Kapitel 6

Neuartige statistische Modellierung für die Klassifikation von Bildsequenzen

Dieses Kapitel beschreibt die Bildsequenzerkennung mit neuartigen pseudo dreidimensionalen Hidden-Markov-Modellen. Anders als im vorhergehenden Kapitel 5 soll hier die Bildfolge in ihrer Gesamtheit analysiert und schließlich klassifiziert werden. Der hierarchische Ansatz des Kapitels 5, der P2DHMMs in Kombination mit einem Kalman-Filter verwendet, dient hingegen lediglich dem Tracking von Personen und nicht der Klassifikation der Bewegungen. Genau dies, nämlich die Klassifikation von menschlichen Aktionen, bzw. Gesten stellt das Anwendungsszenario für die in diesem Kapitel beschriebenen pseudo dreidimensionalen Modelle dar.

Auf Markov-Random-Fields mit einer Nachbarschaftsbeziehung, die die drei Dimensionen einer Bildsequenz (x,y,t) berücksichtigt, sowie auf daraus abgeleiteten *echten* dreidimensionalen Hidden-Markov-Modellen wird in diesem Kapitel nicht ausführlich eingegangen. Die Theorie der MRFs ist allgemein formuliert und schließt den dreidimensionalen Fall ein (siehe [Li95]). Durch die Einführung einer kausalen Nachbarschaftsbeziehung und eines zweiten statistischen Prozesses, der zur Ausgabe von Merkmalen führt, geht das dreidimensionale Hidden-Markov-Modell aus dem Markov-Random-Field hervor (vgl. Kapitel 4.1 und 4.2). Beim dreidimensionalen HMM werden somit Zustandsübergänge mit folgender Definition verwendet:

$$A_{ijk,lmn,opq,uvw} = P(q_{(x,y,t)} = S_{(u,v,w)} | q_{(x-1,y,t)} = S_{(i,j,k)}, q_{(x,y-1,t)} = S_{(l,m,n)}, q_{(x,y,t-1)} = S_{(o,p,q)}) \tag{6.1}$$

Obwohl mit den dreidimensionalen Hidden-Markov-Modellen eine theoretisch sehr geeignete Modellierungsmethode zur Verfügung steht, ist es wiederum sehr problematisch, daß keine effizienten Trainings- und Klassifizierungsalgorithmen bekannt sind. Somit wird wie im zweidimensionalen Fall auf eine Modellierung, die Musterverzerrungen in allen Dimensionen gemeinsam betrachtet, verzichtet (siehe auch Kapitel 4.2). Dies führt zu der Einführung der *pseudo* dreidimensionalen Modelle. Durch den Verzicht auf die in Gleichung 6.1

vorgestellten Modellgrößen können für diese Modelle effiziente Algorithmen für das Training und die Erkennung gefunden werden. Die P3DHMMs werden im folgenden ausführlich vorgestellt.

6.1 Pseudo dreidimensionale Hidden-Markov-Modelle

Pseudo dreidimensionale Hidden-Markov-Modelle wurden im Rahmen dieser Arbeit entwickelt und sind erstmalig in den Arbeiten [Mul99c] und [Mul00a] erwähnt und verwendet worden. Die Bezeichnung *pseudo dreidimensionales Hidden-Markov-Modell* wurde vom Autor gewählt, da es sich um einen leicht zu merkenden Begriff handelt und zudem die methodische Verwandtschaft zu dem pseudo zweidimensionalen HMM betont wird. Es sei an dieser Stelle jedoch darauf hingewiesen, daß der Begriff P3DHMM die Existenz ähnlicher statistischer Vereinfachungen impliziert, wie dies bei P2DHMMs der Fall ist. Dies trifft jedoch nicht zu, denn die bei den P3DHMMs gemachten Vereinfachungen sind schwerwiegender als bei den P2DMMs. Bei den P3DHMMs wird neben dem Verzicht auf eine Modellierung, die Musterverzerrungen in beiden Bilddimensionen gemeinsam betrachtet, auch eine Unabhängigkeit zeitlich benachbarter Bildpunkte angenommen. Die statistische Modellierung besser beschreibende Bezeichnungen wären somit *pseudo pseudo dreidimensionales HMM* oder *doppelt pseudo dreidimensionales HMM*. Es ist offensichtlich, daß diese Bezeichnungen aufgrund ihrer Länge ungeeignet sind.

6.1.1 Modelldefinition

Pseudo dreidimensionale HMMs modellieren die Abhängigkeiten von Merkmalen einer Bildsequenz durch einen dreistufigen, hierarchischen Prozeß. Der in dieser Hierarchie am höchsten stehende Prozeß modelliert die Abhängigkeiten von aufeinanderfolgenden Bildern mit einem Markov-Modell erster Ordnung. Die Bilder selbst werden mit pseudo zweidimensionalen HMMs modelliert (siehe auch Kap. 4.3), die in das übergeordnete Modell eingebunden sind. Abb. 6.1 zeigt die Darstellung eines P3DHMMs, die das Modell als dreistufigen statistischen Automaten interpretiert. Dargestellt sind drei mit $S_1,\ldots,S_3$ bezeichnete übergeordnete Zustände, sowie den übergeordneten Zuständen zugeordnete pseudo zweidimensionale Modelle. Die P2DHMMs in Abb. 6.1 bestehen jeweils aus drei Metazuständen (z.B. $S_1^1,\ldots,S_3^1$ für das S_1 zugeordnete Modell), die wiederum aus jeweils vier Zuständen bestehen (z.B. $S_{\,1}^{\,1\,1},\ldots,S_{\,4}^{\,1\,1}$ für den Metazustand S_1^1). Die Zustände $S_1,\ldots,S_3$ werden im folgenden als *Hyperzustände* des P3DHMMs bezeichnet. Ein dreidimensionales Muster O_{XYT}, das in Form einer $X \times Y \times T$ Matrix vorliegt, kann auf folgende Weise modelliert werden: Jedes Einzelbild des Musters $(o_{xyt}, t = \text{const.})$ wird einem Hyperzustand zugeordnet. Dies ermöglicht eine nichtlineare Musterverzerrung in der Zeitdimension. Darüber hinaus werden die Einzelbilder selbst von pseudo zweidimensionalen Hidden-Markov-Modellen modelliert, was

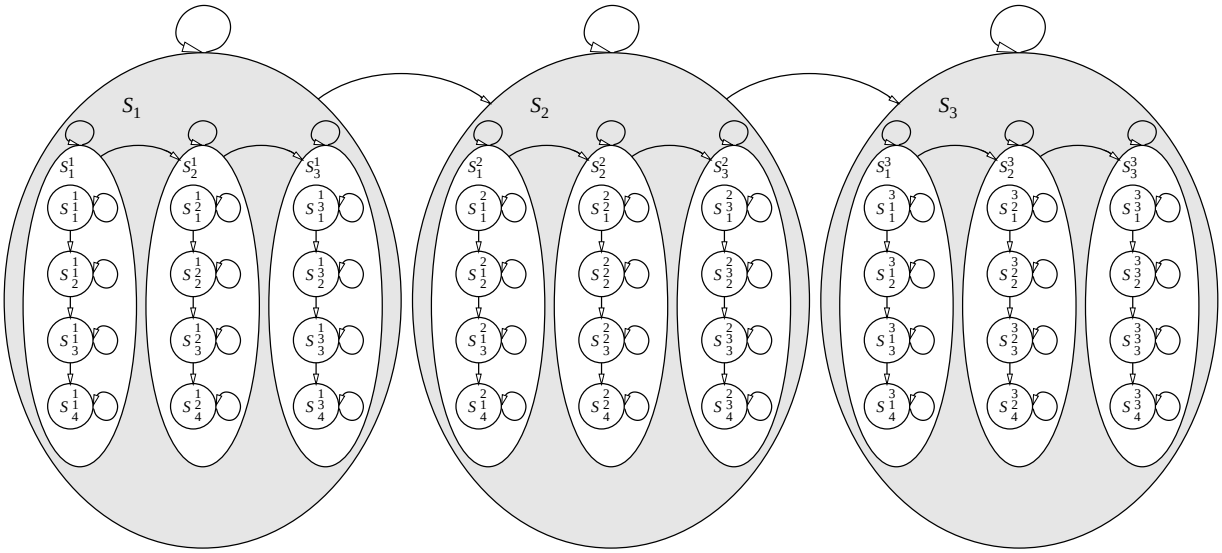

Abbildung 6.1: Pseudo dreidimensionales Hidden-Markov-Modell

eine nichtlineare Musterverzerrung in den beiden Bilddimensionen ermöglicht (siehe auch Kapitel 4.3).

Ein pseudo dreidimensionales HMM (L) ist durch die folgenden Parameter bestimmt: Zunächst ist die Anzahl K der Hyperzustände festzulegen. Diesen Zuständen sind Übergangswahrscheinlichkeiten a_{ij} zugeordnet, die in gleicher Weise, wie im eindimensionalen Fall, definiert sind (siehe Gleichung 2.2). Ebenso entspricht die Definition der Wahrscheinlichkeit für den Anfangszustand (π_j) der Gleichung 2.5. Jedem Hyperzustand des P3DHMMs ist ein P2DHMM zugeordnet. Diese können wie in Kapitel 4.3 dargestellt, definiert werden. Jeder Modellparameter erhält einen zusätzlichen hochgestellten Index, der die Zugehörigkeit zum entsprechenden Hyperzustand angibt (siehe auch Abb. 6.1). Es ergeben sich somit die folgenden Parameter für die pseudo zweidimensionalen Modelle $\Lambda_1, \ldots, \Lambda_K$, die den K Hyperzuständen zugeordnet sind:

- L^j ist die Anzahl der Metazustände $(S^j_1, \ldots, S^j_{L^j})$ des dem j-ten Hyperzustand zugeordneten P2DHMMs.

- Diesen Metazuständen sind Übergangswahrscheinlichkeiten a^j_{kl} zugeordnet:

$$a^j_{kl} = P(q_{x,t} = S^j_l | q_{x-1,t} = S^j_k) \tag{6.2}$$

  Dabei ist mit $q_{x,t}$ die Zufallsvariable für die Einnahme eines Metazustandes für die Bildspalte x zum Zeitpunkt t bezeichnet.

- Die Wahrscheinlichkeit für die Anfangszustände sind gegeben durch:

$$\pi^j_i = P(q_{1,t} = S^j_i) \tag{6.3}$$

Jedem Metazustand ist wiederum ein eindimensionales HMM zugeordnet. Dieses wird durch die folgenden Parameter bestimmt:

- M_i^j ist die Anzahl der Zustände $(S_{\,1}^{\,j}{}_i, \ldots, S_{\,M_i^j}^{\,j}{}_i)$ des dem j-ten Hyperzustand und dem i-ten Metazustand zugeordneten eindimensionalen Hidden-Markov-Modells.

- Diesen Zuständen sind die folgenden Übergangswahrscheinlichkeiten $a_{\,kl}^{\,j}{}_i$ zugeordnet:

$$a_{\,kl}^{\,j}{}_i = P(q_{x,y,t} = S_{\,l}^{\,j}{}_i | q_{x,y-1,t} = S_{\,k}^{\,j}{}_i) \tag{6.4}$$

  Dabei ist mit $q_{x,y,t}$ die Zufallsvariable für die Einnahme eines Zustandes für die Bildspalte x und die Bildzeile y zum Zeitpunkt t bezeichnet.

- Die Wahrscheinlichkeit für die Anfangszustände sind gegeben durch:

$$\pi_{\,k}^{\,j}{}_i = P(q_{x,1,t} = S_{\,k}^{\,j}{}_i) \tag{6.5}$$

- Die Ausgabeverteilungen können wiederum diskret oder kontinuierlich sein. Eine diskrete Ausgabeverteilung $b_{\,k}^{\,j}{}_i(l)$ über einem für das gesamte P3DHMM festgelegten Alphabet kann angegeben werden als:

$$b_{\,k}^{\,j}{}_i(l) = P(v_l | q_{x,y,t} = S_{\,k}^{\,j}{}_i) \tag{6.6}$$

  Die in Kapitel 2.2.4 dargestellten Gaußschen Mischverteilungen können alternativ verwendet werden, um kontinuierliche P3DHMMs zu erhalten.

Es sei an dieser Stelle angemerkt, daß die üblichen statistischen Randbedingungen für die vorgestellten Modellgrößen eingehalten werden müssen (siehe auch Gleichung 2.4). Durch die Spezifizierung der angegebenen Modellgrößen wird ein P3DHMM vollständig beschrieben. Es wurde in dieser Arbeit bereits an mehreren Stellen darauf hingewiesen, daß durch den Viterbi-Algorithmus die Möglichkeiten gegeben sind, HMMs für die Klassifikation einzusetzen und zudem auch die HMMs an Trainingsdaten anzupassen (siehe auch Kapitel 2 und 4). Für den Fall, daß ein verallgemeinerter Viterbi-Algorithmus existiert, ist es möglich, die P3DHMMs zu trainieren und Bildsequenzen mit den Modellen zu klassifizieren. Solch ein verallgemeinerter Viterbi-Algorithmus existiert in Form des dreifachverschachtelten Viterbi-Algorithmus. Dieser Algorithmus basiert auf dem zweifachverschachtelten Viterbi-Algorithmus für P2DHMMs (siehe Kapitel 4.3.2) und geht aus diesem hervor, indem ein weiterer, übergeordneter Viterbi-Durchlauf für die Zeitdimension hinzugefügt wird. Da es wie schon im zweidimensionalen Fall wiederum möglich ist, eine den P3DHMMs gleichwertige eindimensionale HMM-Struktur zu finden, wird an dieser Stelle auf die ausführliche Darstellung des dreifachverschachtelten Viterbi-Algorithmus verzichtet. Das folgende Unterkapitel beschreibt eine gleichwertige eindimensionale Modellierung, die es ermöglicht, die in Kapitel 2 vorgestellten Trainings- und Klassifikationsalgorithmen zu verwenden.

6.1.2 Umformung in gleichwertige eindimensionale Hidden-Markov-Modelle

Auf ähnliche Weise, wie im zweidimensionalen Fall (siehe Kapitel 4.3.3) ist im Rahmen dieser Arbeit mit Hilfe von Markierungszuständen und -merkmalen eine eindimensionale Modellierung entwickelt worden, die gleichwertig mit der pseudo dreidimensionalen Modellierung ist. Die grundlegende Idee dabei ist, die von Samaria in [Sam94b] vorgeschlagene Modellierungstechnik zweimal anzuwenden. Dies bedeutet, daß außer den Markierungsmerkmalen für den Anfang einer Bildspalte auch Merkmale eingefügt werden müssen, die den Anfang eines Einzelbildes markieren. Zusätzlich sind auch Markierungszustände dem eindimensionalen HMM hinzuzufügen, deren Ausgabeverteilungen mit den Markierungsmerkmalen korrespondieren.

Die Abbildungen 6.2 und 6.3 stellen schematisch die gleichwertige eindimensionale Modellierung dar. In Abb. 6.2 ist das zu modellierende dreidimensionale Muster gezeigt, bei dem es sich um einen Ausschnitt aus einer Bildsequenz, die einer Gestendatenbasis entnommen wurde, handelt. Um die Modellierung mit eindimensionalen HMMs zu ermöglichen, müssen die Merkmale der Bildsequenz in eine Merkmalsequenz überführt werden. Dies geschieht in der gleichen Weise, wie im zweidimensionalen Fall: Merkmale werden mit einem Abtastfenster für jedes Einzelbild von oben nach unten und links nach rechts entnommen (vgl. auch Abb. 4.3). Der Anfang einer jeden Bildspalte wird durch ein in Abb. 6.2 grau dargestelltes Markierungsmerkmal angezeigt. Die auf diese Weise erhaltenen Merkmale der Einzelbilder werden unter Berücksichtigung der zeitlichen Reihenfolge und unter Verwendung von weiteren Markierungsmerkmalen aneinandergehängt. Diese zusätzlichen Markierungsmerkmale sind in Abb. 6.2 schwarz dargestellt und zeigen den Anfang eines Einzelbildes an.

Abb. 6.3 illustriert die eindimensionale Modelltopologie, mit der ganze Bildsequenzen modelliert werden können. Die grau-schattierten Zustände sind die Markierungszustände, die beim Auftreten eines Markierungsmerkmals, das den Anfang einer Bildspalte anzeigt, hohe Wahrscheinlichkeiten ausgeben. Die schwarz dargestellten Markierungszustände geben hohe Wahrscheinlichkeiten beim Auftreten eines Markierungszustandes, das den Anfang eines Einzelbildes anzeigt, aus. Da die Markierungsmerkmale sowohl für die Bildspalten als auch für die Einzelbilder nur einzeln in der Merkmalsequenz auftreten, sind die Selbstübergänge der zugehörigen Markierungszustände auf Null zu setzen (z.B. $a_{100,100} = P(q_k = S_{100}|q_{k-1} = S_{100}) = 0, a_{110,110} = 0, a_{120,120} = 0, \ldots$). Die Übergangswahrscheinlichkeiten, die zu den Markierungszuständen führen, also z.B. $a_{114,110}$ und $a_{134,100}$ modellieren die statistischen Abhängigkeiten aufeinanderfolgender Bildspalten und Einzelbilder. Sie entsprechen somit den Übergangswahrscheinlichkeiten der Metazustände, bzw. der Hyperzustände.

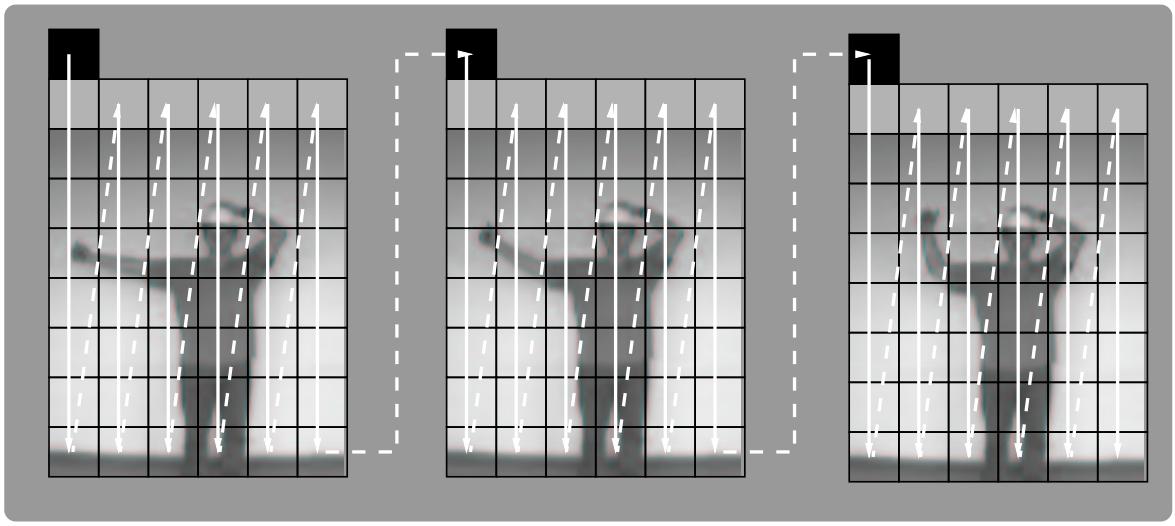

Abbildung 6.2: Überführung der Merkmale einer Bildfolge in eine eindimensionale Sequenz

Obwohl es die unterschiedliche farbliche Gestaltung der Markierungszustände und -merkmale in den Abb. 6.2 und 6.3 vermuten läßt, müssen für diese keine unterschiedlichen Ausgabefunktionen bzw. Werte gewählt werden. Sowohl für die Markierungsmerkmale der Einzelbilder, als auch der Bildspalten können dieselben Werte verwendet werden und somit die Parameter der Markierungszustände in Abb. 6.3 alle miteinander verknüpft werden. Dies ist möglich, da der Anfang eines Einzelbildes durch zwei aufeinanderfolgende Markierungsmerkmale stets eindeutig gekennzeichnet ist. Wie schon im zweidimensionalen Fall ist bei dieser Modellierungsmethode darauf zu achten, daß die Markierungsmerkmale ausschließlich den Markierungszuständen zugeordnet werden. Um dies zu erreichen, können unverändert die in Unterkapitel 4.3.3 beschriebenen Methoden verwendet werden.

Da es sich bei der in diesem Unterkapitel beschriebenen Modelltopologie um eine eindimensionale handelt, können die Trainings- und Klassifikationsalgorithmen des Kapitels 2 verwendet werden. Im folgenden werden nach einer kurzen Einführung in das Gebiet der Bildsequenzklassifikation experimentelle Ergebnisse präsentiert, die mit der vorgestellten Modellierungstechnik erreicht wurden.

6.2 Klassifikation von Bildsequenzen

Die populärste Anwendung von Hidden-Markov-Modellen ist der Bereich der Klassifikation von sich zeitlich ändernden Mustern. Es liegt somit nahe, diese Modelle auch für die Klassifikation von Bildsequenzen einzusetzen. Dies erfolgte in den verschiedensten Anwendungsszenarien, wie beispielsweise Videoindexierung oder Gestikerkennung. Der letztgenannte Bereich, die Gestikerkennung, ist als Anwendungsszenario für die Evaluierung der

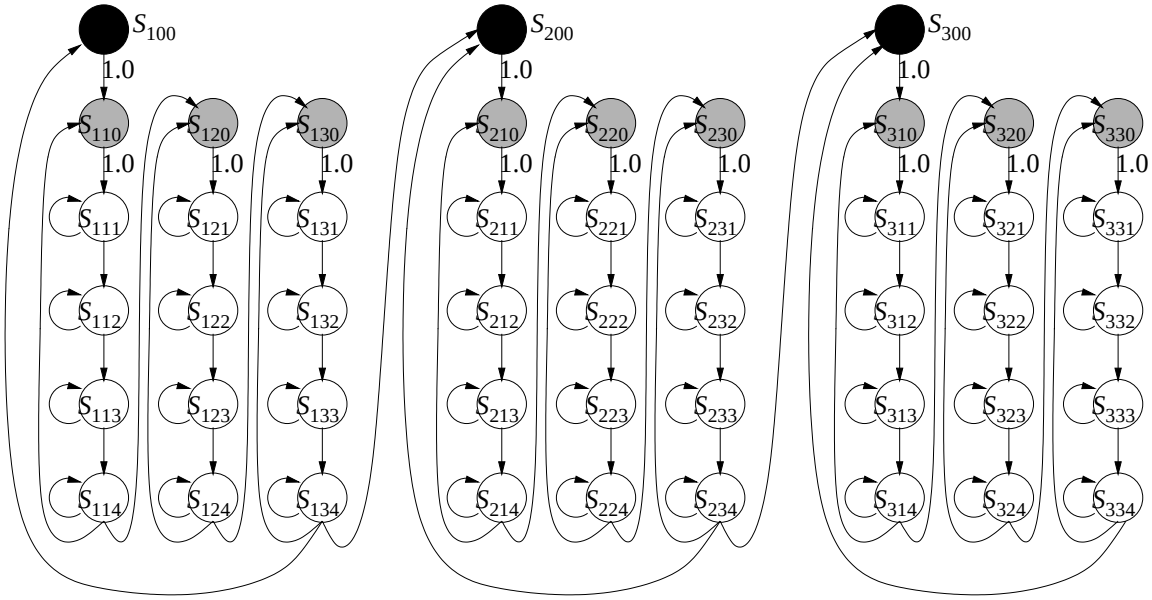

Abbildung 6.3: Eindimensionale HMM-Struktur mit Markierungszuständen, die eine im Vergleich mit den P3DHMMs gleichwertige Modellierung ermöglicht.

neuartigen pseudo dreidimensionalen Hidden-Markov-Modelle gewählt worden, da bereits Ergebnisse mit einem alternativen Verfahren am Lehrstuhl Technische Informatik vorliegen und sich somit eine Vergleichsmöglichkeit bietet. Bevor der Einsatz der P3DHMMs auf dem Gebiet der Gestikerkennung beschrieben wird, werden im folgenden Unterkapitel zunächst relevante Arbeiten anderer Autoren vorgestellt.

6.2.1 Relevante Arbeiten anderer Autoren

Es existiert eine große Zahl von Arbeiten, vorwiegend aus den 90er Jahren, die über die Erkennung von Gesten mit Hidden-Markov-Modellen berichten. Die wahrscheinlich erste Arbeit zu diesem Thema stammt von Yamato et al. [Yam92] und beschreibt Experimente mit diskreten eindimensionalen Hidden-Markov-Modellen, die für die Klassifikation von Tennis-Schlagtechniken eingesetzt werden. Es wurden die folgenden sechs Tennis-Schlagtechniken verwendet, die die zu erkennenden Klassen bilden: Rückhand-Volley, Rückhand-Schlag, Vorhand-Volley, Vorhand-Schlag, Schmettern und Aufschlag. Es wird eine Vielzahl von Vorverarbeitungsschritten, wie z.B. Tiefpaßfilterung, Hintergrundsubtraktion und Binarisierung auf jedes einzelne Bild der Sequenz angewendet. Das Ergebnis dieser Vorverarbeitung ist ein zweiwertiges Bild, das im wesentlichen die extrahierte Pose der Person darstellt. Vor der Berechnung der Merkmale wird zusätzlich eine Größennormalisierung und eine Zentrierung vorgenommen. Die Merkmale sind die Anzahlen von schwarzen Bildpunkten in den Segmenten eines Abtastrasters. Diese Merkmale werden anschließend vektorquantisiert und somit ergibt sich eine Sequenz von Symbolen, die die Bildsequenz repräsentiert. Diese Se-

quenz von Symbolen kann mit einem eindimensionalen diskreten Hidden-Markov-Modell weiterverarbeitet werden.

Schuster und Rigoll verwenden in der Arbeit [Sch96] ebenfalls diskrete Hidden-Markov-Modelle für die Erkennung von Bildsequenzen. Der Hauptunterschied zu [Yam92] liegt in der Verwendung einer wesentlich simpleren Vorverarbeitung, was den Echtzeiteinsatz ermöglicht. Die Vorverarbeitung in [Sch96] besteht lediglich aus der Unterabtastung der RGB-Kanäle der einzelnen Farbbilder einer Sequenz. Horizontale bzw. vertikale Streifen dieser Abtastwerte werden anschließend, ohne weitere Verarbeitungsschritte, vektorquantisiert und zusammen mit den diskreten HMMs für die Klassifikation eingesetzt. Alternativ wurden dieselben Schritte auf Differenzbildern angewendet. Das echtzeitfähige System wurde auf einer Gestendatenbasis evaluiert, die aus zehn selbstdefinierten Klassen besteht. Beispiele für diese Klassen sind: klatschen, sich verbeugen, nicken und den Kopf schütteln.

Das oben beschriebene System ist durch die Verwendung von kontinuierlichen Hidden-Markov-Modellen zusammen mit geometrischen Momenten, die auf Differenzbildern berechnet werden, weiter verbessert worden. Dieses System verwendet, wie in [Rig96] berichtet wurde, 24 verschieden Klassen, die mit einer Genauigkeit von mehr als 90% erkannt werden.

Die Kombination aus kontinuierlichen Hidden-Markov-Modellen und geometrischen Momenten wurde ebenfalls von Starner et al. in der Arbeit [Sta98] verwendet. Das System in [Sta98] erkennt amerikanische Zeichensprache und verwendet die folgenden Vorverarbeitungschritte: Die Hände der Person werden in den einzelnen Bildern der Sequenz lokalisiert und basierend auf diesen Regionen werden Momente berechnet. Neben diesen Merkmalen werden dynamische Merkmale, wie die Positionsveränderung der Hände zwischen den Einzelbildern verwendet.

Die bisher kurz vorgestellten Systeme sind sehr stark abhängig von der Existenz von Bewegung, da sehr oft Merkmale, die auf Differenzbildern basieren, eingesetzt werden. Diese Einschränkung kann durch den Einsatz von pseudo dreidimensionalen Hidden-Markov-Modellen überwunden werden. Dies wird im folgenden Kapitel erläutert.

6.2.2 Klassifikation von Bildsequenzen mit P3DHMMs

Mit Hilfe der pseudo dreidimensionalen Hidden-Markov-Modelle können sowohl auf Bewegung basierende Merkmale als auch statische, auf den Einzelbildern berechnete Merkmale gemeinsam verwendet werden. Die Integration in ein einzelnes Modell erfolgt über die Merkmalströme (siehe auch Kapitel 2). Zusätzlich ermöglicht die Modellierung mit P3DHMMs ein flexibles Erkennungsverhalten auf den einzelnen Bildern der Sequenz. Dies ist ein großer Unterschied zu den im vorhergehenden Unterkapitel vorgestellten Ansätzen, da hier entweder VQ-Indices ([Yam92, Sch96]) oder globale Merkmale ([Rig96, Sta98]) auf den einzelnen Bildern der Sequenz berechnet werden und somit die Bilder auf starre Weise

modelliert werden. Der Vorteil des flexiblen Erkennungsverhaltens auf den Einzelbildern bei Verwendung der P3DHMMs ist, neben der besseren Erkennungsleistung, die Toleranz gegenüber Positionsveränderungen. Ist die Position einer gestikulierenden Person in einer Bildsequenz beispielsweise relativ zu der Position in einer Trainingssequenz verändert, so wird dies durch die flexible Modellierung der Einzelbilder durch die P2DHMMs ausgeglichen. Die P2DHMMs ermöglichen eine nichtlineare Musterverzerrung in beiden Bilddimensionen und daher wird insbesondere die Verschiebung in x-Richtung, also eine translatorische Positionsveränderung, kompensiert. Dies erlaubt die Erkennung von Gesten auch für den Fall, daß sich die gestikulierende Person selbst in einer translatorischen Bewegung befindet und somit die Position im Bild verändert.

Die zusammen mit den P3DHMMs verwendete Merkmalextraktion basiert, wie im Fall der Bildklassifikation mit P2DHMMs (siehe Unterkapitel 5.1), auf der Diskreten-Kosinus-Transformation. Die DCT-Koeffizienten werden sowohl auf den Einzelbildern der Sequenz als auch auf den Differenzbildern berechnet und somit ergeben sich statische und dynamische Merkmale. Durch die Verwendung der Merkmalströme können diese Merkmale zu heterogenen Merkmalvektoren zusammengefaßt werden. Die Merkmalstrom-Gewichte ermöglichen ferner, den Einfluß der statischen und der dynamischen Merkmale zu kontrollieren.

6.3 Experimentelle Ergebnisse

Die P3DHMMs wurden anhand einer aus 12 Klassen bestehenden Gesten-Datenbasis evaluiert. Zusätzlich sind die erzielten Ergebnisse mit denen, die mit einem alternativen Verfahren erreicht wurden, verglichen worden. Bevor diese Ergebnisse detailliert vorgestellt werden, wird zunächst der Aufbau der verwendeten Datenbasis erläutert.

6.3.1 Gesten-Datenbasis

Die in den Experimenten verwendete Gesten-Datenbasis wurde am Fachgebiet Technische Informatik vom Autor erstellt. Die Datenbasis besteht aus 12 verschiedenen Gesten, die der Steuerung von Baukränen dienen. Diese Gesten ermöglichen das Manövrieren von Kränen in vom Kranführer schwer einsehbarem Gelände. Eine zweite Person, deren Sicht auf eine zu positionierende Last besser ist, kann durch Ausführen der Gesten dem Kranführer assistieren. Dieses Gestenvokabular ist wohldefiniert und ist z.B. in dem Nachschlagewerk für Mechanik [Par80] zu finden. Abb. 6.4 illustriert die verschiedenen Klassen, die folgendermaßen benannt sind: links-herumdrehen, rechts-herumdrehen, näherkommen, entfernen, Lastarm ausfahren, Lastarm einziehen, Lastarm hoch, Lastarm runter, hochwinden, herunterlassen, halt, nothalt. Die beiden letztgenannten Klassen stellen Beispiele für statische Gesten dar, da sie durch Körperhaltungen und nicht durch Bewegungsabläufe definiert sind, was Abb. 6.4 entnommen werden kann. Fünf Personen führten die in Abb. 6.4 definierten

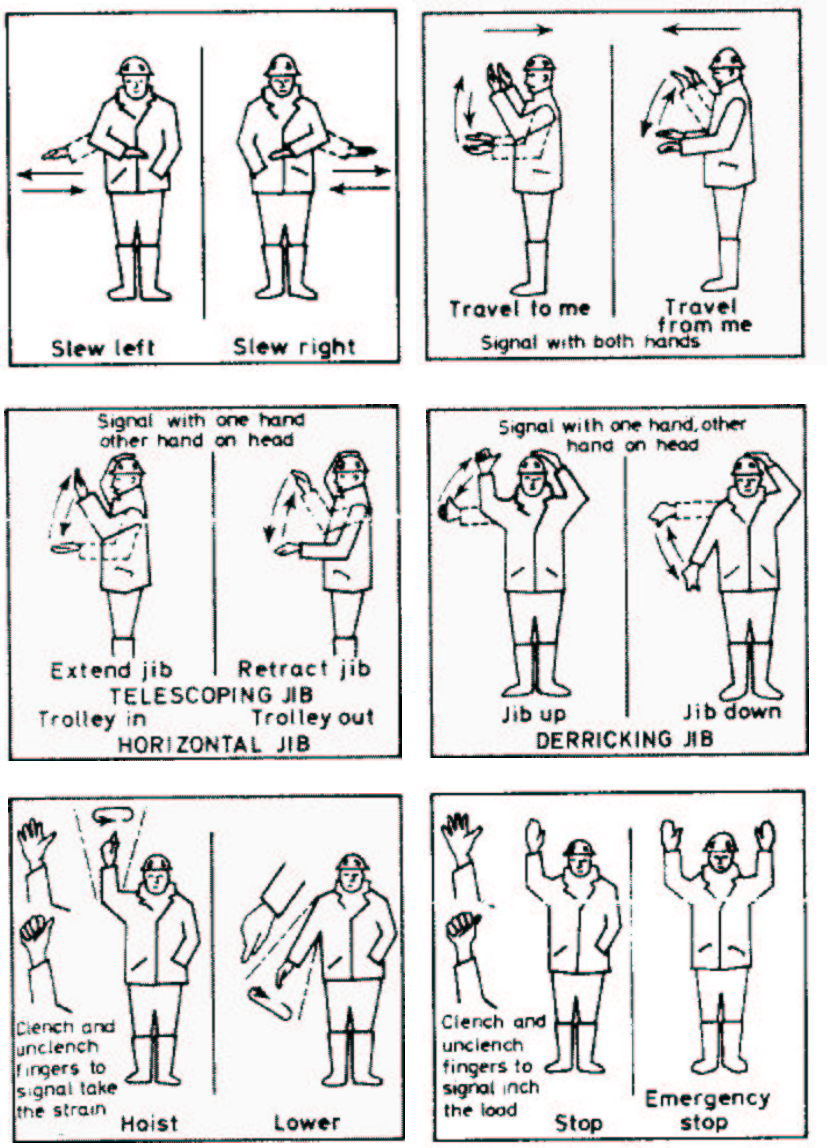

Abbildung 6.4: Darstellung der 12 Gesten, mit denen Baukräne manövriert werden können. Ins Deutsche übersetzt ergeben sich die folgenden Klassenbezeichnungen (von links nach rechts und oben nach untern): links-herumdrehen, rechts-herumdrehen, näherkommen, entfernen, Lastarm ausfahren, Lastarm einziehen, Lastarm hoch, Lastarm runter, hochwinden, herunterlassen, halt, nothalt (aus [Par80])

Abbildung 6.5: Ausschnitte aus Sequenzen, die die Gesten *Lastarm hoch* und *Lastarm runter* darstellen

Bewegungsabläufe mehrere Male durch. Zwei Beispiele für jede Gesten-Klasse dienen als Trainingsdatensatz, die verbleibenden Beispiele werden in der Erkennungsphase verwendet. Die Abb. 6.5 zeigt zwei Ausschnitte aus Sequenzen, die zu den Klassen *Lastarm hoch* (obere Zeile) und zur Klasse *Lastarm runter* (untere Zeile) gehören. Die Bildsequenzen sind mit einer Auflösung von 192×144 Bildpunkten und einer Bildwiederholrate von 25 Bildern pro Sekunde digitalisiert worden. Die Aufnahme selbst erfolgte mit einer analogen Videokamera.

6.3.2 Quantitative Ergebnisse

Es wurde bei der Durchführung der Experimente die folgende P3DHMM-Topologie verwendet: Das Modell bestand aus 4 Hyperzuständen, denen jeweils ein (5×5) P2DHMM zugeordnet war. Die Größe der Abtastfenster, auf die eine diskrete Cosinustransformation angewendet wurde, betrug 16×16 Bildpunkte. Die DCT-Koeffizienten wurden sowohl auf Abtastfenstern, die Grauwerte enthalten, als auch auf Abtastfenstern, die Differenzen benachbarter Einzelbilder enthalten, berechnet. Die dabei entstehenden Merkmalströme wurden auf gleiche Weise mit den Werten $\gamma_1 = \gamma_2 = 1$ gewichtet. In Tabelle 6.1 sind in der zweiten Spalte die mit dieser Konfiguration erzielten Erkennungsgenauigkeiten angegeben. Diese Erkennungsgenauigkeiten wurden jeweils getrennt für die einzelnen Personen ermittelt, die in Tabelle 6.1 in der ersten Spalte aufgelistet sind. Zusätzlich sind zur besseren Bewertbarkeit dieser Ergebnisse in der dritten Spalte Erkennungsgenauigkeiten angegeben, die mit einem alternativen Ansatz erzielt worden sind. Dieser alternative Ansatz verwendet geometrische Momente, die auf den Differenzbildern berechnet werden, sowie eindimensionale kontinuierliche Hidden-Markov-Modelle zur zeitlichen Modellierung. Eine ausführliche Darstellung dieses Ansatzes ist in [Rig97] zu finden. Der Tabelle 6.1 kann entnommen werden, daß der neuartige P3DHMM basierte Ansatz im Vergleich zur eindimensionalen Modellierung eine höhere durchschnittliche Erkennungsgenauigkeit erzielt hat. Darüber hinaus existieren noch zwei weitere Vorteile der pseudo dreidimensionalen Hidden-Markov-Modelle: Zum einen können mit dieser Methode statische und dynamische Gesten gemeinsam und unter Verwendung eines Ansatzes modelliert und klassifiziert werden. Ein zweiter Vorteil ist, daß durch die elastische Modellierung der Einzelbilder durch die P3DHMMs eine positions- und größen-

Person	P3DHMM	1DHMM
ste	88,6%	100%
stm	91,2%	85,3%
ank	100%	100%
bw	94,1%	88,2%
jmr	80,5%	80,5%
Durchschnitt	90,88%	90,74%

Tabelle 6.1: In den Experimenten erzielte Erkennungsgenauigkeiten

tolerante Erkennung durchgeführt werden kann. Dies wird durch Experimente belegt, die in [Yal00b] und [Yal00a] dokumentiert sind. Neben der Anwendung in der Gestikerkennung können P3DHMMs auch auf anderen Gebieten eingesetzt werden. So sind z.B. in [Hue01] Experimente beschrieben, die die Eignung der P3DHMMs für die Erkennung menschlicher Gesichtsausdrücke bzw. Gemütszustände belegen.

6.4 Ausblick auf einen integrierten Ansatz zur Klassifikation und Segmentierung mit P3DHMMs

In Kapitel 5 wurde für den zweidimensionalen Fall zunächst die Klassifikation von Einzelbildern mittels pseudo zweidimensionaler Hidden-Markov-Modelle beschrieben und, im Anschluß daran, die integrierte Segmentierung und Klassifikation von komplexen Szenen durch erweiterterte P2DHMMs vorgestellt. Für den in diesem Kapitel betrachteten dreidimensionalen Fall wurde bisher die Klassifikation von Bildsequenzen mit neuartigen pseudo dreidimensionalen HMMs vorgestellt. Obwohl die in Kapitel 5 vorgestellten Modellierungstechniken nahezu unverändert auf den dreidimensionalen Fall übertragen werden können, werden in dieser Arbeit keine Experimente mit den um Umgebungszustände erweiterten P3DHMMs vorgestellt. Der Grund hierfür ist der sehr hohe Rechenbedarf, der für solche Experimente benötigt wird. Als Ausblick soll an dieser Stelle auf die möglichen Einsatzgebiete für die integrierte Segmentierung und Klassifikation mit P3DHMMs hingewiesen werden. Die Modelltopologie für diesen Ansatz ist in Abb. 6.6 dargestellt. Wie schon in Abbildung 5.4 sind die Umgebungszustände grau schattiert, während die klassenbeschreibenden Zustände weiß ausgefüllt sind. Die Parameter der Umgebungszustände können, wie es in Kapitel 5.3.1 dargestellt wurde, z.B. unter Verwendung aller Merkmale einer zu analysierenden Sequenz bestimmt werden. Der integrierte Segmentierungs- und Klassifikationsansatz mit P3DHMM kann für die folgenden Aufgaben eingesetzt werden:

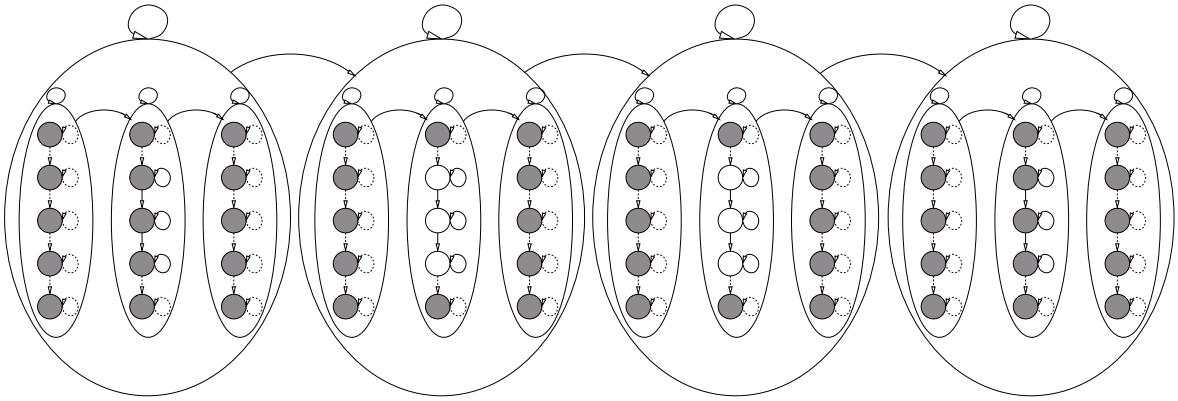

Abbildung 6.6: Pseudo dreidimensionales Hidden-Markov-Modell mit Umgebungszuständen

- Auffinden von Personen in Bildsequenzen aufgrund charakteristischer Bewegungen und des Aussehens
- Erkennen von Gesten einer sich bewegenden Person
- Abfrage von Filmdatenbanken mit folgenden Aufgaben
  - Auffinden von Actionszenen in Spielfilmen
  - Erkennen von Bewegungsabläufen im Sport
  - Anfrage an Filmdatenbanken mit Beispielgesten

Das erstgenannte Beispiel stellt eine um die Zeitdimension erweiterte Version des bereits in der Einleitung vorgestellten Problems dar, ein bekanntes Muster zu erkennen, das in eine komplexen Umgebung eingebettet ist (siehe Abb. 1.1). Da nun das zu erkennende Muster auch in der Zeitdimension aufzufinden ist, besteht der erste und der letzte Hyperzustand des erweiterten P3DHMM in Abb. 6.6 ausschließlich aus Umgebungszuständen. Durch diese Maßnahme wird es ermöglicht, ein Muster zu erkennen, welches am Anfang bzw. am Ende einer Bildsequenz möglicherweise *nicht* vorkommt.

6.5 Kapitelzusammenfassung

Es wurde die Klassifikation von Bildfsequenzen mit neuartigen pseudo dreidimensionalen Hidden-Markov-Modellen vorgestellt. Diese Modellierung ermöglicht es, Merkmale, die auf Einzelbildern berechnet wurden, gemeinsam mit Merkmalen zu verwenden, die aus der temporalen Abfolge der Bilder bestimmt wurden. Somit können dynamische und statische Gesten gemeinsam mit einem Modell erkannt werden.

Es wurde in die Theorie der P3DHMMs eingeführt und dargestellt, wie gleichwertige eindimensionale Modelle konstruiert werden können. Die Klassifikation von Bildsequenzen wurde anhand einer Gestendatenbasis demonstriert. Die verwendete Datenbasis besteht

aus 12 Gesten, die zur Steuerung von Baukränen dienen. In Experimenten wurden auf dieser Datenbasis höhere Erkennungsgenauigkeiten erzielt, als mit einem alternativen Ansatz, der eindimensionale Hidden-Markov-Modelle in Kombination mit geometrischen Momenten verwendet. Als Ausblick wurde auf die integrierte Segmentierung und Klassifikation von Bildsequenzen hingewiesen, die z.B. eine positionsunabhängige und gleichzeitig hintergrundunabhängige Gestenerkennung ermöglicht.