive bestimmt wurden. Bei ersterem wurde die Anzahl der Gaußglocken für alle Bewegungsprimitive identisch in Zweierpotenzen zwischen 1 und 64 optimiert. Um eine an die Bewegungstypen angepasste Anzahl der Gaußglocken zu erreichen, wurde statt die Anzahl der Gaußglocken manuell festzulegen, deren Anzahl während der Initialisierung mit Hilfe des Split & Merge Algorithmus festgelegt (Ueda, 2000).

Bei der Anzahl an Zuständen pro Bewegungsprimitiv wurde die Länge für alle identisch zwischen 1 und 12 Zuständen optimiert. Um eine an den Typ der Bewegungsprimitive angepasst Länge zu erreichen wurde die durchschnittliche Länge der Trainingsdaten jedes Typs von Bewegungsprimitive mit verschiedenen Faktoren multipliziert um die Anzahl der Zustände zu berechnen. Für eine Übersicht über die Längen der Primitive siehe Abb. 1.

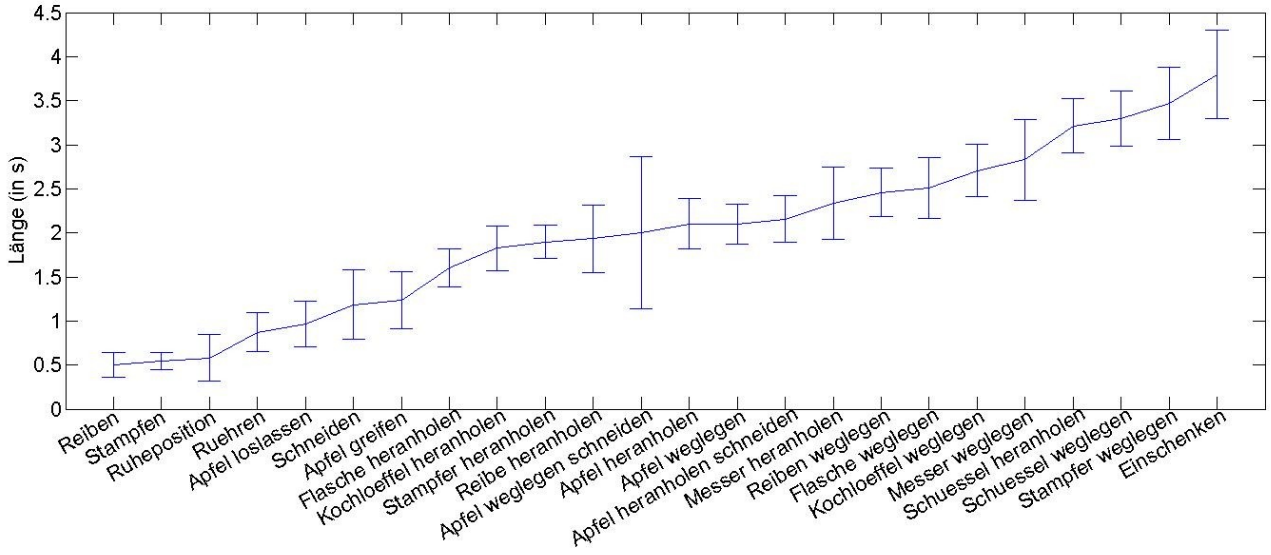

Abb.1: Durchschnittliche Länge der manuell segmentierten Bewegungsprimitive.

Die besten Erkennungsraten für die verschiedenen Kombinationen aus identischen und angepassten Parametern sind wie folgt:

Zustände	**Gaußglocken**	**Beste Primitivfehlerrate**	
		Markerbasiert	**Videobasiert**
Identisch	Identisch	4.2 %	6.9 %
Identisch	Angepasst	5.3 %	---
Angepasst	Identisch	5.1 %	---
Angepasst	Angepasst	6.3 %	---

Tab. 1: Ergebnisse der Erkennung mit identischer und angepasster Anzahl an Zuständen und Gaußglocken.

Das beste markerbasierte Erkennungssystem lieferte lediglich einen Fehler von 4,2% bei der Erkennung der Bewegungsprimitiven. Dieses Ergebnis wurde mit

identischer Anzahl an Zuständen und Anzahl an Gaußglocken für alle Bewegungsprimitive erreicht. Dieses Experiment wurde deshalb auch mit den unpräzisen Video-Daten des Roboterkopfes durchgeführt. Die beste Fehlerrate lag hierbei bei 6,9%.

Diskussion

Sowohl mit identischer als auch mit angepasster Anzahl der Zustände pro Typ von Bewegungsprimitiv als auch Anzahl der Gaußglocken pro Zustand wurden sehr gute Erkennungsergebnisse erreicht. Die beste Erkennung wurde bei identischer Anzahl an Zuständen und Gaußglocken für alle Bewegungsprimitive erreicht und liegt auf den 24 zu erkennenden Bewegungsprimitiven bei 4.2%. Auch mit Video wurden sehr gute Ergebnisse in der Erkennung von menschlichen Bewegungsprimitiven erzielt. Die Ergebnisse waren dabei etwas schlechter, als die Ergebnisse der Vicon-basierten Erkennung. Im nächsten Schritt müssen die Erkennungsergebnisse nochmals genauer untersucht werden, um die ausgebliebene Verbesserung durch angepasste Parameteranzahl erklären zu können.

Danksagung

Diese Arbeit wird von der DFG im Rahmen des Teilprojektes M3 „Bewegungs- und Handlungsmodelle“ des Sonderforschungsbereiches 588 „Humanoide Roboter – Lernende und kooperierende multimodale Roboter“ gefördert.

Literatur

Rabiner, L. (1989). A tutorial on HMM and selected applications in speech recognition. Proceedings of the IEEE.

Gehrig, D., Kühne, H., Wörner, A. & Schultz, T. (2009). HMM-based Human Motion Recognition with Optical Flow Data. Humanoids 2009.

Fischer, A., Stein, T., Gehrig, D., Schultz, T. & Schwameder, H. (2009). Training und Erkennung mit Hidden Markov Modellen bei unterschiedlichen Geh-/Laufgeschwindwindigkeiten. Biomeachanik, DVS 2009.

Shi, J. & Tomasi, C (1994). Good features to track. Proceedings of the Conference on Computer Vision and Pattern Recognition, 593–600.

Simonidis, C. & Seemann, W. (2008). MkdTools - human models with Matlab. In Wassink, R. (Ed.), The 10th International symposium on 3D Analysis of Human Movement - Fusion Works

Lucas, B.D. & Kanade T. (1981). An iterative image registration technique with an application to stereo vision.

Ueda, N., Nakano, R., Ghahramani, Z. & Hinton, G.E. (2000). Split and Merge EM Algorithm for Improving Gaussian Mixture Density Estimation. Journal of VLSI Signal Procesing 26, 133-140.