artec Paper Nr. 42

Gestenerkennung mit einem Datenhandschuh

Volker Brauer
Forschungszentrum Arbeit und Technik (artec)
Universität Bremen
März 1996

Einleitung

Gestik ist ein Teil der menschlichen Kommunikation. Wir setzen Gestik ein, um wort- und geräuschlos Signale auszusenden oder um sprachergänzend miteinander zu kommunizieren. Oftmals sind wir uns unserer Körpersprache jedoch nicht bewußt und üben gestisches Verhalten im Verständigungsprozeß automatisch aus. Inwieweit Gestik für die Interaktion mit Computern genutzt werden kann, ist bisher nur wenig erforscht, auch fehlt es an zuverlässigen, universellen Verfahren, mit denen menschliche Gesten vom Rechner erkannt und analysiert werden können. In diesem Artikel wird ein auf statistischen Methoden basierender Klassifizierer für Gesten vorgestellt.

1 Gestik im Kontext der Mensch-Maschine Kommunikation

Der Begriff „Gestik“ findet in der Informatik vielfältige Verwendung. Im Bereich des Pen-Computing und der Schrifterkennung werden handschriftliche Eingaben eines Benutzers als Gesten bezeichnet [KIM88, CARR91]. Mit der Erkennung von Handbewegungen, die mit einer konventionellen Maus aufgezeichnet werden, haben sich u.a. Rubine [RUBI91] sowie Wilson und Bobick [WILS95] beschäftigt. Aus sprachwissenschaftlicher Sicht ist Gestik eine Verhaltensform des Menschen, die alle Körperteile bzw. die gesamte Körperhaltung einbezieht, wobei den Händen die größte Bedeutung zukommt [EKMA79]. Einen Ansatz für die Gestenerkennung im gesamtheitlichen Zusammenhang von Körpersprache und Sprache hat Wexelblatt am MIT entwickelt [WEXE94]. Aufgrund praktischer Überlegungen werden im folgenden nur einhändige Hand- und Fingerbewegungen untersucht.

Bei der Betrachtung von Gestik als Eingabe für den Computer müssen sowohl die besonderen Voraussetzungen ihrer Verwendung als Medium zur Übermittlung von Informationen an eine Maschine (anstatt eines Menschen), als auch die rein physiologischen und anatomischen Eigenschaften der Hand berücksichtigt werden. Anders als bei zwischenmenschlicher Kommunikation, bei der Gestik als Ausdrucksform von so unterschiedlichen Faktoren wie Kultur, Sozialisation und Gefühlszustand beeinflußt und durch ein gegenseitiges Verständnis dafür interpretierbar wird, ist gestische Kommunikation zwischen Mensch und Maschine anwendungs- bzw. zweckgebunden und durch das primitive „Wahrnehmungssystem“ des Rechners gekennzeichnet. Daher ist es notwendig, einfache und formalisierbare Eigenschaften von Gesten zu finden, die deren algorithmisches Erkennen ermöglichen. Zudem ist der rein sprachliche Gebrauch von Gestik, wie es unter Menschen der Fall ist, hinsichtlich möglicher Computer-Anwendungen allgemeiner zu fassen. Im Alltagsleben werden viele manuelle Tätigkeiten wie das Betätigen eines Schalters nicht mit einem kommunikativen Hintergrund ausgeführt, sondern nur, um einen bestimmten Effekt zu erzielen. Aber gerade die Handlungen, die auch reale Objekte der Umgebung einbeziehen, sind aus Anwendungssicht interessant [BRUN93]. Daher ist es sinnvoll, den Gestik-Begriff aus dem Kontext von Zeichensprache oder Sprache zu lösen und ihn um allgemeine, zielgerichtete Handbewegungen zu erweitern.

In der Literatur (vergl. u.a. [STUR92, BORD93]) wird häufig zwischen statischen Gesten (Posen) und dynamischen Gesten unterschieden, wobei Bewegungen der Hand oder der Finger in einem Zeitraum als dynamische Gesten bezeichnet werden, und unter einer Pose der Zustand (Konfiguration der Gelenke) der Hand zu einem bestimmten Zeitpunkt t verstanden wird. Diese Differenzierung ist im Sinne der Gestenerkennung zweckmäßig, weil das Berücksichtigen der Bewegungsdynamik der Hand anspruchsvollere Anforderungen an ein Erkennungsverfahren stellt als bei statischen Zuständen.

2 Eingabegeräte für Gesten

Die menschliche Hand gehört zu den flexibelsten Greifwerkzeugen, die in der Natur vorkommen. Mit Hilfe eines komplizierten Systems von Gelenken, Sehnen und Muskeln wird eine hohe Beweglichkeit der Finger und des Handgelenkes erreicht (vergl. Abb. 1), an die ein durchschnittlicher Industrieroboter nicht im Entferntesten heranreicht. Bei seiner Zählung der Freiheitsgrade (FG), die ein wichtiges Maß für die Beweglichkeit sind, kommt Sturmann auf 23, ohne die Positionierbarkeit und Orientierung der Hand im Raum (6 FG) mitzuzählen [STUR92]. Die komplexen Bewegungsabläufe, die mit der Hand ausgeführt werden können, stellen hohe Anforderungen an ein Aufzeichnungssystem, das die Bewegungs- und Zustandsdaten in für den Rechner weiterzuverarbeitende Werte umwandelt. Für Zwecke der Gestenerkennung werden im wesentlichen zwei Techniken eingesetzt. Zum einen gibt es eine Reihe von Ansätzen, bei denen die Handbewegungen mit Videokameras aufgenommen und mit Bildverarbeitungsverfahren analysiert werden. Andererseits haben sich Spezialhandschuhe, die mit einer Aufzeichnungssensorik für die Handkonfiguration ausgestattet sind [STUR94], weit verbreitet. Der Vorteil der bildverarbeitenden Verfahren gegenüber den Datenhandschuhen liegt darin, daß der Beobachtete keine behindernde technische

Ausstattung tragen muß. Zudem kann das Bildmaterial auch zur Überwachung anderer Körperteile oder Objekte genutzt werden [DARR95]. Nachteilig an dieser Technik ist der relativ hohe Bedarf an Rechenleistung und die im Vergleich zu sensorbasierten Systemen wie dem Datenhandschuh ungenauen Daten über Krümmungs- oder Spreizwinkel der Finger. Darüberhinaus kann ein Datenhandschuh ohne großen Aufwand an verschiedene Rechner und Benutzer angepaßt werden und ist gegenüber den eher stationär verwendeten Kameras mobil einsatzfähig.

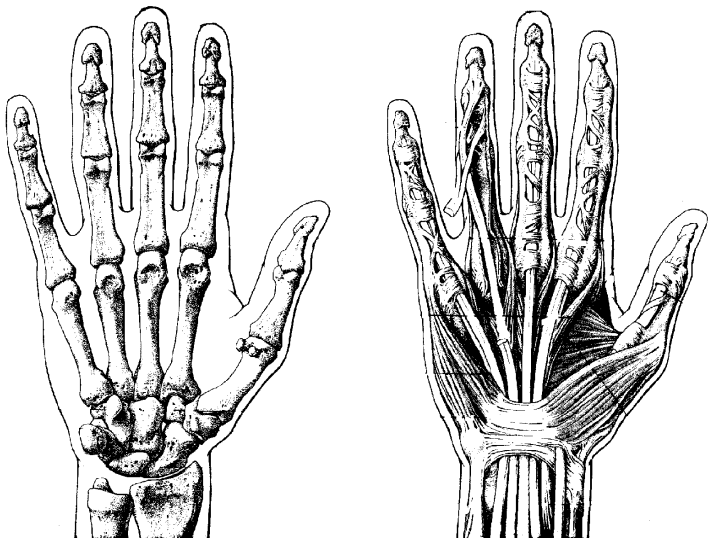

Abb. 1: Anatomie der Hand

Zusätzlich zu den Informationen über die Konfiguration der Hand, werden insbesondere für die Erkennung dynamischer Gesten Daten über die Orts- und Lageveränderung der Hand benötigt. Für diesen Zweck sind auf dem Markt sog. „Position Tracker“ (Positionsverfolger) erhältlich. Je nach verwendeter Technik wird entweder eine Sende- oder eine Empfangseinheit auf das Objekt, das verfolgt werden soll, montiert. Bei den professionellen Systemen von Polhemus oder Ascension Technologies wird ein Empfänger für elektromagnetische Wellen, die von einer stationären Quelle emittiert werden, auf dem Trägerobjekt installiert. Abhängig von der gemessenen Feldstärke kann auf die Entfernung zur Quelle und damit die relative Position im Raum zu einem festgelegten Fixpunkt errechnet werden. Indem der Einfallswinkel der Wellen auf den Empfänger gemessen wird, kann außer auf die Position auch auf die Orientierung des Objektes bzw. des Empfängers zurückgeschlossen werden. Als Alternative zu Magnetwellen wird bei anderen Trackern auch Ultraschall [HOMM94] oder Infrarotlicht eingesetzt.

Für das hier vorgestellte System wurde der primitive, aber sehr kostengünstige Mattel PowerGlove verwendet. Dieser Datenhandschuh verfügt über ein einfaches Ultraschall Ortungssystem und liefert für die Finger (außer dem kleinen Finger) je vier Beugungswerte, die den Grad ihrer Krümmung bestimmen. Zudem wird die Rotation um die Tiefenachse des Raumes erfaßt. Insgesamt werden acht verwertbare Daten über den Zustand der Hand an den Rechner übermittelt, die mit einer Frequenz von ca. 10 bis 12 Hz aktualisiert werden. Obwohl dieses Eingabegerät nur ein Minimum an Zuverlässigkeit, Komfort und Datenmaterial bietet, wird es für diverse PC-basierte Virtual Reality Systeme eingesetzt und wurde auch schon für Zwecke der Gestenerkennung benutzt [HARL93]. Eine weitere Anwendung für diesen Datenhandschuh ist das Erfassen und Erkennen von Griffmustern, die beim Greifen realer Objekte entstehen. In einem sog. „Real Reality“ Projekt an der Universität Bremen wird untersucht, wie durch Manipulation gegenständlicher Modelle simultan Änderungen in einem rechnerinternen (virtuellen) Abblild vorgenommen werden können [BRUN93, BRAU96].

3 Ein Werkzeug zum Definieren, Trainieren und Analysieren von Handgesten

Die Tatsache, daß es nur sehr selten gelingt, bezüglich ihrer Meßwerte identische Handbewegungen auszuführen, macht deren Erkennung zu einem Problem, das schwierig zu lösen ist. Mit Hilfe mathematischer Verfahren kann lediglich festgestellt werden, ob eine Handbewegung einer dem Erkennungssystem be-

kannten Menge von Handbewegungen ähnelt. Aus dieser Erkenntnis lassen sich funktionale und strukturelle Anforderungen an ein System zur Gestenerkennung ableiten.

- Damit die Gestenerkennung für verschiedene Anwendungen verwendbar ist, müssen Gestiken als anwendungsbezogene Menge unterscheidbarer Gesten frei definiert werden können. Eine statisch vorgegebene Menge von Gesten ist nicht sinnvoll, da es keine allgemeinen Standards gibt und benutzerspezifische Eigenarten bei Handbewegungen zu berücksichtigen sind.
- Aus der Voraussetzung, daß Gesten aus einer dem System bekannten Menge von Gesten (wieder)erkannt werden müssen, ergibt sich die Notwendigkeit eines Werkzeuges, mit dem Gestiken eingegeben bzw. trainiert werden können.
- Dem erfolgreichen Wiedererkennen von Gesten liegt eine Datenmenge zugrunde, die einigen Qualitätsanforderungen genügen muß. Verrauschte Meßwerte oder große Varianzen beim Trainieren einer Gestik lassen sich nur mit adäquaten Analyseverfahren feststellen.
- Weil es wenig Sinn ergibt, ein Erkennungsverfahren zum Selbstzweck zu konstruieren, sollte dieses leicht in übergeordnete Anwendungssysteme integrierbar sein.

Das hier beschriebene System zur Gestenerkennung besteht im wesentlichen aus einem Trainingswerkzeug (Trainer) mit integriertem Analysemodul und einem Erkennungsmodul (Klassifizierer), der zum Trainieren und für spezifische Anwendungen verwendet wird. Abb. 2 stellt die Struktur grafisch dar.

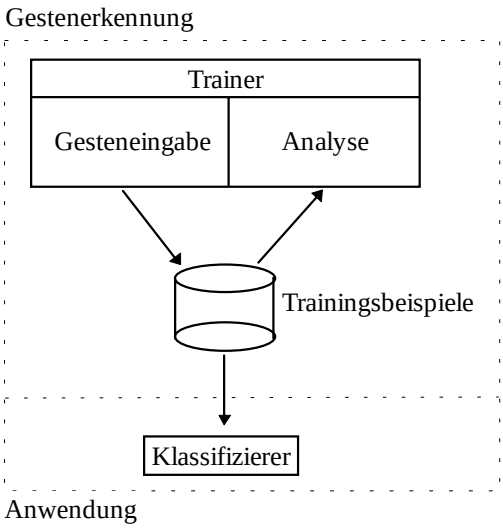

Abb. 2: Struktur des Gestenerkennungssystems

Der Klassifizierer besteht aus einer Reihe von C++-Klassen, die einem Anwendungssystem eine wohldefinierte Schnittstelle bieten. Bevor der Klassifizierer benutzt werden kann, muß mit Hilfe des Trainingswerkzeuges eine Gestik definiert und trainiert werden, die in einer Datei permanent gespeichert wird. Im folgenden wird kurz die Funktionalität des Trainers erläutert (vergleiche auch [BRAU94]).

3.1 Definition von Gestiken

Eine Gestik bzw. Gestensprache besteht aus einer Menge von Posen und einer Menge dynamischer Gesten, wobei jeder Pose bzw. dynamischen Geste eine Nummer, ein Name und eine Kurzbeschreibung zugeordnet wird. Diese Definition erfolgt interaktiv innerhalb des Trainers und kann jederzeit verändert werden.

3.2 Das Trainieren von Gesten

Im Eingabemodus des Trainers wird dem Benutzer eine grafische, perspektivische Abbildung einer Hand auf dem Bildschirm angezeigt, die sich analog zu den Handbewegungen ändert. Zudem können Informationen über den Status des Datenhandschuhs und über bereits erfaßte Trainingsbeispiele abgerufen werden. Um Eingaben über die Tastatur zu vermeiden, werden im Trainingsmodus die Funktionstasten des Datenhandschuhs zur Steuerung des Programms benutzt.

Das Trainieren von Posen erfolgt sehr einfach, indem der Benutzer die Hand entsprechend der gewählten Pose formt und eine Taste des Datenhandschuhs drückt. Dadurch werden die Konfigurationsdaten des Handschuhs dem gewählten Beispiel zugeordnet. Sollte bereits ein Trainingsbeispiel mit gleichen Werten existieren, wird die Eingabe solange ignoriert bis der Benutzer durch geringfügiges Ändern der Konfiguration abweichende Werte erzeugt oder der Vorgang abgebrochen wird. In der Regel sind je Pose nicht mehr als fünf Trainingsbeispiele erforderlich. Praktische Erfahrungen haben gezeigt, daß das Trainieren einer Pose durchschnittlich ca. 2-3 Minuten dauert.

Im Vergleich zur simplen Trainingsmethode für Posen, erfordert das Trainieren dynamischer Gesten etwas mehr Zeit und Sorgfalt. Um ein zuverlässiges Erkennen zu gewährleisten, müssen mehr Beispiele eingegeben werden, die zudem ein gewisses Maß an „Qualität“ zu erfüllen haben. Das Klassifizierungsverfahren beruht auf einer nicht zufälligen Verteilung der Merkmalswerte einer Geste im n-dimensionalen Merkmalsraum, d.h. mit der Eingabe von Trainingsbeispielen wird ein Bereich im Merkmalsraum lokalisiert, der den spezifischen Eigenschaften einer Klasse von Gesten entspricht (vergl. Abschnitt 4.2). Je größer die Anzahl der Trainingsbeispiele ist, um so genauer kann durch Mittelwertbildung die „durchschnittliche“ Geste einer Gestenklasse ermittelt werden. Im vorliegenden praktischen Fall hat sich gezeigt, daß 15 Beispiele pro Geste für die Klassifizierung ausreichen. Aufgrund einiger technischer Mängel des Datenhandschuhs können die Meßdaten selbst bei subjektiv gleichen Handbewegungen stark voneinander abweichen. Daher ist es unter Umständen notwendig, einzelne Beispiele während der Trainingsphase wiederholt einzugeben. Dennoch wurden im Versuch selten mehr als 5 Minuten benötigt, um eine Geste mit 15 „guten“ Beispielen zu trainieren. Damit möglicherweise untaugliche Trainingsbeispiele schnell identifiziert werden können, wurden diverse Analyseverfahren entwickelt, die im nächsten Abschnitt erläutert werden.

3.3 Analyse der Trainingsdaten

Die durch das Trainieren erzeugten Zahlenwerte, die den Bewegungsablauf der Hand charakterisieren, können mit statistischen Methoden untersucht werden. Das Ziel einer Datenanalyse ist es, festzustellen, ob mit dem Datenmaterial eine erfolgreiche Gestenerkennung zu erwarten ist, und ggf. untaugliche Trainingsbeispiele zu lokalisieren. Die Untersuchung der Daten erfolgt auf den drei Ebenen Trainingsbeispiel, Geste und Gestik. Es werden, ausgehend von den grundlegenden Werten der Beispiele, Annahmen über die abstrakteren Ebenen Geste und Gestik gemacht. Die folgende Tabelle zeigt die Meßwerte von drei Trainingsbeispielen für eine dynamische Geste.

Nr.	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}
0.	77,63	6,17	29,50	3,83	1,74	37,13	193,56	26,83	203,30	30,79	24,72	37,00	197,00	4,28	0,30
1.	58,99	4,67	21,33	3,67	1,81	31,05	146,74	21,00	175,82	22,05	25,96	28,00	138,00	3,63	0,87
2.	65,49	5,67	18,33	5,67	1,85	34,29	140,15	9,09	192,31	26,27	25,57	34,00	148,00	3,36	0,20

Die Spalten f1 bis f15 entsprechen den Merkmalen (Features) einer Geste, wie sie in Abschnitt 4.3 definiert sind. Weil die Daten in dieser Darstellung kaum zu bewerten sind, werden sie dem Benutzer zusammengefaßt in anderer Form präsentiert.

Feature	Mean	Min.	Ex.	Max.	Ex.	Std.Dev.	Relation
BoxDia	78,34	58,45	[10]	99,19	[14]	13,1551	16,8%
InOut	5,54	4,67	[01]	6,17	[00]	0,3468	6,3%
UpDown	27,94	18,33	[02]	40,17	[06]	6,5228	23,3%
LeftRight	4,38	3,33	[11]	6,00	[09]	0,7946	18,2%

Über alle Trainingsbeispiele einer Geste wird für jedes Merkmal der Mittelwert, das Minimum, das Maximum und die Standardabweichung ermittelt. In der rechten Spalte wird die prozentuale Abweichung der Standardabweichung vom Mittelwert angezeigt, was ein gutes Maß für die Streuung der Werte ist. Je kleiner dieser Wert ist, um so weniger variieren die Werte. Die Zahlen in den eckigen Klammern geben an, welches Beispiel ein Minimum- oder Maximumwert verursacht. Ungeeignete Beispiele besitzen Werte, die von den Mittelwerten stark abweichen. Die Folge ist eine breite Streuung und eine relativ häufige Kennzeichnung eines solchen Beispiels in der Tabelle. Wie oft ein Beispiel mit Extremwerten (Minimum oder Maximum) in der Tabelle vertreten ist, wird durch die Datenanalyse ersichtlich. Um den Anwender auf weniger geeignete Eingaben hinzuweisen, wird eine Liste der Beispiele und deren Anzahl von Extremwerten ausgegeben.

Beispiel Nr.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Extrema	3	3	2	1	0	2	3	1	1	2	5	2	1	0	4

Der oben dargestellten Tabelle ist zu entnehmen, daß Beispiel Nr. 10 fünf Extremwerte aufweist und die anderen Extrema sich relativ gleichmäßig auf die einzelnen Beispiele verteilen. Dieses ist ein Hinweis für den Benutzer, ggf. die Eingabe für das Beispiel Nr. 10 zu wiederholen.

Um zu testen, ob sich die Gesten bezüglich ihrer Trainingsdaten untereinander beeinflussen, wurde ein Selbsttestverfahren für die Gesten entwickelt. Mit dem Selbsttest kann geprüft werden, ob die Trainingsbeispiele einer Geste korrekt klassifiziert werden, d.h. jedes Beispiel wird als Eingabe für den Klassifizierungsalgorithmus benutzt und es wird festgestellt, als welche Geste das Beispiel erkannt wird. Das Ergebnis einer solchen Analyse wird nachstehend für eine Geste namens „Greifen“ dargestellt.

Gesture Nr: 0 Name: Greifen

Example: 0 in Class: 0 ok
Example: 1 in Class: 0 ok
Example: 2 in Class: 0 ok
Example: 3 in Class: 0 ok
Example: 4 in Class: 0 ok
Example: 5 in Class: Ziehen error
Example: 6 in Class: 0 ok
Example: 7 Rejected because of Distance Measure
•
•

Diesem Selbsttest ist zu entnehmen, daß die Beispiele Nr. 5 und 7 nicht erfolgreich klassifiziert werden konnten. Beispiel Nr. 5 wurde als Geste „Ziehen“ identifiziert und Nr. 7 wurde vom Klassifizierer keiner Geste zugeordnet. Es ist zweifellos eine grundlegende Voraussetzung für eine zuverlässige Klassifizierung, daß wenigstens die Trainingsbeispiele fehlerfrei erkannt werden. Sollte es auffällig viele falsche Zuordnungen der Beispiele zwischen einzelnen Klassen geben, so ist das ein Hinweis darauf, daß die Gesten sich bezüglich ihrer Werte nicht hinreichend abgrenzen lassen. Die Ursache dafür ist entweder eine zu hohe Streuung der Werte oder eine große Ähnlichkeit zwischen den Handbewegungen. Damit die Daten detaillierten Analysen unterzogen werden können, ist im Trainer eine Schnittstelle enthalten, mit der sie an externe Programme, wie z.B. einer Tabellenkalkulation, übertragen werden können. Abb. 3 zeigt eine grafische Darstellung ausgewählter Merkmale für eine Testgestik.

Das Diagramm (Abb. 3) verdeutlicht die Verteilung der Werte zweier Merkmale in ihrem Wertebereich. Entlang der X-Achse sind senkrechte Linie aufgetragen, die jeweils eine Geste (Stichprobenklasse) eingrenzen. Es ist erkennbar, wie sich innerhalb der Klassen die Werte häufen, und daß die Häufungen zwischen den Klassen sich vielfach unterscheiden, d.h. verschiedene Schwerpunkte gebildet werden. Die Gesten, deren Punktewolken beider Merkmale im gleichen Wertebereich liegen, können bezüglich dieser Merkmale nicht unterschieden werden.

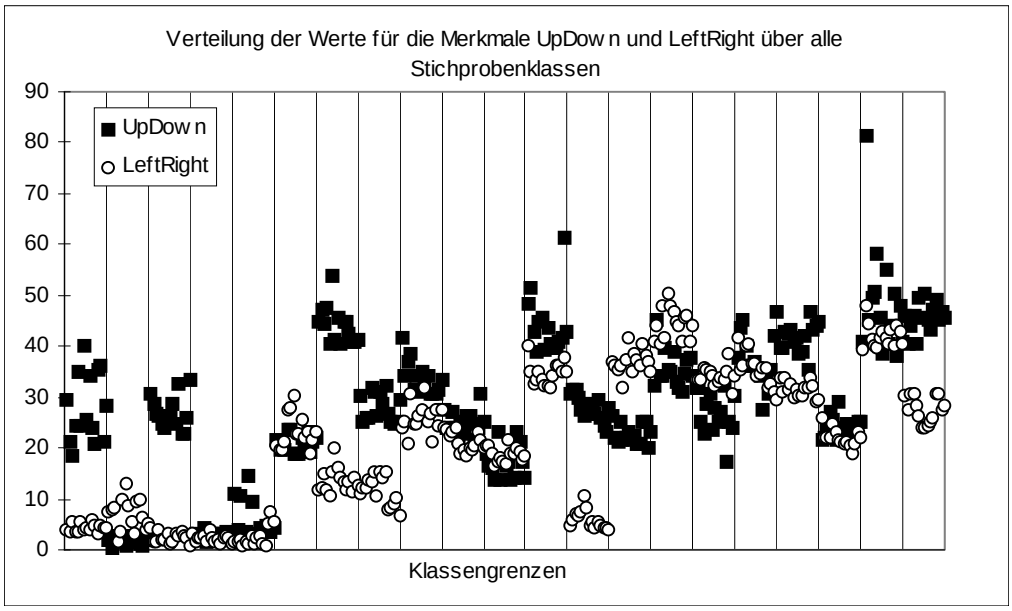

Abb. 3: Werteverteilung ausgewählter Merkmale

4 Handgestenerkennung

Insbesondere seitdem die ersten Datenhandschuhe kommerziell verfügbar sind, wird auf dem Gebiet der Handgestenerkennung geforscht. In ersten Ansätzen wurde daran gearbeitet, die Hand als Ersatz für konventionelle Eingabegeräte zu verwenden, um beispielsweise durch virtuelle Räume zu navigieren und einfache Kommandos auszuführen. Für diese Zwecke wurden vorwiegend Verfahren für die Erkennung von Posen entwickelt. Das Erkennen dynamischer Gesten ist eine wesentlich komplexere Aufgabe, weil eine Reihe weiterer Parameter wie Zeit, Ortsveränderungen, Geschwindigkeit usw. zu berücksichtigen sind. Für Anwendungen wie die Erkennung von Zeichensprache oder die Interaktion mit (virtuellen) Objekten ist aber Wissen über die Bewegungsdynamik erforderlich.

4.1 Stand der Technik

Betrachtet man Handgestenerkennung als einen Anwendungsfall für Mustererkennungsverfahren, so gibt es eine Reihe von Techniken, mit denen Problemlösungen denkbar sind. In bisherigen Veröffentlichungen werden jedoch zwei favorisiert. Zum einen sind dies Neuronale Netze und zum anderen statistische Klassifizierungsverfahren, die auf Diskriminanzanalysemethoden beruhen. Einen weiteren vielversprechenden Ansatz haben Starner und Pentland unter Verwendung von Markov Modellen entwickelt [STAR95].

Ein mit Gesten steuerbares elektronisches Präsentationssystem, wobei statistische Methoden zur Gestenerkennung eingesetzt wurden, hat Baudel entwickelt [BAUD93]. Bordegoni und Hemmje haben eine Gestik zur Interaktion mit dreidimensionalen virtuellen Objekten entworfen, die mit Neuronalen Netzen erkannt wird [BORD93]. In ihren Glove-Talk Projekten haben Fels und Hinton erfolgreich Neronale Netze verwendet, um Zeichensprache mit Hilfe eines Sprachgenerators in Sprache umzuwandeln [FELS93, 95]. Roberts hat mit einem statistischen Verfahren ein Programm für die Erkennung des australischen Zeichenalphabets entwickelt und damit eine Erkennungsrate von 94% erreicht [ROBE94]. Etwas ähnliches, aber mit Neuronalen Netzen, haben Takahashi und Kashino für das japanische Zeichenalphabet versucht [TAKA91]. Sie konnten aber lediglich 34 von insgesamt 46 Posen korrekt erkennen. Ein System zur Gestenerkennung, das außer den Handbewegungen auch andere Körperteile einbezieht, hat Wexelblatt am MIT realisiert [WEXE94].

Die bisherigen Entwicklungen auf dem Gebiet der Gestenerkennung sind vorwiegend auf spezielle Anwendungsfälle ausgerichtet. Es fehlen z.Z. allgemeinere Ansätze, durch die gestische Eingabe für einen größeren Kreis von Anwendungen zugänglich wird. Auch bezüglich der bisher erzielten Erkennungsraten sind noch Verbesserungen notwendig, damit eine ausreichende Akzeptanz bei den Benutzern erreicht wird.

4.2 Klassifizierung von Handgesten

Aufgrund der ungleichen Eigenschaften von Posen und dynamischen Gesten wurden im hier verfolgten Ansatz zu deren Erkennung zwei voneinander unabhängige, in ihrer Funktionsweise unterschiedliche Techniken entwickelt. Dynamische Gesten werden mit einem multivariaten Diskriminanzanalyseverfahren klassifiziert. Posen werden erkannt, indem festgestellt wird, ob für eine zu klassifizierende Pose p ein Trainingsbeispiel b existiert, mit b = p, wobei eine ungefähre Gleichheit der Meßwerte des Datenhandschuhs zu einem Zeitpunkt t geprüft wird. Diese einfache Methode zur Posenerkennung ist angesichts der sehr geringen Auflösung des Datenhandschuhs bezüglich der Beugungswerte der Finger ausreichend genau und effizient.

Der vom Algorithmus anspruchsvollere und interessantere Fall ist der der dynamischen Gesten. Während eine Pose als eine Momentaufnahme des Zustandes der Hand aufgefaßt werden kann, ist eine dynamische Geste eine Menge zeitlich aufeinanderfolgernder Informationen über die Position und Konfiguration der Hand. Die Erkennung dynamischer Gesten erfordert daher eine gesamtheitliche Betrachtung der Daten, die während der Phase des Gestikulierens aufgezeichnet werden. Dadurch, daß ganze Meßreihen untersucht werden - beim PowerGlove ca. 10-12 Meßtupel pro Sekunde - können aus den Rohdaten des Datenhandschuhs charakteristische Eigenschaften von Handbewegungen berechnet werden. Eine wichtige Aufgabe besteht darin, diese Eigenschaften (Features) zu definieren und Berechnungsfunktionen dafür zu implementieren. Ist dies vollbracht, kann mit statistischen Standardverfahren ein Klassifizierer erzeugt werden, der – basierend auf einer Menge von Trainingsdaten – Gesten erkennt.

Die als Zahlenwerte (Merkmalsausprägungen) repräsentierten Merkmale der Trainingsbeispiele können in einer Matrix, der sog. Beobachtungsmatrix X, dargestellt werden. Jede Zeile u (Vektor) dieser Matrix entspricht einem Trainingsbeispiel und jede Spalte f enthält die Ausprägungen eines Merkmals bezüglich aller Traingsbeispiele. Eine Menge von Trainingsbeispielen für eine einzelne Geste wird Stichprobenklasse c für eine Gestenklasse genannt. Vorausgesetzt die Stichprobenklassen unterscheiden sich signifikant in ihren Merkmalsausprägungen, so können klassenspezifische Gewichtungsfaktoren für die Merkmale berechnet werden. Mit diesen kann für jede Stichprobenklasse eine Diskriminanzfunktion y_c mit

$$y_c = w_0^c + \sum_{i=1}^{F} x_i . w_i^c$$

angegeben werden, die für einen zu klassifizierenden Vektor x (eine Geste) einen Zahlenwert liefert. w_0 ist die Gewichtungskonstante und die w_i sind die F Gewichtungsfaktoren für die Merkmale einer Stichprobenklasse. Einzelheiten zur Berechnung der Gewichtungsfaktoren und mathematische Grundlagen der Diskriminanzanalyse findet der interessierte Leser in [RUBI91] und [KRZA88]. Die Gewichtung wird so gewählt, daß das y_c mit dem größten Funktionswert dem Vektor x (statistisch) am ähnlichsten ist, so daß gilt: $c = \max(y_1, y_2 ..., y_C)$. Da für jede Eingabe ein y_{max} existiert, muß entschieden werden, unter welchen Bedingungen eine Eingabe als nicht klassifizierbar abgelehnt wird. Bei dem hier vorgestellten Verfahren werden Gesten als nicht klassifizierbar bezeichnet, wenn entweder die Entfernung zwischen der Eingabe und der ihr zugeordneten Gestenklasse im Merkmalsraum (Distanz von Mahalanobis) einen Grenzwert überschreitet, oder die Entfernungen der Eingabe zu den y_i mit den beiden größten Werten sich nicht signifikant unterscheiden. Im zweiten Fall liegt eine Konfliktsituation vor, die vom Klassifizierer nicht entschieden werden kann.

Die Diskriminanzanalyse ist eine effiziente und zuverlässige Klassifizierungsmethode, wenn tatsächlich unterscheidbare Stichprobenklassen vorliegen. Außer der erwähnten Datenqualität, wirkt sich die Wahl der Features entscheidend auf den Erfolg des Verfahrens aus. Die zur Gestenerkennung verwendeten Merkmale werden im anschließenden Abschnitt vorgestellt.

4.3 Features für die Gestenerkennung

Zwischen den Freiheitsgraden der Hand bzw. des Datenhandschuhs und den Features für die Klassifizierung besteht ein enger Zusammenhang. Nur Freiheitsgrade, die in die Berechnung von Merkmalswerten eingehen, sind auch für die Gestenerkennung relevant. Beispielsweise ist die Rotation der Hand nur dann ein Unterscheidungsmerkmal, wenn die entsprechenden Meßwerte in einem Rotations-Merkmal verwendet werden. Ist dies nicht der Fall, könnten augenscheinlich verschiedene Handbewegungen von einem Klassifizierer nicht eindeutig identifiziert werden, wenn sie sich in allen anderen Merkmalen gleichen würden. Da der Aufwand des Klassifizierens unmittelbar von den Features abhängt, sollten diese möglichst effizient berechenbar und in ihrer Anzahl minimiert sein. Zudem müssen sie leicht reproduzierbar

sein, d.h. geringe Abweichungen im Bewegungsablauf dürfen keine großen Abweichungen in den Werten verursachen.

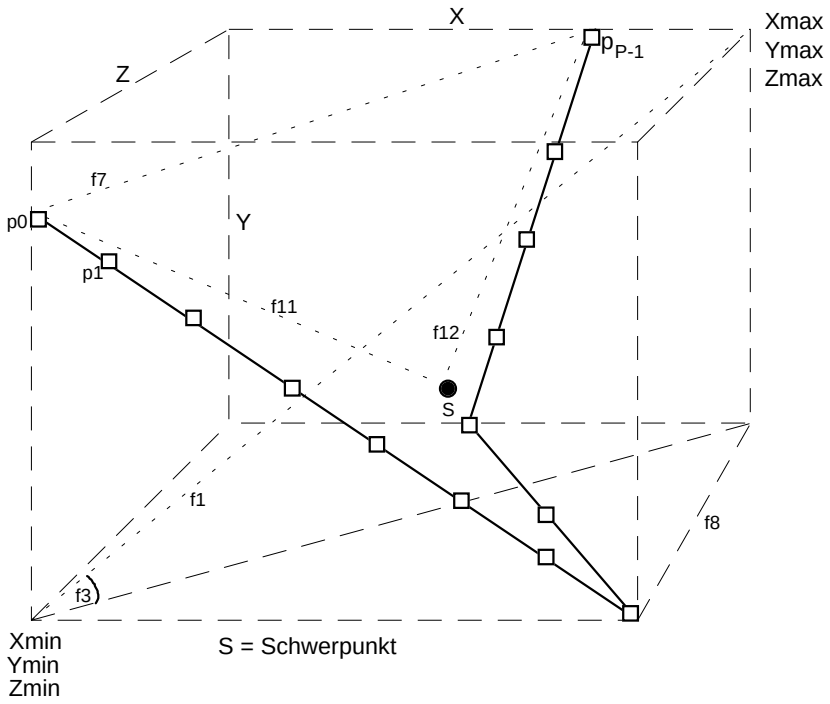

Abb. 4: Der Quader und der Pfad einer Geste

Ein Teil der im folgenden aufgezählten 15 Features wurde von Rubine, der einen Klassifizierer für zweidimensionale Mausbewegungen implementiert hat, übernommen bzw. angepaßt [RUBI91]. Durch die Bewegung der Hand wird im dreidimensionalen Raum ein virtueller Quader aufgespannt, der durch die minimalen und maximalen X-, Y-, und Z-Koordinaten spezifiziert ist. Einige Merkmale beziehen sich auf die räumliche Ausdehnung und auf den Weg, der durch die Bewegung zurückgelegt wird (s. Abb. 4). Da die Informationen des Datenhandschuhs in einem konstanten Zeittakt an den Rechner übertragen werden, läßt sich die mit der Hand zurückgelegte Wegstrecke (Pfad) als Polylinie darstellen, wobei ein Segment der Linie durch zwei Punkte begrenzt wird. Ein Pfad besitzt P Punkte, die von 0 bis P-1, der zeitlichen Reihenfolge entsprechend, numeriert sind. Um Verzögerungen im Bewegungsablauf herauszufiltern, werden sich überdeckende oder sehr dicht nebeneinander liegende Punkte, die zeitlich aufeinanderfolgen, ignoriert. Die Messung der Fingerkonfigurationen und der Rotation unterliegt nicht diesen Einschränkungen, so daß eine Geste H Werte für den Zustand der Finger und für die Rotation besitzt, wobei die Finger mit F_1, F_2, F_3 und F_4 bezeichnet werden[1] .

1) Die Länge der Raumdiagonalen zwischen dem minimalen und dem maximalen Punkt:

$$f_1 = \sqrt{(x_{max} - x_{min})^2 + (y_{max} - y_{min})^2 + (z_{max} - z_{min})^2}$$

Diese ist ein Maß für die räumliche Ausdehnung der Geste.

2) Die absolute Länge des Pfades der Geste:
Im Unterschied zu f_1 , enthält dieser Wert die Länge der zurückgelegten Wegstrecke. Auch Gesten mit einem kleinem Wert für f_1 , können einen großen Wert für dieses Merkmal erhalten, wenn die Bewegung sich auf einen kleinen Radius beschränkt.

[1] Der „kleine Finger“ wird beim PowerGlove nicht berücksichtigt.

$$\Delta x_p = x_p - x_{p-1}$$
$$\Delta y_p = y_p - y_{p-1}$$
$$\Delta z_p = z_p - z_{p-1}$$

$$f_2 = \sum_{p=1}^{P-1} \sqrt{\Delta x_p^2 + \Delta y_p^2 + \Delta z_p^2}$$

3) Der Winkel der Raumdiagonalen:

$$f_3 = \operatorname{acos}\left(\frac{(x_{min} x_{max}) + (y_{min} y_{max}) + (z_{min} z_{max})}{\sqrt{x_{min}^2 + y_{min}^2 + z_{min}^2} + \sqrt{x_{max}^2 + y_{max}^2 + z_{max}^2}} \right)$$

Mit diesem Merkmal wird die Höhe des Quaders (der Geste) ins Verhältnis zur Breite und zur Tiefe gesetzt. Ein spitzer Winkel weist auf eine „flache“ Geste mit Hauptbewegungsrichtung entlang der X- oder Z-Achse hin. Nähert sich der Wert der 90°-Marke, so ist dies ein Hinweis darauf, daß die Handbewegung hauptsächlich in Y-Richtung verlief. Werte um 45° ergeben sich bei einem ausgeglichenem Verhältnis.

4) Das Verhältnis der Bewegung in X-Richtung zur absoluten Weglänge:

$$\Delta x_p = x_p - x_{p-1}$$

$$f_4 = \frac{\sum_{p=1}^{P-1} \left| \Delta x_p \right|}{f_2}$$

Dieser Wert gibt an, wie groß der Anteil der Bewegung in X-Richtung an der gesamten Bewegung ist. Zusammen mit den Merkmalen (5) und (6), wird das Verhältnis der Bewegungsrichtungen zueinander berechnet.

5) Das Verhältnis der Bewegung in Y-Richtung zur absoluten Weglänge:

Berechnung wie (4), aber mit Y-Koordinaten.

6) Das Verhältnis der Bewegung in Z-Richtung zur absoluten Weglänge:

Berechnung wie (4), aber mit Z-Koordinaten.

7) Die Distanz zwischen dem ersten und letzten Punkt:

$$f_7 = \sqrt{(x_{P-1} - x_0)^2 + (y_{P-1} - y_0)^2 + (z_{P-1} - z_0)^2}$$

Die Distanz zwischen dem ersten und letzten Punkt gibt einen Hinweis darauf, ob zum Ende der Geste an den Ausgangspunkt zurückgekehrt wurde.

8) Die „Tiefe“ der Geste:

$$f_8 = \left| z_{max} - z_{min} \right|$$

Viele Bewegungsabläufe beschränken sich auf eine oder zwei Dimensionen des Raumes, d.h. eine bzw. zwei der Raumkoordinaten unterliegen nur geringen Schwankungen, während die anderen Koordinaten um so mehr variieren. In der Praxis muß häufig zwischen Gesten unterschieden werden, die in der Z_0-Ebene, also flächenartig vor dem Körper, ausgeführt werden, und denen, die die Tiefe des Raumes ausnutzen.

9) Die maximale lokale Geschwindigkeit (quadriert):

$$\Delta t_p = t_p - t_{p-1}$$

$$f_9 = \max_{p=1}^{P-1} \frac{\Delta x_p^2 + \Delta y_p^2 + \Delta z_p^2}{\Delta t_p^2}$$

Dieses Merkmal erhält den Wert für die maximale Geschwindigkeit zwischen zwei Punkten. Um die Unterschiede zwischen den Handbewegungen zu verstärken, wird die Geschwindigkeit quadriert.

10) Die Dauer der Geste:

$$f_{10} = t_{P-1} - t_0$$

ist die Differenz zwischen dem zeitlichen Ende und dem Anfang der Geste. Da Handbewegungen mit unterschiedlichen Geschwindigkeiten ausgeführt werden können, ist die Pfadlänge (f_2) nicht mit der Dauer gleichbedeutend.

11) Die Distanz zwischen dem ersten Punkt und dem „Schwerpunkt" der Geste:

$$x_s = \frac{\sum_{p=0}^{P-1} x_p}{P} \quad \text{wenn } P > 0, \text{sonst } 0$$

$$y_s = \frac{\sum_{p=0}^{P-1} y_p}{P} \quad \text{wenn } P > 0, \text{sonst } 0$$

$$z_s = \frac{\sum_{p=0}^{P-1} z_p}{P} \quad \text{wenn } P > 0, \text{sonst } 0$$

$$f_{11} = \sqrt{(x_0 - x_s)^2 + (y_0 - y_s)^2 + (z_0 - z_s)^2}$$

Der Schwerpunkt S der Geste gibt an, in welchem Bereich des Raumes die Hand durchschnittlich am meisten bewegt wurde. Da Gesten an beliebigen Stellen im Raum ausgeführt werden können, ergibt es keinen Sinn, absolute Koordinaten miteinander zu vergleichen. Daher wird der Schwerpunkt mit dem ersten und dem letzten Punkt (f_{12}) der Geste über die Entfernung zwischen den Punkten ins Verhältnis gesetzt.

12) Die Distanz zwischen dem letzten Punkt und dem „Schwerpunkt" der Geste:

$$f_{12} = \sqrt{(x_{P-1} - x_s)^2 + (y_{P-1} - y_s)^2 + (z_{P-1} - z_s)^2}$$

Beschreibung siehe (11).

13) Die Summe der Bewegungsänderungen der Finger:

$$\Delta F_h^1 = \left| F_h^1 - F_{h-1}^1 \right|$$
$$\Delta F_h^2 = \left| F_h^2 - F_{h-1}^2 \right|$$
$$\Delta F_h^3 = \left| F_h^3 - F_{h-1}^3 \right|$$
$$\Delta F_h^4 = \left| F_h^4 - F_{h-1}^4 \right|$$

$$f_{13} = \sum_{h=0}^{H-1} \sum_{i=1}^{4} \Delta F_h^i$$

Mit diesem Merkmal wird die Aktivität der Finger während der Ausführung einer Geste gemessen. Die Praxis hat gezeigt, daß es viel Konzentration erfordert, individuelle Finger präzise zu steuern, wenn die Hand bzw. der ganze Arm bewegt wird. Diese Erfahrung entspricht den Beobachtungen von Hauptmann, der festgestellt hat, daß die Finger häufig gemeinsam bewegt werden [HAUP93]. Zudem sind die Informationen über die Finger, die der Datenhandschuh liefert, oftmals von geringer Qualität. Um zu vermeiden, daß das Klassifizierungsverfahren aufgrund wenig aussagekräftiger oder sogar falscher Daten versagt, wird die Aktivität der Finger nur mit zwei Merkmalen berücksichtigt (f_{13} und f_{14}).

14) Die absoluten Fingerwerte im Verhältnis zur Anzahl der Konfigurationsmessungen:

$$f_{14} = \frac{\sum_{h=0}^{H-1}\sum_{i=1}^{4} F_h^i}{H} \text{ wenn } H > 0, \text{sonst } 0$$

Werden die Finger während des Bewegungsablaufes nicht bewegt, können Gesten bezüglich ihrer summierten Fingerkonfigurationen unterschieden werden.

15) Der durchschnittliche Drehwinkel der Hand:

$$f_{15} = \frac{\sum_{h=0}^{H-1} Rot_h}{H} \text{ wenn } H > 0, \text{sonst } 0$$

Der PowerGlove unterscheidet 12 Grade der Rotation um die Z-Achse mit einem ganzzahligen Wertebereich von [-5...6]. Da die Werte vorzeichenbehaftet sind, tendieren ausgeglichene Drehbewegungen gegen null.

Die aufgeführten Features sind 15 allgemeine Unterscheidungsmerkmale für Handbewegungen, wobei einige Merkmale die speziellen Eigenschaften des Datenhandschuhs berücksichtigen. Es sind Gesten denkbar, die mit einer Teilmenge der Features klassifiziert werden können, und andere, für die alternative Merkmale besser geeignet sind. Die Trainingsbeispiele einer jeden Gestik erzeugen jeweils eine eigene Verteilung der Features im n-dimensionalen Raum, so daß die Klassifizierung unter wechselnden Voraussetzungen erfolgt. Die trainierten Daten einer Gestik können weiteren Untersuchungen wie der Hauptkomponentenanalyse unterzogen werden, um die für eine spezielle Gestik relevantesten Merkmale zu finden [MARI77]. Eine weitere Möglichkeit zur Optimierung des Klassifizierers besteht in der Vergabe einer sog. a priori Wahrscheinlichkeit für jede Geste. Im konkreten Anwendungsfall ist nicht zu erwarten, daß alle Gesten mit der gleichen Häufigkeit bzw. Wahrscheinlichkeit gebraucht werden. Dieser Umstand kann bei nicht eindeutigen Klassifizierungen (ähnliche Werte für die y_i) genutzt werden, um dann die Geste mit der größten a priori Wahrscheinlichkeit zu wählen.

4.4 Evaluation des Klassifizierers

Mit dem Trainingswerkzeug wurde eine Testgestik, bestehend aus 8 Posen und 21 dynamischen Gesten, definiert und trainiert. Um die Klassifizierungsmethode zu Testen, wurden Gesten von einer Testperson eingegeben, die vom Klassifizierer zu erkennen waren. Es wurden Erkennungsraten von 85% für Posen und 88% für dynamische Gesten erreicht. Einzelne Gesten konnten im Versuch sogar zu 100% korrekt klassifiziert werden.

In dem Experiment wurde deutlich, daß der verwendete Datenhandschuh nur sehr bedingt als Eingabegerät für Gesten tauglich ist. Während des Betriebes kommt es immer wieder zu Aussetzern bei der Datenübertragung oder die Dehnungsmeßstreifen reagieren träge bzw. überhaupt nicht. Dennoch konnte prinzipiell nachgewiesen werden, daß der implementierte Klassifizierungsalgorithmus auch ohne gestikspezifische Optimierungen brauchbare Ergebnisse liefert.

5 Zusammenfassung und Ausblick

Mit dem hier vorgestellten System wurde eine Umgebung zum Definieren, Trainieren und Analysieren von Gestiken realisiert. Unter Verwendung konventioneller und kostengünstiger Ausstattung – Standard PC und low-cost Datenhandschuh – kann Gestik als neue Eingabeform genutzt werden. Der dafür entwickelte Klassifizierer ist kompakt, effizient sowie durch seine objekt-orientierte Schnittstelle erweiterbar und einfach zu integrieren.

Für den praktischen Einsatz sind die bisher erzielten Erkennungsraten noch zu steigern. Eine bessere Qualität des Datenhandschuhs und verfeinerte Optimierungen beim Klassifizierer sind die Faktoren, die höhere Erfolgsquoten beim Erkennen der Gesten erwarten lassen. Es ist auch denkbar, andere Verfahren aus dem Gebiet der Mustererkennung einzusetzen. In unsere weiteren Forschungstätigkeiten werden diese Erkenntnisse eingehen.

6 Literatur

[BAUD93] Baudel, T.: CHARADE - Remote Control of Objects using Free-Hand Gestures, Communications of the ACM, Vol.36, No.7, July 1993

[BORD93] Bordegoni, M. & Hemmje, M.: A Dynamic Gesture Language and Graphical Feedback for Interaction in a 3D User Interface, EUROGRAPHICS ´93, Vol. 12, No. 3

[BRAU94] Brauer, V.: Feature-basierte Erkennung dynamischer Gesten mit einem Datenhandschuh, Diplomarbeit im Fach Informatik, Universität Bremen, 1994

[BRAU96] Brauer, V. & Bruns, F.W.: Greifendes und begreifendes Modellieren im Realen und Virtuellen, in Neugebauer u. Wiesener (Hrsg.), 7. Workshop Hypermedia und KI, Report FR-1996-001, FORWISS, Erlangen, 1996

[BRUN93] Bruns, F. W.: Über die Rückgewinnung von Sinnlichkeit - Eine neue Form des Umgangs mit Rechnern Universität Bremen, Forschungszentrum Arbeit und Technik (artec) artec Arbeitspapier Nr. 19, Januar 1993

[CARR91] Carr, R.: The Point of the Pen, Byte, Feb. 1991, S. 211-221

[DARR95] Darrell, T & Pentland, A. P.: Attention-driven Expression and Gesture Analysis in an Interactive Environment, Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR), Zurich, 1995

[EKMA79] Ekman & Friesen in Scherer, R. (Hrsg.): Handbewegungen, Nonverbale Kommunikation: Forschungsberichte zum Interaktionsverhalten, Beltz Verlag, Weinheim, 1979

[FELS93] Fels, S. & Hinton, G.: Glove-Talk: A Neural Network Interface between a Data Glove and a Speech Synthesizer, IEEE Transactions on Neural Networks, Vol 4, No. 1, January 1993

[FELS95] Fels, S. & Hinton, G.: Glove-Talk II: An Adaptive Gesture-to-Formant Interface, ACM Proceedings of the CHI 1995

[HARL93] Harling, P.A.: Gesture Input using Neural Networks, Project Report, University of York, UK, Department of Computer Science, March 1993

[HAUP93] Hauptmann, Alexander & McAvinney, Paul: Gestures with Speech for Graphic Manipulation, International Journal of Man Machine Studies, Vol. 38, No. 2, (1993), S. 231-249

[HOMM94] Hommel, Hofmann & Henz: The TU-Berlin High-Precision Sensorglove, Fourth International Scientific Conference on Work With Display Units (WWDU), Milan, Oct. 1994, Book of Short Papers Vol. 2

[KIM88] Kim, J.: On-line Gesture Recognition by Feature Analysis, Proceedings of Vision Interface, 1988, S. 51-55

[KRZA88] Krzanowski, W. J.: Principles of Multivariate Analysis, Oxford University Press, Oxford, 1988

[MARI77] Marinell, G.: Multivariate Verfahren, Oldenbourg Verlag, München, 1977

[ROBE94] Roberts, Geoff D.: Statistical Pattern Classification in Computer Recognition of Sign Language (MS Thesis) R. Roberts Advanced Computer Graphics Centre, RMIT, Melbourne, Australia, 1994

[RUBI91] Rubine, Dean: The Automatic Recognition of Gestures, Ph.D. Thesis, Carnegie-Mellon University, 1991

[STAR95] Starner T. & Pentland A.: Visual Recognition of American Sign Language Using Hidden Markov Models, Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR), Zurich, 1995

[STUR92] Sturman, D.: Whole-Hand Input, Ph. D. Thesis, Media Arts and Sciences, MIT 1992

[STUR94] Sturman, D. & Zeltzer, D.: A Survey of Glove-Based Input, IEEE Computer Graphics & Applications, Vol. 14, No. 1, January 1994

[TAKA91] Takahashi, T. & Kashino, F.: Hand Gesture Coding Based on Experiments using a Hand Gesture Interface Device SIGCHI Bulletin, Vol. 23, No. 2, April 1991, S. 67-74

[WEXE94] Wexelblatt, A.: A Feature-Based Approach to Continous-Gesture Analysis, MS Thesis, Massachusetts Institute of Technology (MIT), June, 1994

[WILS95] Wilson, A. D. & Bobick, A. F.: Configuration States for the Representation and Recognition of Gesture, Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR), Zurich, 1995