Visuelle Gestenerkennung zur Interaktion zwischen Mensch und Roboter

Florian A. Bertsch

Lehr- und Forschungsgebiet Kognitive Robotik
Institut für Informatik
Humboldt-Universität zu Berlin
Unter den Linden 6
D-10099 Berlin

Das visuelle Erkennen und Lernen von Gesten bildet die Grundlage für eine gestenbasierte Interaktion mit mobilen Robotern. Die zugrunde liegende Bildverarbeitung muss dabei auf der oft langsamen Hardware eines mobilen Roboters in Echtzeit berechnet werden. Das vorgestellte Verfahren modelliert dazu die menschliche Hautfarbe unter variierenden Beleuchtungsbedingungen, um hautfarbene Regionen als mögliche Positionen von Kopf und Händen einer gestikulierenden Person zu lokalisieren.

1 Einführung

Im Fall der Interaktion mit mobilen Robotern kann die gestenbasierte Kommunikation sehr vorteilhaft sein. Insbesondere wenn humanoide Roboter in zunehmendem Maße im alltäglichen menschlichen Umfeld zum Einsatz kommen, ist eine möglichst natürliche Interaktion nach dem Vorbild der zwischenmenschlichen Interaktion wünschenswert. Dieser Artikel beschäftigt sich mit dem Aspekt der Farbbildverarbeitung bei einem Verfahren zur Gestenerkennung für den Einsatz zur Interaktion mit mobilen Robotern.

2 Übersicht über gängige Verfahren

Für das Erkennen von Gesten in Videosequenzen wurde bereits eine Vielzahl von Verfahren entwickelt und vorgestellt, die sich spezielle Eigenschaften des Problems zu Nutze machen. So wurden spezifische Aspekte wie der menschliche Körperbau, die menschliche Hautfarbe oder das typische Aussehen von Gesichtern als Vorwissen genutzt, um effizient geeignete Merkmale aus den Videosequenzen zu bestimmen. Gemeinsames Ziel dieser Verfahren ist es, Merkmale zu bestimmen, die die Position und Haltung der Person beschreiben, die in der Videosequenz dargestellt ist. Diese Information kann dann als Grundlage für die Gestenerkennung genutzt werden, da die Geste einer Person durch die Änderung ihrer Haltung über die Zeit beschrieben wird.
An dieser Stelle soll ein Einblick in die Vielfalt der Verfahren zum Ermitteln der Haltung und Bewegung einer Person in einer Videosequenz gegeben werden. Dabei werden nur einige grundsätzliche Unterschiede skizziert, um das in dieser Arbeit entwickelte Verfahren in diesem Kontext besprechen zu können. Für einen weiterführenden Überblick über den Stand der Forschung und die grundlegenden Unterschiede der Verfahren sei hier auf [9] und [7] verwiesen.

2.1 Vereinfachende Annahmen

Mithilfe verschiedener Annahmen kann die Problemstellung vereinfacht werden. Diese lassen sich in Annahmen über Bewegung und Annahmen über Erscheinung unterteilen. Eine Übersicht über typische Annahmen zur Vereinfachung des Problems bietet Tabelle 2.1. Durch die Annahmen wird das Aussehen der Person im Bild oder ihre möglichen Haltungen vereinfacht, so dass das Lokalisieren der Person bzw. das Schätzen ihrer Haltung leichter wird. Die Annahmen, die zur Vereinfachung des in diesem Artikel vorgestellten Verfahrens gemacht werden, sind in der Tabelle durch * markiert.

Annahmen über Bewegung	Annahmen über Erscheinung
Keine Bewegung der Kamera	Konstante Beleuchtung
Nur eine Person im Bild*	Statischer Hintergrund
Personen Richtung Kamera orientiert*	Markierungen an der Person
Bewegung parallel zu Kamera-Ebene*	Spezielle Kleidung*

Tabelle 2.1: Vereinfachende Annahmen bei der Schätzung von Personen in Videosequenzen

2.2 Modellbasierte Verfahren

Die meisten Verfahren basieren auf einem kinematischen Modell des menschlichen Körpers (siehe Abbildung 2.1 links). Dieses beschreibt den menschlichen Körper als eine aus Geraden-Segmenten bestehende Struktur. Die Geraden-Segmente sind mit Gelenken verbunden, die sich hinsichtlich der möglichen Bewegungsrichtungen unterscheiden [9]. Neben den kinematischen Modellen gibt es noch die formbasierten Modelle (siehe Abbildung 2.1 rechts). Bei diesen werden die Segmente des Modells als volumetrische Formen, wie beispielsweise Kugeln oder Zylinder, modelliert [4].

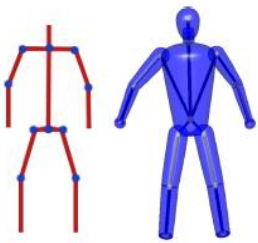

Abbildung 2.1: Kinematisches (links) und volumetrisches (rechts) Modell des menschlichen Körpers. Quelle: Volumetrisches Modell aus [4]

Um die aktuelle Konfiguration des Modells aus einer Videosequenz zu schätzen, werden zunächst geeignete Merkmale durch eine Vorverarbeitung mit Hilfe von Verfahren der Bildverarbeitung gewonnen. Typische Merkmale, die auf diesem Wege gewonnen werden, sind Silhouetten, Kanten, Bereiche ähnlicher Farbe bzw. Textur sowie Bewegung bzw. optischer Fluss [9].

Die Vorgehensweisen, um die aktuelle Konfiguration des Modells zu schätzen, lassen sich in top-down- und bottom-up-Verfahren unterscheiden. Bei den top-down-Verfahren wird ein formbasiertes Modell in das Bild projiziert und mit der Beobachtung im Bild verglichen. Bei den bottom-up-Verfahren werden zunächst einzelne Körperteile im Bild lokalisiert. Aus der Lage der gefundenen Körperteile wird dann die Haltung einer Person zusammengesetzt.

2.3 Verfahren ohne Modell

Eine wichtige Klasse von Verfahren, die kein explizites Modell des menschlichen Körpers zugrunde legen, sind Verfahren, die keine Schätzung der gesamten Körperhaltung vornehmen. Insbesondere bei vielen Verfahren zur gestenbasierten Interaktion in Echtzeit beschränkt man sich auf das Lokalisieren und zeitliche Verfolgen einzelner Körperteile. In vielen Verfahren wird beispielsweise ausschließlich eine Hand in der Videosequenz verfolgt, um Zeichen, die mit dieser Hand ausgeführt werden, zu erkennen [3] [13].

3 Anforderungen an das Verfahren zur Merkmalsextraktion

Nun soll das in diesem Artikel vorgestellte Verfahren zur Extraktion gestenrelevanter Merkmale aus Videosequenzen vorgestellt werden. Als Designkriterien für den Entwurf des Verfahrens wurden die Anforderungen für die gewünschte Anwendung im Bereich der gestenbasierten Interaktion zwischen Mensch und humanoidem Roboter herangezogen:

1. **Echtzeitfähigkeit** bei beschränkter Hardware eines mobilen Roboters
2. Erkennung von **Gesten aus Videosequenzen** ohne zusätzliche Hilfsmittel

Die beiden ersten Punkte beeinflussen die Auswahl des Verfahrens zur Merkmalsextraktion. Aufgrund der gewünschten Echtzeitfähigkeit muss das Verfahren auf zeitaufwändige Berechnungen verzichten. Typischerweise sind Verfahren, die nach dem top-down Ansatz funktionieren, aufwändig zu berechnen. Daher wurde ein Verfahren gewählt, das darauf beruht, einzelne Körperteile in der Videosequenz zu lokalisieren und ihre relative Lage auszuwerten.
Da ein solches Verfahren jedoch besonders große Schwierigkeiten mit stark variierenden Perspektiven hat, wird die Aufgabe im Weiteren auf den Fall einer Frontalansicht des Menschen beschränkt. Zusätzlich wird angenommen, dass die Gestikulation im Wesentlichen parallel zur Kamera-Ebene stattfindet. Dies scheint für eine gestenbasierte Interaktion keine sehr problematische Einschränkung, da es natürlich ist, sich dem Gesprächspartner zuzuwenden. Trotzdem beschränkt diese Voraussetzung erheblich die Menge der Gesten, die mit dem Verfahren erkannt werden können. Zudem wird davon ausgegangen, dass der Benutzer langärmlige Oberbekleidung mit hohem Kragen und keine kurzen Hosen trägt. Diese Annahme ist notwendig, da das schnelle Erkennen von Kopf und Händen im weiteren Verlauf dieser Arbeit auf dem Erkennen von Farbregionen beruht, die hautfarben sind. Daher sollen zusätzliche hautfarbene Regionen, wie nackte Arme und Beine, vermieden werden.

4 Lokalisierung von Personen

Zusätzlich wird zur Vereinfachung angenommen, dass sich nur eine Person im Blickfeld des Roboters befindet. Dadurch wird vermieden, dass man gefundene Körperteile verschiedenen Personen zuordnen muss. Es wird davon ausgegangen, dass die Person, die mit dem Roboter interagieren möchte, sich diesem frontal zuwendet. Bei einer solchen Ausrichtung ist das Gesicht der Person gut zu erkennen und bietet sich als Merkmal zum Erkennen der Person an, die sich im Blickfeld des Roboters befinden.

4.1 Lokalisieren von Gesichtern

Einen Überblick über gängige Verfahren zum Lokalisieren von Gesichtern findet man in [12]. Ein Verfahren, das sich durch seine hohe Geschwindigkeit auszeichnet und sich somit besonders gut für den Einsatz bei den beschränkten Hardware-Ressourcen eines mobilen Roboters eignet, ist der Viola-Jones Detector [11]. Dieses Verfahren basiert auf der Extraktion einfacher Merkmale aus einem Graustufenbild. Zur Konstruktion dieser Merkmale werden die Werte in rechteckigen Bereichen des Bildes aufsummiert. Die Merkmale werden dann durch gewichtete Differenzen einiger Rechtecksummen gebildet. So führen Kontrastunterschiede in der Form dieser rechteckigen Merkmale zu hohen Differenz- und damit Merkmals-Werten.

Abbildung 4.1: Beispiele wichtiger Merkmale zum Lokalisieren von Gesichtern mittels des *Viola-Jones Detector*

In Abbildung 4.1 sind zwei Merkmale dargestellt, die typischerweise von einem Viola-Jones Detector verwendet werden. Man kann hier beispielhaft sehen, dass ein Merkmal die dunkle Augenpartie gegenüber dem helleren Bereich darunter beschreibt (links), ein anderes die helle Nase im Vergleich zu den dunklen Augen daneben (rechts). Der Viola-Jones Detector erreicht durch das Bestimmen und Auswerten vieler tausend solcher Merkmale eine hohe Trefferquote. Dabei können die Merkmale aufgrund ihrer rechteckigen Grundstruktur effizient mit Hilfe von *Integral Images* berechnet werden.
Um Gesichter an verschiedenen Positionen in einem Bild zu finden, wird ein Suchfenster über das gesamte Bild geschoben. Der Inhalt des Fensters wird mit dem Viola-Jones Detector auf das Vorhandensein eines Gesichtes überprüft. Ist das gesamte Bild abgesucht, wird der Suchbereich vergrößert und erneut über das Bild geschoben, um Gesichter verschiedener Größen zu lokalisieren.

4.2 Berücksichtigung der Hautfarbe bei der Lokalisierung

Zum Lokalisieren von Personen kann zusätzlich die menschliche Hautfarbe als Merkmal herangezogen werden. Während die Farbe der Kleidung stark variiert, ist die Färbung der menschlichen Haut ein typisches Merkmal. Auch für Menschen mit verschiedener Hautfarbe ergeben sich große farbliche Ähnlichkeiten gegenüber den sonstigen Farbvariationen im menschlichen Lebensumfeld. Während große Teile der Haut häufig mit Kleidung bedeckt sind, sind Gesicht und Hände typischerweise unbekleidet. Somit kann die Farbe ein leistungsfähiges Merkmal zum Lokalisieren von Gesichtern sein.

4.2.1 Wahl des Farbmodells

Welches Farbmodell sich am besten für die Kodierung und anschließende Klassifikation der menschliche Hautfarbe eignet, wurde in vielen Arbeiten untersucht [10]. Es wird deutlich, dass sich die Frage nach dem besten Farbmodell nicht eindeutig beantworten lässt und auch von der Kombination mit einem Klassifikator abhängt. In vielen Ansätzen wird jedoch davon ausgegangen, dass die menschliche Hautfarbe in Bildern besonders stark in der Helligkeit und weniger im Farbton variiert. Wird also ein Farbmodell gewählt, das die Helligkeit getrennt von zwei Farbkanälen modelliert, so bildet die menschliche Hautfarbe einen vergleichsweise kompakten Bereich in den beiden Farbkanälen. Man betrachtet die Farbwerte also nur noch im zweidimensionalen Farbraum, der den Farbton beschreibt, und ignoriert die Farbhelligkeit. In [5] und [6] wir das Normalisierte-RGB-Modell als Farbraum gewählt, da die menschliche Hautfarbe verschiedener Ethnien bei unterschiedlichen Beleuchtungsbedingungen dort einen kompakten und einfach zu beschreibenden Bereich bilden. Die Farben aus dem gängigen RGB-Modell werden wie folgt in das Normalisierten-RGB-Modell überführt:

$$r = R / (R + G + B) \qquad g = G / (R + G + B) \qquad b = B / (R + G + B)$$

Da die Summe der drei Komponenten mit $r + g + b = 1$ bekannt ist, kann ohne einen Informationsverlust auf eine Komponente verzichtet werden und man erhält einen zweidimensionalen Farbraum. Dazu werden wie in [6] die Komponenten r und b gewählt.

4.2.2 Bereich der Hautfarbe im Farbraum

Abbildung 4.2 zeigt die Lage der Farben eines hautfarbenen Bildbereiches im Normalisierte-RGB-Farbraum. Dazu wurde ein ellipsenförmiger Bereich des Bildes, der das Gesicht enthält, ausgewählt (links). Dieser Bereich führt im reduzierten zweidimensionalen Farbraum der r- und b-Komponente zu einem kompakten Histogramm h (Mitte). Alle Farben, die im ellipsenförmigen Bereich enthalten sind, bedecken einen kleinen und kompakten Bereich des gesamten Farbraumes, der rechts im Bild dargestellt ist.

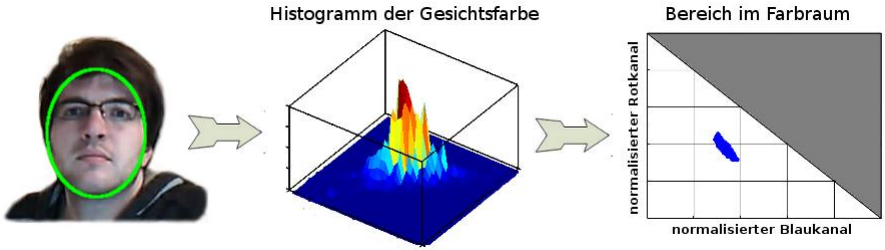

Abbildung 4.2: Histogramm der Farbwerte eines Gesichtes und ihre Lage im Normalisierten-RGB-Farbraum

Der Bereich des Normalisierten-RGB-Farbraumes, der alle möglichen Hautfarben verschiedener Personen unter unterschiedlichen Beleuchtungsbedingungen umfasst, wird skin locus genannt [10]. Er lässt sich experimentell bestimmen, indem hautfarbene Bereiche in Beispielbildern unter variierenden Beleuchtungsbedingungen manuell selektiert werden. Obwohl der skin locus bereits in vielen Untersuchungen für variierende Beleuchtungsbedingungen und Personen ermittelt wurde, kann nur eingeschränkt auf diese Ergebnisse zurückgegriffen werden, da das Ergebnis von der benutzten Kamera abhängt [6].

4.2.3 Umgebung des white point ausschließen

Typischerweise liegt auch die Farbe Weiß (r = 0.33 und b = 0.33) im Bereich des skin locus. Der Grund dafür sind helle Lichtreflexionen in hautfarbenen Bereichen. Dass Weiß im skin locus enthalten ist, führt dazu, dass auch helle bzw. gräuliche Bereiche als möglicherweise hautfarben angenommen werden. Da diese Farben aber auch häufig in der Umgebung auftreten, werden sie oft explizit aus dem skin locus ausgeschlossen [5]. Dies erfolgt, indem ein kreisförmiger Bereich um den white point aus dem skin locus entfernt wird.

4.2.4 Verwendung zum Lokalisieren von Gesichtern

Bei der Lokalisierung von Gesichtern kann der skin locus genutzt werden, um zu überprüfen, ob ein Bereich im Bild genügend potentiell hautfarbene Pixel enthält, um als Gesicht in Frage zu kommen. Als eigenständiges Merkmal zur Erkennung von Gesichtern ist die Farbe nicht gut geeignet, da der skin locus, aufgrund der starken farblichen Variation durch wechselnde Beleuchtungsbedingungen, zu umfangreich ist. In alltäglicher Umgebung funktioniert die Differenzierung zwischen wirklich hautfarbenen und anderen Objekten daher nicht ausreichend gut. Das Merkmal Hautfarbe kann jedoch gut mit weiteren Merkmalen zur Lokalisierung von Gesichtern kombiniert werden.

5 Verfolgen von Personen

Nachdem eine Person im Bild lokalisiert wurde, müssen ihre Bewegungen verfolgt werden. Die Analyse der Bewegungen einer Person über die Zeit bildet die Grundlage für das Erkennen von Gesten. Um ein Tracking der Bewegung von Personen in Echtzeit auf den begrenzten Hardware-Ressourcen eines mobilen

Roboters zu ermöglichen, werden ausschließlich die Bewegungen des Kopfes und der beiden Hände verfolgt. Aus der in Abschnitt 2 beschriebenen Vielzahl von Möglichkeiten zum Schätzen von Haltung und Bewegung einer Person wird somit ein einfaches Verfahren ausgewählt. Die Effizienz des Verfahrens beruht darauf, dass die Lokalisierung der Positionen von Kopf und Händen auf der Lokalisierung hautfarbener Bereiche beruht. Da sich solche Bereiche schnell lokalisieren lassen, ist das Verfahren echtzeitfähig und wird in vielen Arbeiten angewandt [1] [8].

5.1 Adaptives Modell der Hautfarbe

Die Lokalisierung eines Gesichtes mit Hilfe eines Viola-Jones-Detector-basierten Verfahrens ist zuverlässiger als seine Lokalisierung nur aufgrund der Hautfarbe. Dabei ist das Viola-Jones-Detector-basierte Verfahren jedoch zu rechenaufwändig, um zum Tracking des Gesichtes in Echtzeit bei langsamer Hardware geeignet zu sein. Es wird daher in dieser Arbeit nur zum initialen Lokalisieren verwendet, bei dem das Verarbeiten nur sehr weniger Bilder pro Sekunde unproblematisch ist. Mit Hilfe der Farben im Bereich des initial lokalisierten Gesichtes kann dann ein Modell der Hautfarbe unter den aktuellen Beleuchtungsbedingungen erstellt werden, das zum effizienten Lokalisieren des Gesichtes als hautfarbenem Bereich beim Tracking genutzt wird.
Als Modell der Hautfarbe unter den aktuellen Beleuchtungsbedingungen kann beispielsweise ein Histogramm oder eine 2-dimensionale Normalverteilung verwendet werden. In beiden Fällen wird jeder Farbe (r, b) die Wahrscheinlichkeit, unter den aktuellen Beleuchtungsbedingungen eine Hautfarbe zu sein, zugeordnet.

Modell-Wahrscheinlichkeiten für einzelne Pixel

Mit Hilfe des Modells der Hautfarbe kann jeder Position (x, y) im Bild I die Wahrscheinlichkeit $P_{skin}(x, y)$ zugeordnet werden, bei den aktuellen Beleuchtungsbedingungen in einem hautfarbenen Bereichen zu liegen. In Abbildung 5.1 ist eine solche Zuordnung dargestellt. Jedem Pixel aus dem linken Bild ist seine Wahrscheinlichkeit, zu einem hautfarbenen Bereich zu gehören, zugeordnet. Die Zuordnung erfolgt mit dem Histogramm als Farbmodell, das im mittleren Bild dargestellt ist. Das Ergebnis ist das rechte Bild, wobei dunkle Pixel eine hohe Wahrscheinlichkeit, zu einem hautfarbenen Bereich zu gehören, anzeigen.

Abbildung 5.1: Jedem Pixel wird über ein adaptives Farbmodell die Wahrscheinlichkeit zugeordnet zu einem hautfarbenen Bereich zu gehören.

Adaption des Modells an variierende Beleuchtungsbedingungen
Ein Farbmodell muss sich Veränderungen in den Beleuchtungsbedingungen anpassen. So ergibt sich eine Folge von Farbmodellen FM_t, wobei sich das Modell in jedem Zeitschritt t = 1...T ändern kann. Zur Initialisierung des adaptiven Modells wird auf das Ergebnis der initialen Lokalisierung eines menschlichen Gesichts zurückgegriffen.
Wie die Lage des Gesichts im aktuellen Bild bestimmt wird, wird im weiteren Verlauf des Artikels beschrieben. Ist die Lage des Gesichts bestimmt, so kann sie genutzt werden, um eine Adaption des Farbmodells an langsame Veränderungen der Beleuchtungsbedingungen zu erreichen. Dies geschieht, indem das Farbmodell in jedem Zeitschritt aus dem bisherigen Farbmodell und den Farben innerhalb des aktuell lokalisierten Gesichtes kombiniert werden.
Die Zuverlässigkeit der Lokalisierung von Gesicht und Händen mittels eines adaptiven Farbmodells wird gesteigert, wenn man sicherstellt, dass nur Farben aus dem Bereich des skin locus der benutzten Kamera zur Adaption genutzt werden [6]. Alle Farben, die nicht im skin locus liegen, werden daher beim Bestimmen des Farbmodells nicht berücksichtigt.

5.2 Bestimmen hautfarbener Bereiche

Mit Hilfe des adaptiven Farbmodells sollen zusammenhängende hautfarbene Bereiche im Bild bestimmt werden, da diese die Position und Größe von Gesicht und Händen beschreiben. Dazu werden mit Hilfe von Schwellwert-Operationen die Pixel bestimmt, die unter den gegebenen Beleuchtungsbedingungen als hautfarben angenommen werden und zu connected components zusammengefasst.

Abbildung 5.2: Bestimmen von hautfarbenen Bereichen als Connected Components von Pixeln, die mit hoher Wahrscheinlichkeit zu diesen gehören.

6 Erkennen von Gesten

Im bisherigen Teil wurden Techniken der Farbbildverarbeitung beschrieben, die die Basis für eine visuelle Gestenerkennung bilden. Im Folgenden werden die weiteren Schritte zum Erkennen von Gesten beschrieben. Da sich dieser Teil weit vom Thema Farbbildverarbeitung entfernt, werden die verwendeten Methoden nur kurz dargestellt. Für eine ausführlichere Darstellung der gesamten Methodik sei auf [2] verwiesen.

6.1 Extraktion geeigneter Merkmale

Unter den hautfarbenen Bereichen werden der Kopf und die Hände identifiziert und deren Bewegung wird mit Hilfe eines adaptiven Kalman-Filters von Messfehlern bereinigt. Verwechslungen zwischen den Körperteilen werden mittels einiger Regeln vermieden, die typische Konfigurationen des Kopfes und der Hände zueinander berücksichtigen. Das Ergebnis ist eine Zeitreihe von ellipsenförmigen Körperteil-Modellen, die die Lage von Kopf und Händen im Bild beschreiben.

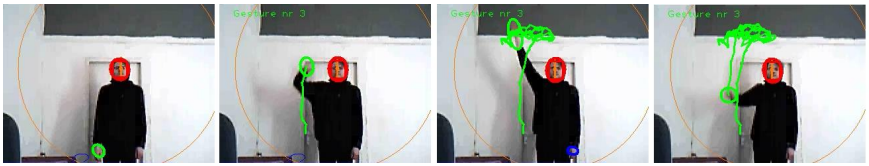

Abbildung 6.1: Ergebnis des Verfolgens von Kopf und Händen beim Winken mit der nach oben ausgestreckten rechten Hand

6.2 Methoden zur Klassifikation

Die extrahierten Merkmale werden bezüglich der Lage und Position der gestikulierenden Person normiert. Es folgt eine Segmentierung der Merkmalssequenz in die Segmente, die jeweils der Durchführung einer einzelnen Geste entsprechen. Dazu wird eine Grundposition erkannt, die eine gestikulierende Person zwischen den einzelnen Gesten einnehmen muss. Die resultierenden Merkmalssequenzen zu jeder Geste werden dann in unterschiedlicher Form codiert und mit typischen Klassifikationsverfahren wie den Hidden-Makrov-Modellen und Support-Vektor-Maschinen klassifiziert.

7 Ergebnisse

7.1 Auswahl der zu Erkennenden Gesten

Die beschriebene Methode zur Extraktion gestenrelevanter Merkmale aus Videosequenzen beschränkt sich auf das Erkennen und Verfolgen von Kopf und Händen in der Bildebene. Viele Gesten lassen sich gut durch die relative Lage der Hände zum Kopf beschreiben. Dabei handelt es sich insbesondere um Gesten, die in der zwischenmenschlichen gestenbasierten Interaktion für die Kommunikation über größere räumliche Distanzen genutzt werden. Während Menschen bei einem normalen Gespräch typischerweise die Mimik einsetzen und mit kleineren Handbewegungen gestikulieren, nutzen sie über größere räumliche Distanzen klare Bewegungen der Arme. Einen Beispiel ist die gestenbasierte Kommunikation beim Einweisen von Fahrzeugen. Dabei soll eine Person als Einweiser einer zweiten Person, die ein Fahrzeug führt, über Gesten vermitteln, wie sie das Fahrzeug steuern soll. Für unsere Zwecke wurden 8 Gesten aus einer Vorschrift für das Einweisen von Kraftfahrzeugen als Grundlage für die Verifikation des Verfahrens gewählt.

Für die Verifikation der Gestenerkennung wurden Videos von der Durchführung von 212 Gesten aufgenommen, die sich zu etwa gleichen Teilen aus den 8 Beispielgesten zusammensetzen und von 9 verschiedenen Personen durchgeführt wurden.

Abbildung 6.2: Satz von Beispiel-Gesten zum Einweisen von Fahrzeugen

7.2 Ergebnisse beim Erkennen der Beispielgesten

Bei einer Vielzahl von Experimenten mit verschiedenen Codierungen der Merkmale und verschiedenen Methoden zur Klassifikation wurde ein Erkennungsrate von bis zu 0.9 erzielt für Gesten, die von einer Person durchgeführt wurden, die beim Training der Klassifikationsverfahren nicht berücksichtigt worden war.

Literatur

[1] Argyros, Antonis A. ; Lourakis, Manolis I. A.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: ECCV, 2004, S. 368–379

[2] Bertsch, Florian A. and Hafner, Verena V.: Real-time dynamic visual gesture recognition in human-robot interaction. In: 9th IEEE-RAS International Conference on Humanoid Robots (2009). accepted

[3] Elmezain, M. ; Al-Hamadi, A. ; Appenrodt, J. ; Michaelis, B.: A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory. In: 19th International Conference on Pattern Recognition (2008), Dec., S. 1–4

[4] Kehl, Roland ; Gool, Luc V.: Markerless tracking of complex human motions from multiple views. In: Computer Vision and Image Understanding 104 (2006), Nr. 2-3, S. 190 – 209

[5] Martinkauppi, B.: Face colour under varying illumination - analysis and applications. University of Oulu, Diss., 2002

[6] Martinkauppi, Birgitta ; Soriano, Maricor ; Pietikäinen, Matti: Comparison of skin color detection and tracking methods under varying illumination. In: Journal of Electronic Imaging 14 (2005), Nr. 4

[7] Moeslund, Thomas B. ; Granum, Erik: A Survey of Computer Vision-Based Human Motion Capture. In: Computer Vision and Image Understanding 81 (2001), Nr. 3, S. 231–268

[8] Ng, Chan W. ; Ranganath, Surendra: Real-time gesture recognition system and application. In: Image and Vision Computing 20 (2002), S. 993–1007

[9] Poppe, Ronald: Vision-based human motion analysis: An overview. In: Computer Vision and Image Understanding 108 (2007), S. 4–18

[10] Vassili, Vladimir V. ; Sazonov, Vassili ; Andreeva, Alla: A Survey on Pixel-Based Skin Color Detection Techniques. In: Proc. Graphicon-2003, 2003, S. 85–92

[11] Viola, Paul ; Jones, Michael J.: Robust Real-Time Face Detection. In: Int. J. Comput. Vision 57 (2004), Nr. 2, S. 137–154

[12] Yang, Ming-hsuan ; Kriegman, David J. ; Ahuja, Narendra: Detecting faces in images: A survey. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002), S. 34–58

[13] Yoon, H.S. ; Soh, J. ; Bae, Y.J. ; Seung Yang, H.: Hand gesture recognition using combined features of location, angle and velocity. In: Pattern Recognition 34 (2001), Nr. 7, S. 1491–1501