

Adaptive Gestenerkennung mit Variationsabschätzung für interaktive Systeme

Maxim Boianetchii¹ und Marian Stein²

¹ Universität Rostock

`maxim.boianetchii@uni-rostock.de`

² Universität Rostock

`marian.stein@uni-rostock.de`

1 Einführung

Dieser Artikel stellt eine Gestenerkennung / Adaptionssystem für die Mensch-Computer-Interaktion Anwendungen. Er beschreibt eine vorlagenbasierte Erkennungsverfahren, die durch ein Anpassungsprozess, die Geste Variation in Echtzeit verfolgen. Die Wichtigsten Vorteile sind geschätzten Parameter und Erkennungsergebnisse. Die Technik wurde auf verschiedene Weise in einer Benutzerstudie mit 3D freien Raum Gesten ausgewertet. Wobei wurde festgestellt, dass die Verfahren robust gegen Rauschen ist und erfolgreich passt sich an Variationsparameter. Dieses Verfahren führt die Anerkennung genauso gut oder sogar besser als andere Methoden.

Die Gesten werden in Mensch-Computer Interaktion mit mehreren und verschiedenen Formen der Aktivitätserkennung verwendet. Es besteht ein Bedarf für aufwendige Interaktionsparadigmen, die sich auf Körperbewegungen oder greifbare Geräte und Schnittstellen basieren. Es gibt bereits Verfahren zur Gestenerkennung und kontinuierliche Gestenerkennung, die erfolgreich umgesetzt wurden. Diese Verfahren wurden entwickelt um die Geste in den meisten Fällen zu markieren. In diesem Experiment wird ganz neue Technik vorgeschlagen, die über Klassifikation durchgeht, die komplementär Geste markiert, die eine Bewegung kennzeichnet und bietet Möglichkeit, die innovative Interaktion Szenarien zu nutzen. Die Benutzer müssen ständig ihre Bewegungen anzupassen und sich zum Beispiel auf optische oder akustische Rückmeldung verlassen. Die Geste die erkannt werden, können in Vergleich zu Geste Referenzen „verzerrt“ angezeigt werden. Auf solche Änderungen anzupassen, muss ein robustes Erkennungssystem in der Lage sein. Außerdem wäre es nützlich die Parameter der Geste Ausführung während der Leistung schrittweise zu schätzen. Weil das ermöglicht dem System die Berücksichtigung von Schwankungen die während der Bewegung und Aktualisierung der Bewegungsmodell auftreten, wahrzunehmen. Außerdem solche Parametrisierung könnte direkt in die Gestaltung der Interaktion verwendet werden. Wir schlagen ein neues Gestenerkennungssystem, die in der Lage ist in Echtzeit die berücksichtigen Bewegung und Variation inkrementell während der Durchführung wahrzunehmen und dem Nutzer Parameter zurückliefern. Es können verschiedene Gestenvariationen innerhalb der Geschwindigkeit, Amplitude oder Orientierung aufgenommen werden. Wichtig ist, dass diese Variationen

können kontinuierlich während der Ausführung der Gesten geschätzt werden. Diese Methode ist ein Zustandsraummodell, in dem Variationen online geschätzt werden. Dazu braucht man ein Partikelfilter um kontinuierlich auf die Variation der Geste anzupassen.

2 Ähnliche Arbeiten

In diesem Abschnitt untersuchen wir Methoden die am meisten benutzt werden um Gesten, die als mehrdimensionale Zeitreihen für die Mensch-Computer-Interaktion vorgestellt sind, zu erkennen. Die mehrdimensionalen Zeitreihen stellen die Bahn eines Punktes auf einer Oberfläche oder in dem 3D-Raum. Rubine [1991] schlägt eine geometrische Abstandmaß basierend auf Beispiele von Einzeltakt-Gesten. Wobbrock et al. [2007] schlagen eine einfache Template-basierte Methode, die Nutzung der euklidische Abstands nach einer vor Verarbeitungsstufe zur Berücksichtigung geometrischer Variationen (wie Skalierung und Rotation) und Drehzahlschwankungen ermöglicht. [Gawrila und Davis, 1995; Liu et al. 2009] schlagen vor Dynamic Time Warping (DTW), das erfordert die Speicherung des gesamten zeitliche Struktur von Geste. Es gibt verschiedene Anwendungen, wie Gestensteuerung [Merrill und Paradiso 2005] kommunikative Krakelfolgen [Heloir et al. 2006], und Abfragen auf Basis menschlichen Bewegung [Forbes und Fiume 2005]. Statistische Methoden, wie Hidden Markov Model (HMM) [Rabiner 1989], die auf einer Wahrscheinlichkeitsinterpretation basiert der Beobachtungen (Geste Proben) und können zeitliche Trajektorie der Geste durch eine kompakte Darstellung zu modellieren. Variationen in der Geste werden so gehandhabt, zum größten Teil durch Methoden wie HMM dass umfassende Datenbanken verwenden, unter Berücksichtigung aller möglichen Variationen. In diesem Artikel verweisen wir auf Variationen innerhalb Geste Klassen auftreten. Wilson und Bobick [1999] schlagen ein Modell, das berücksichtigt parametrische Änderungen, die in der Ausführung führt. Wir werden in diesem Artikel, dass ein adaptiver Ansatz mit einem ausgefahrenen Zustand Modell zeigen und eine andere Decodierschema möglich ist Die Methode, die wir vorschlagen, durch die Arbeit von Black und Jepson [1998a] inspiriert, basierend auf die Kondensation Algorithmus [Isard und Blake 1998] für die Anerkennung der raumzeitlichen Geste Vorlagen. Die Umsetzung für die Verfolgung von Geschwindigkeit und Skalierungsänderung erlaubt. Wir verallgemeinern den Ansatz von Black und Jepson [1998a] durch Schätzen nicht nur die Skalierung, sondern auch andere Parameter, wie Rotation.

3 Interaktionsprinzipien

Wir präsentieren Interaktionsprinzipien, die wir als wichtig für Zielanwendungen wie Ton Manipulation oder Visual Processing in Kontexten wie Spiele, interaktive Kunst oder Rehabilitation. Unsere Interaktionsmodell beinhaltet zwei Arten von Kontrolle: die Ausführung von Geste und wie die durchgeführt werden. Die

Körperbewegungen werden von einem Bewegungserfassungsgerät erfasst. Der Algorithmus ist in der Lage die Geste ausgeführt zuerkennen, an ihre Variation und Veränderungen anzupassen und der Variationsparameter mit Berücksichtigung der zuvor gelernten Vorlagen zurückliefern. Die Erkennung und Adaption werden in der Echtzeit durchgeführt. Der Algorithmus Ausgänge werden fortlaufend aktualisiert. Die schematische Darstellung des Interaktionsmodells ist in Benutzers Körperbewegungen gesehen und wird durch ein Sensor-System erfasst. Die Modelausgang hat zwei Komponenten: einen Index für die anerkannten Gesten und ein Vektor für kontinuierliche Werte der Schätzung der Variation. Zur Erleichterung der schnellen Testsitzungen und um das Lernverfahren so leicht wie möglich zu machen, wird von System nur eine einzige Schablone zum Definieren einer bestimmten Geste Klasse erforderlich.

4 Zugrunde liegendes Modell

Wir definieren Gesten so, dass eine Geste ist eine Körpergliedbewegung die durch eine zeitliche Reihe aus feste Anzahl von Parametern vertreten wird. So wird bei der Eingabe von Geste die Erkennungsaufgabe gestellt, die beste Übereinstimmung mit vorher aufgezeichneten Vorlagen von Gesten auszuwählen.

4.1 Kontinuierliche Zustandsmodell

Das Modell kann mit folgende dynamische System formuliert werden:

$$\begin{cases} X_k = f_{TR}(X_{k-1}, V_{k-1}) \\ Z_k = f_{OB}(X_k, W_k; g) \end{cases} \quad (1)$$

wobei k diskreter Zeitpunkt ist. $-x_k$ ist ein Vektor der den Systemzustand und Zustand der Elemente unterschiedlichen Geste Eigenschaften vorstellt - F_{TR} ist eine Funktion, die die Entwicklung des Systemzustands regelt - $-v_k$ Rauschsequenz - F_{OB} ist eine Funktion, die die Beobachtungen z_k erzeugt, je über den Systemzustand x_k und Messrauschsequenz w_k und eine Vorlage Geste g . Das Problem wird als ein Verfolgungsproblem formuliert

4.2 Zustandsraummodell

Der Zustand des Systems besteht aus unterschiedliche Merkmalen und Eigenschaften von Gesten, die über eine lange Zeit geschätzt werden. Der Zustandsraum umfasst die Merkmale, die online bewertet können und folglich als Dauerleistungen während der Interaktion verwendet werden. Die Funktionen werden in jedem Zeitschritt aktualisiert.

4.3 Zustandsübergang

In diesem Modell, ist die Zustandsübergangsfunktion F_{TR} linear und durch Matrix A gegeben und als Gauß-Verteilung modelliert:

$$\begin{aligned}
p(X_k|X_{k-1}) &= N(X_k|A_{X_{k-1}}, \Sigma) \\
\Sigma &= \text{diag}(\phi_1 \dots \phi_D)
\end{aligned} \tag{2}$$

Durch Einstellen der Beziehung zwischen der Phase und der Geschwindigkeit wählen wir Einschränkung, entsprechend einer ersten Ordnung Bewegungsgleichung:

$$p_k = p_{k-1} + V_k/T + N(0, \phi_1) \tag{3}$$

wobei T Länge der Vorlage und ϕ_1 das erste Element in der Diagonale der Σ . Diese Einschränkung kann durch Einstellen der ersten Zeile der Matrix A in $(1 \frac{1}{T} 0 \dots 0)$ berücksichtigt werden. Die andere Bedingungen werden in der ersten Zeile auf Null gesetzt. Die Übergangsparameter spielen eine ganz wichtige Rolle für die Anpassung. Sie regeln die Dynamik der Veränderungsschätzung: die Geschwindigkeit der Annäherung an die genaue Schätzung und die Präzision der Schätzung.

4.4 Beobachtungsfunktion

Die Beobachtungsfunktion wertet die Genauigkeit der Zustandsschätzung gemäß Eingangsbeobachtung und der Vorlage. Das Verfahren wird wie Diskriminante von der Parameter der Beobachtungsfunktion gesteuert.

$$S_t(Z_k|f(X_k, g(p_k)), \Sigma, v) = C(\Sigma, v) \left(1 + \frac{d^2(Z_k, f(X_k, g))}{v}\right)^{-\frac{v+K}{2}} \tag{4}$$

$$\text{wo: } C(\Sigma, v) = \frac{\Gamma(v/2 + K/2)}{\Gamma(v/2)} \frac{|\Sigma|^{-1/2}}{(v\pi)^{K/2}}$$

wobei $f(X_k, g)$ eine Funktion des Templates g und den Zustandswert bei k ist. $f(X_k, g)$ passt sich die erwartete Vorlage Probe g (Stück), da die Phase p_k bei k gegeben ist. Der Abstand d zwischen dem angepassten Vorlageprobe und dem eingehenden Beobachtung ist gegeben durch:

$$d(Z_k, f(X_k, g)) = \sqrt{[Z_k - f(X_k, g)]^T \Sigma^{-1} [Z_k - f(X_k, g)]}. \tag{5}$$

4.5 Inferenz und Implementation

Inferenz ist die Echtzeitschätzung der Zustandswerte. Hier wird die Zustandsprobe aus einem einfacheren Verteilung gezeichnet und dann entsprechend ihrer Bedeutung bei der Schätzung der "true" Verteilung gewichtet. Bei jedem Schritt k stellt ein Teilchen x_k^i einen möglichen Wert des Zustandsraums, der von seiner Wahrscheinlichkeit w_k^i gewichtet wird. Der Erwartungswert der Merkmale, zu dem Zeitpunkt k ist:

$$X_k = \sum_{i=1}^{N_s} w_k^i X_k^i$$

wobei N_s bezeichnet die Anzahl der Teilchen. Abgeleitete Merkmals-werte x_k bilden den Anpassungsprozess, seit dem Zeitpunkt k . Wir beurteilen die Variationswerten definieren sie als Zustands variablen.

4.6 Bearbeitung der Anerkennung

Schließlich wird das Modell erweitert, um Anerkennung zu behandeln, indem wir mehrere Schablonen in Betracht ziehen. Der Zustandsraum wird geändert um die wahrscheinlichste Schablone zusätzlich zu den variierenden Gersteneigenschaften zu schätzen. Betrachten M Vorlagen jeweilige Länge $L_1 \dots L_M$ bezeichnet durch $g^1 \dots g^M$. Bei der Initialisierung, ordnen wir jedem Zustand Partikel x_k^i eine Geste mit Index zwischen $1 \dots M$ (m_k bezeichnet), auf der Grundlage einer Anfangsverteilung. Im allgemeinen wird eine gleichmäßige Verteilung gewählt. Dies erweitert die Zustandkonfiguration angewand auf jede Partikel wie folgt:

$$X_k^i = \left\{ \begin{array}{c} X_k^i(1) \\ \vdots \\ X_k^i(D) \\ m_k \end{array} \right\} \in \mathbb{R}^D \times \mathbb{N} \quad (6)$$

Die Übergangswahrscheinlichkeit wird dann wie folgt angepasst:

$$\begin{aligned} p(X_k^i | X_{k-1}^i) &= N(X_k^i - k | A(X_{k-1}^i), \Sigma) \\ \Sigma &= \text{diag}(\phi_1 \dots \phi_D 0). \end{aligned} \quad (7)$$

Durch Summieren der Gewichte W_k^i entsprechend der Partikelgeste Indizes, ist es einfach, die Wahrscheinlichkeit jedes Geste zu berechnen:

$$p(g_k^i | g_k^m) = \sum_{j \in J} w_k^j, \forall l \in [1, M], \forall m \in [1, M], m \neq l \text{ wo:}$$

$$J = j \in [1, N_s] / X_k^j(D+1) = l. \quad (8)$$

5 Erkennung von realen 2D-Gesten

Um die Gestenauswertung zu evaluieren, wurde die Gestendatenbank von Wobbrock et al.[1] verwendet. Diese enthält Daten von 16 Stiftgesten, die von zehn Teilnehmern in drei verschiedenen Geschwindigkeiten jeweils zehn mal aufgenommen wurden. Für die Versuche wurden pro Geste zufällig aus den Daten eines Teilnehmers jeweils ein Trainings- und ein anderer Testdatensatz ausgewählt. Insgesamt wurden 4 verschiedene Tests durchgeführt, die jeweils 100 Mal wiederholt wurden:

1. Gleiche Testbedingungen, wie sie von Wobbrock[1] verwendet wurden.

2. Einfluss von geänderten Verteilungsparametern.
3. Verwendung von Trainings- und Testdaten mit unterschiedlichen Geschwindigkeiten.
4. Genauigkeit der Erkennung von Gesten, während sie noch nicht abgeschlossen sind.

Bei allen Tests außer 3. wurden Trainings- und Testdaten mit der gleichen Geschwindigkeit gewählt. Da GVF auch Veränderungen in den Gesten erkennen und ausgeben soll, ergibt sich für diese Tests ein Zustandsraum x_k , der aus der Phase p_k , der Geschwindigkeit v_k , der Skalierung s_k und dem Drehwinkel α_k besteht.

5.1 Erkennung von Beispielen gleicher Geschwindigkeit

Im ersten Test wurde die durchschnittliche Erkennungsrate sowie die Standardabweichung von GVF mit den erreichten Werten von \$1 Recognizer[1], DTW und GF verglichen. Im Vergleich zu GF erreicht GVF eine bessere Erkennungsrate (98,11% gegenüber 95,78%), was sich darauf zurückführen lässt, dass GVF sich an Skalierung und Drehung anpasst. Auch im Vergleich mit den beiden Offline-Methoden erreicht GVF leicht bessere Erkennungsraten (98,11% gegenüber 97,27% bzw. 97,86%).

5.2 Einfluss der Verteilungsparameter auf die Erkennung

Im zweiten Test wurde der Einfluss der Standardabweichung σ und der Parameter v der Student'schen t-Verteilung auf die Erkennungsrate untersucht. Untersucht wurden dabei für σ in Zehnerschritten Werte von 10 bis 150 und für v die Werte 0.5, 1, 1.5 und ∞ (hier entspricht die Verteilung einer Gaussverteilung).

Die Ergebnisse sind in Graph X zusammen mit den Erkennungsraten der beiden Offline-Methoden dargestellt.

Hier sind zwei Beobachtungen festzustellen. Erstens, dass die Erkennungsrate die Beste ist für beschränkte σ und v . Zweitens, dass die Verwendung einer Student'schen t-Verteilung anstatt einer Gaussverteilung den Vorteil hat, dass die Erkennungsrate wesentlich weniger von σ abhängt.

5.3 Erkennung von Beispielen verschiedener Geschwindigkeit

Im dritten Test wurden die Erkennungsraten von GVF mit denen des \$1 Recognizers verglichen. Diesmal wurden allerdings die Vorlage und die Testgeste aus verschiedenen Geschwindigkeitsgruppen verwendet. Tabelle Y zeigt die Erkennungsraten der unterschiedlichen Kombinationsmöglichkeiten. Es wird deutlich, dass insgesamt die durchschnittliche Erkennungsrate beider Algorithmen vergleichbar ist (94,6% gegenüber 94,8%). Ebenso fällt auf, dass die schlechtesten Ergebnisse erreicht werden, wenn Vorlage und Testgeste entgegengesetzte Geschwindigkeiten haben, mit Genauigkeiten von 88,3% bzw. 85,9%.

5.4 Früherkennung der Gesten

Zuletzt wurde die Erkennung der Gesten untersucht, während sie ausgeführt werden. In Graph Z sind die Früherkennungsraten von GVF und GF in Abhängigkeit vom Fortschritt der Geste dargestellt. Es wird deutlich, dass GVF über den ganzen Messbereich bessere Erkennungsraten erreicht als GF. So erzielt GVF im Schnitt bereits nach 10% der Geste eine Genauigkeit 67% und erreicht bereits nach 40% Fortschritt 90% Genauigkeit.

5.5 Ergebnisse der Versuche mit realen Daten

In den Versuchen hat sich gezeigt, dass GVF in der Erkennung von Gesten gleich gut bis besser ist als andere aktuelle Erkennungsmethoden. Darüber hinaus wurde deutlich, dass die im GVF verwendete Student'sche t-Verteilung gegenüber der im GF verwendeten Gaussverteilung weniger auf eine gute Abschätzung von σ angewiesen ist. Somit ist diese Methode besser in Anwendungsfällen verwendbar, in denen wenig Trainingsdaten zur Verfügung stehen, und somit σ schlecht abgeschätzt werden kann. Eine Verbesserung gegenüber bisher üblichen Methoden stellt der GVF auch dadurch dar, dass bereits während der Gestenausführung Ergebnisse geliefert werden, und die Früherkennung gut genug funktioniert, dass schon nach 10% der Geste Erkennungsgenauigkeiten von über 60% erreicht werden.

6 Abschätzung der Varianz anhand synthetischer Daten

Um die Anpassung an Varianz der Gesten auszuwerten, wurden synthetische Daten verwendet, da menschliche Eingaben keine exakten Werte für die Varianzparameter liefern würden, mit denen die Ergebnisse des GVF verglichen werden könnten. Es wurden zwei Fälle betrachtet: Im ersten Fall wurde nur die Phase und Skalierung der Gesten variiert, im zweiten Fall zusätzlich auch die Rotation. Für die Generierung der Daten wurde eine Viviani-Kurve verwendet:

$$C(t) = \begin{cases} x(t) = a(1 + \cos(t)) \\ y(t) = a \sin(t) \\ z(t) = 2a \sin(t/2) \end{cases} \quad (9)$$

6.1 Auswertung der Phasenabschätzung

Für diesen Fall wurde als Vorlage eine lineare Abtastung der Viviani-Kurve, und als Testdaten eine kubische Abtastung($t \rightarrow t^3$) zu der ein gleichmäßig verteiltes Rauschen hinzugefügt wurde, verwendet. Der Zustandsraum ist hier dreidimensional, bestehend aus der Phase $p_k \in [0, 1]$, der Geschwindigkeit $v_k \in \mathbb{R}$ und der Skalierung $s_k \in \mathbb{R}$, wobei v_k und s_k so normalisiert sind, dass ein Wert von 1 jeweils der Geschwindigkeit bzw. Skalierung der Vorlage entspricht. Um einen Vergleich zu GF machen zu können, wurde $v \rightarrow \infty$ gewählt.

Im Vergleich erreichten beide Verfahren gute Abschätzungen mit durchschnittlichen Fehlern von 1,3 Abtastungen beim GVF, bzw. 2,3 Abtastungen beim GF.

Weiter wurde der Einfluss von σ auf die Abschätzungen untersucht. Die Ergebnisse sind in Graph Z dargestellt. Für alle getesteten Werte für σ lieferte GVF bessere Abschätzungen als GF, obwohl GF eine genauere Inferenztechnik verwendet. Dies liegt daran, dass GVF ein besseres kontinuierliches Modell verwendet, um die Daten wiederzuspiegeln, sowie an dem Zusammenhang zwischen Geschwindigkeit und Phase, welcher von GF ignoriert wird.

Ebenso stellte sich heraus, dass GVF bessere Abschätzungen in Abhängigkeit von der Stärke des Rauschens liefert, als GF.

6.2 Auswertung der Rotationsabschätzung

Für diesen Versuch wurde die Rotation der Geste mit der Zeit variiert. Die Drehwinkel ϕ , θ und ψ für die Drehung entlang der x , y und z -Achse wurden pro Zeitschritt berechnet durch:

$$\begin{aligned}\phi(t) &= t^2 \\ \theta(t) &= t \\ \psi(t) &= -t^{1/3}\end{aligned}\tag{10}$$

Der Zustandsraum ist hier entsprechend 5-dimensional, bestehend aus p_k , v_k , ϕ_k , θ_k und ψ_k . Es wurden die gleichen Tests durchgeführt wie bei der Phasenabschätzung. Es stellt sich heraus, dass σ wenig Einfluss auf die Genauigkeit der Abschätzung hat, aber dass bei niedrigen Werten für σ die Fehlerrate stärker schwankt als bei großen Werten.

6.3 Ergebnisse der Auswertung synthetischer Daten

In beiden Experimenten wurde deutlich, dass σ wenig Einfluss auf die Abschätzung hat. Daher ist der Algorithmus auch anwendbar, wenn es nur sehr wenig Trainingsdaten gibt. Ebenso stellte sich heraus, dass die Phasenabschätzung bei festem Rauschen und σ mit einem Durchschnittsfehler von 2,3 Abtastungen sehr genau ist. Zuletzt hat die Stärke des Rauschens einen zu erwartenden Einfluss auf die Abschätzung, wobei sie noch gut genug bleibt, dass der Algorithmus auch bei signifikantem Rauschen verwendet werden kann.

7 Nutzerstudie

7.1 Durchgeführte Experimente

Es wurde eine Nutzerstudie durchgeführt, um die Einbindung von GVF in reale Anwendungen zu untersuchen. Hierfür wurde eine Anwendung entwickelt, wo eine Gestenausführung die Wiedergabe eines vordefinierten Tons hervorruft, welcher durch Variationen in der Geste verändert wurde. Hierbei wird die Früherkennung des GVF verwendet, um den Ton abzuspielen, sobald die Wahrscheinlichkeit der Erkennung über 50% beträgt. Ab dann werden für den Rest der

Geste die Abschätzung der Varianzen für die Manipulation des Tons verwendet. Hierbei sorgt eine schnellere/langsamere Ausführung der Geste für eine schnellere/langsamere Wiedergabe des Tons, größere/kleinere Gesten sorgen für eine lautere/leisere Wiedergabe und eine Drehung der Geste regelt die Cut-Off-Frequenz eines Hochpassfilters. Den Teilnehmern werden 7 Aufgaben gestellt, die alle daraus bestehen, einen bestimmten Ton abzuspielen, und ihn dann zu manipulieren. Hierbei ist zu beachten, dass Aufgabe 1 nur das Abspielen eines Tons, Aufgabe 2-4 eine globale Änderung eines Aspekts, Aufgabe 5 eine kontinuierliche Änderung eines Aspekts und Aufgabe 6 und 7 die globale Änderung zweier Aspekte beinhalten. Die Gesten werden von einem infrarotbasiertem System aufgenommen, welches Fingerbewegungen im freien Raum misst. Die Studie wurde mit 10 Teilnehmern durchgeführt, die jede der 3 Gesten 3 Mal pro Aufgabe durchführten. Somit ergeben sich 630 Datensätze.

Literatur

1. O. Wobbrock, A. D. Wilson, and Y. Li. Gestures without libraries, toolkits or training: A \$1 recognizer for user interface prototypes. *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, 2007.