Automatische Detektion menschlichen Kopfnickens und Wege zur Interpretation

Dissertation

Eduard Wall

Erklärung der Urheberschaft

Gemäß der Promotionsordnung der Universität Bielefeld S 8 (1) g: erkläre ich hiermit, von der Rahmenpromotionsordnung der Universität Bielefeld, sowie der Promotionsordnung der Technischen Fakultät Kenntnis genommen zu haben. Darüber hinaus bestätige ich, dieses Thema selbstständig bearbeitet und die Dissertation eigenständig verfasst zu haben. Die Verwendung von Arbeiten anderer Personen habe ich stets gekennzeichnet. Dritte haben weder direkt noch indirekt finanzielle Vorteile im Zusammenhang mit dem Inhalt dieser Arbeit erhalten. Für diese Arbeit habe ich keine laufenden oder vorausgegangenen Promotionsgesuche gestellt.

Eduard Wall	Ort, Datum

Automatische Detektion menschlichen Kopfnickens und Wege zur Interpretation

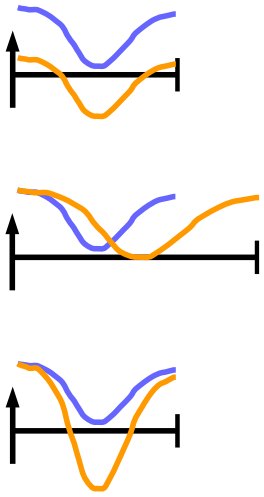
Dissertation

Eduard Wall

Juli 2021

Dissertationsschrift zur Erlangung des akademischen Grades Doktor der Ingenieurwissenschaften (Dr.-Ing.) der

Technischen Fakultät
Universität Bielefeld
Angewandte Informatik
Inspiration 1
33619 Bielefeld
Deutschland



Gutachter

Prof. Dr.-Ing. Franz Kummert
Prof. Dr.-rer.nat. habil. Christian Wöhler

Prüfungsausschuss

Prof. Dr.-Ing. Ulrich Rückert
Dr. rer. nat. Alexander Schulz

Verteidigt und anerkannt am 11. Juni 2021

Gedruckt auf alterungsbeständigem Papier nach DIN-ISO 9706.

Danksagung

Hiermit möchte ich die Möglichkeit nutzen, einigen Leuten Danke zu sagen.

Ich danke meinem Betreuer Franz Kummert für die Möglichkeit der Mitarbeit in der AGAI und für die unkomplizierte und sehr motivierende, zuverlässige Begleitung.

Lars Schillingmann hat mich einige Jahre sehr kompetent unterstützt und insbesondere danke ich ihn für die Mithilfe an der Publikation für die RO-MAN 2017. Ich danke auch Jakob Naurath und Kirsten Kästel, die als wissenschaftliche Hilfskräfte die ein oder andere Aufgabe mit übernommen haben, sowie dem KOMPASS-Projekt-Team.

Vielen Dank an meine Eltern nicht nur für die Unterstützung auf meinem Bildungsweg.

Ein besonderer Dank gilt meiner wunderbaren Frau Manuela, die mir für die Fertigstellung dieser Arbeit immer wieder Zeit eingeräumt, mich geschoben und motiviert hat.

Über alles aber danke ich Gott.

Kurzfassung

Die Dissertation steht unter der Fragestellung, inwiefern automatisches Erkennen und Interpretieren von menschlichem Kopfnicken die Mensch-Agenten-Interaktion bereichern kann. Dafür werden die Grundlagen menschlichen Kopfnickens im kommunikationswissenschaftlichen Kontext erörtert. Es wird zunächst ein Detektionssystem sowohl auf Basis von Dynamic Time Warping (DTW) als auch einer Support Vektor Maschine (SVM) realisiert. Dieses wird mit Hilfe von realitätsnahen Anwendungsszenarien und einer erweiterten Kreuzvalidierung getestet; In den umgesetzten Varianten zeigt SVM eine geringfügig höhere Erkennungsrate als DTW. Dennoch werden eher bei DTW vielversprechende Erweiterungsmöglichkeiten deutlich. Im Vergleich mit einer führenden Veröffentlichung erreicht das System nicht ganz dessen Leistungsfähigkeit, kommt dafür aber mit weniger Information und ohne spezielle Hardware aus. Zur Untersuchung von automatischer Interpretation und Kategorisierung werden einige Theorien und Modelle zur Klassifikation von Kopfnicken herausgearbeitet und anhand des vorliegenden Datenmaterials nachvollzogen; Ein 3-Kategorien-Modell zeigt hierbei Anwendungspotential im Erkennen von nonverbaler Ankündigung von Äußerungen und kann somit harten Gesprächsunterbrechungen vorbeugen. Auch anhand eines Commitment-Stärken-Modells werden Hinweise auf einen Zusammenhang zwischen physikalischer Nick-Ausführung und Bedeutung von Kopfnicken geliefert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Seit Beginn der Menschheit ist die zwischenmenschliche Interaktion und Kommunikation ein zentrales und allgegenwärtiges Thema des menschlichen Miteinanders. Treten Menschen in einen Dialog, findet Verständigung stets auf verschiedenen Ebenen statt. Neben den gesprochenen Wörtern und der prosodischen Information sind Hand- und Kopfgesten sowie Gesichtsausdrücke von großer Bedeutung. Bereits Kinder äußern sich in Form von Kopfbewegungen schon bevor sie zu sprechen beginnen. So zeigte M. Guidetti 2005 an französischen Kindern, wie sie bereits im Alter von einem Jahr Nicken als Zustimmung und Kopfschütteln als Ablehnung einsetzten (Thelen et al. (2001)). Die Nutzung von Gesten wird später durch Sprache nicht ersetzt, sondern ergänzt und wird ein Leben lang beibehalten. So wird der Anteil an durch nonverbales Verhalten transportierten Information häufig unterschätzt.

Diagramm 1.1 zeigt die Ergebnisse einer bekannten Studie des Psychologen A. Mehrabian (Mehrabian and Wiener (1967)). Er untersuchte, wie viel Information über welchen Kommunikationskanal übertragen wird. Einschränkend sollte dabei stets erwähnt werden, dass sich seine Studien lediglich auf den Ausdruck eines ausgewählten Repertoires an Gefühlen und Gesinnungen beziehen. Dennoch geben uns die Zahlen eine Idee davon, dass nonverbale Signale in der Kommunikation wichtig sind, zumal eine menschliche Äußerung wohl nie ganz unabhängig von Gefühlen und Gesinnungen hervorgebracht wird.

Der sprachliche Kanal, zumindest reduziert auf die sprachwissenschaftliche Interpretation der ausgesprochenen Wörter, lässt sich recht klar untersuchen. Wörter werden bewusst gewählt, eingesetzt und haben oft klare Definitionen. Dagegen findet nonverbale Kommunikation meist unbewusst statt, wird in der Regel nicht gezielt eingesetzt und ist daher schwierig zu erkennen und zu verstehen.

In einer natürlichen Gesprächssituation wird neben Körperhaltung und händischen Gesten den Kopfgesten besondere Bedeutung beigemessen, da

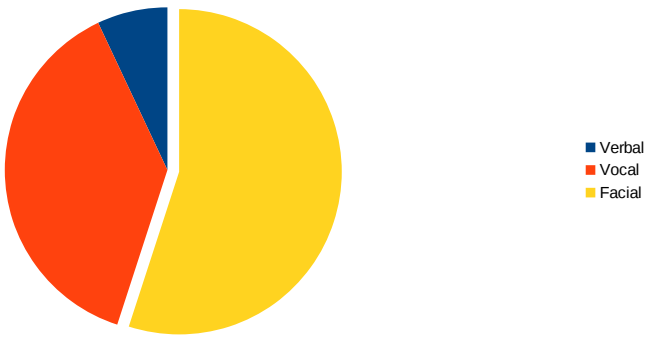


Abbildung 1.1: Dieses Diagramm zeigt die Anteile an Information dreier Kommunikationskanäle: Verbale Äußerungen (Verbal) mit 7%, prosodischer Anteil (Vocal) mit 38% und Gesichtsausdrücke (Facial) mit 55% als Ergebnis einer Studie von Mehrabian, 1972.

nicht zuletzt durch Blickkontakt und Sprachäußerungen der Kopf Fokus der Aufmerksamkeit ist. Harrigen machte im Jahr 2005 die Feststellung, dass fast 80 Prozent der publizierten Forschungsartikel im Bereich der Kopfbewegungen von Kopfnicken handeln (Harrigan et al. (2005)). Dies lässt die Annahme zu, dass Kopfnicken eines der wichtigsten und interessantesten nonverbalen Signale in der zwischenmenschlichen Kommunikation ist.

In heutiger Zeit findet Kommunikation und Interaktion jedoch zunehmend auch mit nichtmenschlichen Dialogpartnern statt. Virtuelle Assistenz- und Interaktionssysteme sowie Roboter werden unseren Alltag in Zukunft stärker prägen und somit wird die Frage immer wichtiger, wie man Mensch-Maschine-Interaktion natürlicher und effizienter gestalten kann. Beispiele des täglichen Lebens sind etwa automatisierte Telefon-Hotlines oder generell Spracherkennungssoftware wie Siri der Firma Apple (Meyer and Stiller (2011)). Während sich die meisten Dialogsysteme auf die sprachliche Ebene beschränken, bleibt viel Potential zu effizienter Verständigung ungenutzt. Daher wird zunehmend versucht, nonverbale Signale zu erkennen und zu

verarbeiten, um weitere Möglichkeiten zum Informationsgewinn zu nutzen. Automatische Detektionssysteme im Bereich des Kopfnickens wurden bisher kaum mit dem Anwendungsfall eines Dialogkontextes untersucht. Dabei ist auch die Evaluationsmethodik veröffentlichter Ansätze sehr unterschiedlich und lässt meist wenig Rückschlüsse auf die tatsächliche Leistungsfähigkeit in konkreten Dialogkontexten zu. Hier soll diese Arbeit einen Beitrag in Richtung Anwendungstauglichkeit in Dialogkontexten leisten.

1.1 Mensch-Agenten-Interaktion

Um die Interaktion zwischen einem Menschen und einem nicht-menschlichen System ranken sich viele Begriffe. Je nach Beschaffenheit dieses Systems spricht man von Mensch-Agenten-Interaktion, Mensch-Maschine-Interaktion, Mensch-Roboter-Interaktion oder von Mensch-Computer-Interaktion. Dabei haben die Begriffe große gemeinsame Schnittmengen. Mensch-Maschine-Interaktion ist der wohl umfassendste Begriff. Seit es Maschinen gibt, werden Methoden entwickelt, um eine Schnittstelle zwischen Mensch und Maschine herzustellen. Waren es vor Jahren noch mechanische Konfigurationswerkzeuge, wandelten sich die Interaktionsmöglichkeiten mit der Digitalisierung vermehrt durch elektronische Eingabegeräte. Mit der Entwicklung von bedienbaren Computern entstand so der Unterbegriff der Mensch-Computer-Interaktion. Durch immer leistungsstärkere Computertechnologien wurden so immer komplexere Interaktionen möglich, so dass dadurch als Untergruppe der neue Forschungszweig der Mensch-Agenten-Interaktion entstand.

Mensch-Agenten-Interaktion bezieht sich im Speziellen auf die Interaktion mit einem Agentensystem. Dabei besitzt das System meist ein verkörperlichtes, oft virtuelles Erscheinungsbild eines kompetenten Assistenten. Dieser ist in der Lage, durch Sensorik Nutzereingaben zu erfassen und besitzt Weltwissen, um diese Informationen im Hinblick auf ein internes Ziel zu verarbeiten. Dazu werden dem Nutzer selektiv entsprechende Informationen bereitgestellt, auf die der Nutzer reagieren kann. So entsteht die Interaktion. Dabei hebt sich die Mensch-Roboter-Interaktion hauptsächlich durch Attribute wie Mobilität, das Vorhandensein von physikalischem Körper, Gliedmaßen und mechanischen Manipulatoren von der Mensch-Agenten-Interaktion ab.

Die GI (Gesellschaft für Informatik) definiert auf deren Webseite (https://gi.de/, Zugriff 2020) allgemeine Ziele und Aufgaben von Mensch-Computer-Interaktion. Dabei werden als Schwerpunkte Kommunikation, Dialogsteuerung, Turn-Taking, Sprach- und Bildverarbeitung genannt. Diese Aufgaben spielen besonders beim Design eines kompetenten Assistenz-Systems eine Rolle mit dem Ziel, der Effizienz einer natürlichen Mensch-zu-Mensch-Interaktion nahe zu kommen. Dies schließt die intensive Erforschung von Mensch-zu-Mensch-Interaktion als Idealvorstellung mit ein. Eine wichtige Fragestellung dabei ist, inwiefern sich das Verhalten eines Nutzers mit einem menschlichen Gesprächspartner von einem virtuellen Agenten unterscheidet. Untersuchungen haben gezeigt, dass Menschen auch in nicht-menschlichen Dialog-Kontexten nonverbale Signale aussenden (Bavelas et al. (2008)). Dies gilt selbst dann, wenn dem Sender bewusst ist, dass der Adressat diese gar nicht erfassen kann. Eindrucksvoll gezeigt wurde der Effekt in der Studie von J. Bavelas (Bavelas et al. (2008)), in der anhand von Telefongesprächen gezeigt wurde, dass die Menge an ausgeführten Gesten unabhängig davon ist, ob die Gesprächspartner sich tatsächlich visuell wahrnehmen können oder nicht. Auch im direkten Kontext von Kopfnicken zeigt C.L. Sidner, dass Nutzer während einer Interaktion mit einem Roboter Nickverhalten zeigen, auch wenn der Roboter diese gar nicht registrieren kann (Sidner et al. (2006)).

Des Weiteren zeigte beispielsweise Goldin-Meadow, dass Kopfnicken nicht nur zusätzlich zur Sprache, sondern auch als Ersatz von Sprache verwendet wird (Goldin-Meadow (1999)).

Diese Effekte motivieren die Forschung im Bereich der Mensch-Agenten-Interaktion, nicht nur Sprache und Eingabe sensorisch zu erfassen, sondern auch gesendete visuelle Signale parallel zu erkennen und auszuwerten.

1.2 Zielsetzung

Das Ziel dieser Promotionsschrift widmet sich der Fragestellung, inwiefern das automatische Erkennen und Interpretieren von Kopfnicken in einer Mensch-Agenten-Interaktion dieselbe messbar bereichern kann. Im Blick auf dieses Ziel beschäftigt sich diese Arbeit einerseits mit den Grundlagen des menschlichen Kopfnickens in kommunikativen Kontexten, andererseits wird aber auch die praktische Realisierung und Anwendung einer automatischen

Erkennung und Interpretation von Kopfnicken verwirklicht.

Als Grundlage für automatische Verarbeitungsschritte soll ein System entwickelt werden, das zunächst allgemein ein menschliches Kopfnicken detektiert. Ein solches Detektionssystem sollte mit marktüblicher Hardware ausreichend leistungsfähig sein. Kopfnicken muss innerhalb kurzer Zeit erkannt werden, um eine kontext-entsprechende direkte Reaktion darauf zu ermöglichen. Es ist zu erwarten, dass in natürlichen Interaktionen das Vorkommen von Nicht-Nicken weitaus größer ist als Nicken. Daher ist für einen praxisnahen Einsatz vor Allem eine geringe Falsch-Positiv-Rate eine wichtige Anforderung an ein Detektionssystem. In diesem Kontext ist eine aussagekräftige Evaluationsstrategie wichtig. Dabei soll der Frage nachgegangen werden, ob Kopfnicken ausreichend effizient und sicher für die Mensch-Agenten-Interaktion erkannt werden kann.

Es schließt sich die Frage an, ob auf Basis einer soliden Erkennung auch eine automatische Interpretation von Kopfnicken zu einer verbesserten Mensch-Agenten-Interaktion führt. Aus der Literatur gewonnene Erkenntnisse zu Interpretationsansätzen menschlichen Kopfnickens in der Mensch-Mensch-Interaktion sollen dahingehend untersucht werden, ob sie sich in messbarer Ausprägung innerhalb von Mensch-Agenten-Interaktionen wiederfinden. Das Ziel dabei ist, eine fundierte Aussage treffen zu können, ob und inwiefern entsprechende automatische Interpretationsansätze für den praktischen Einsatz zu empfehlen sind.

1.3 Aufbau der Arbeit

Im Rahmen dieser Arbeit soll ein Kopfnick-Erkennungssystem entwickelt und auf Interpretationsansätze hin untersucht werden. Somit werden mit dem Kapitel 2 zuerst grundlegende relevante Methoden der Bildverarbeitung und Verfahren zur automatischen Detektion erklärt.

Neben den technischen Aspekten der Bildverarbeitung und Gestenerkennung wird anschließend im Kapitel 3 das menschliche Kopfnicken hinsichtlich der sozialen Bedeutsamkeit und der Rolle in der zwischenmenschlichen Kommunikation erörtert. Verschiedene Erkenntnisse über die Interpretation von Kopfnicken und dessen funktionelle Eigenschaften werden aus aktuellen wissenschaftlichen Untersuchungen herausgearbeitet. Die gewonnenen Erkenntnisse sollen dahingehend untersucht werden, ob sie derart automa-

tisiert erkannt und verarbeitet werden können, so dass ein virtueller Agent daraus einen für den Kontext relevanten Informationsgewinn erzielt.

Kapitel 4 beschreibt das Anwendungsszenario und die aufgezeichnete Datenbasis. Als Datenbasis dienen dabei Aufzeichnungen von Nutzerstudien eines virtuellen Agenten für Menschen mit Unterstützungsbedarf. Sämtliches Nickverhalten der Nutzer wurde nicht durch Instruktionen beeinflusst. Außerdem wird der in den Studien verwendete virtuelle Agent 'Billie' erläutert. Die Aufzeichnungen werden bezüglich des beobachtbaren Nickverhaltens annotiert und bieten somit die Datengrundlage für praktische Tests des entwickelten Detektionssystems.

Die Entwicklung eines Systems zum Erkennen von Kopfnicken wird in Kapitel 5 in mehreren Schritten aufgezeigt. So wird eine Support Vector Regression (SVR) zur Kopfwinkelschätzung trainiert. Mithilfe von Merkmalen aus der Kopfinkelschätzung werden zwei Verfahren zur Kopfnickerkennung umgesetzt. Zum einen wird ein neuer Algorithmus auf Basis von Dynamic Time Warping (DTW) entwickelt und zum anderen eine Support Vector Machine (SVM) als alternatives Modell zum Erkennen von Kopfnicken beschrieben und genutzt.

Das darauffolgende Kapitel 6 befasst sich anschließend mit der Auswertung und Performanz des im vorigen Kapitel entwickelten Systems. Hierzu werden Problematiken von gängigen Auswertungsverfahren diskutiert und ein aussagekräftiges Evaluationsverfahren ausgearbeitet. Dieses soll an dem entwickelten System angewendet werden.

Anschließend werden im Kapitel 7 Ansätze zur automatischen Interpretation erörtert und an den vorhandenen Daten getestet. Die Ergebnisse und Folgerungen werden schließlich im Kapitel 8 zusammengefasst und diskutiert.

2 Grundlagen der automatischen Kopfgestenerkennung

Zum besseren Verständnis dieser Arbeit werden im Folgenden grundlegende Techniken aufgezeigt, die für die Entwicklung eines Erkennungssystems für Kopfgesten Anwendung finden. Hierfür wird zuerst der Begriff 'Geste' eingeordnet und prinzipielle Grundlagen zum typischen Ablauf von Gestenerkennung beschrieben. Daraus folgend ergeben sich weitere Abschnitte, die die einzelnen Teilbereiche der Kopfgestenerkennung behandeln. Im Abschnitt 'Sensorik' wird relevante Hardware vorgestellt, bevor Methoden zur Gesichtserkennung und der Extraktion von Merkmalen von Kopf und Gesicht näher erläutert werden. Nach der Vorstellung von Verfahren zur Kopfwinkelschätzung werden entsprechende Arbeiten zur Kopfnickdetektion ausführlich dargelegt. Wichtige und insbesondere für diese Arbeit relevanten Methoden werden abschließend detailliert eingeführt. Zu den einzelnen Verarbeitungsschritten wird jeweils auf aktuelle Literatur verwiesen.

2.1 Prinzipien der Gestenerkennung

Eine Geste ist nach Kurtenbach und Hulteen eine Bewegung des Körpers, die im Normalfall Information beinhaltet (Kurtenbach and Hulteen (1990)). Automatische Gestenerkennung ist somit die Technologie, aus sensorisch erfassten Bewegungsabläufen mittels Algorithmen deren Information zu entschlüsseln. Die Untersuchung von Bewegungsabläufen erfordert die Verarbeitung von Merkmalen mehrerer zeitlich direkt aufeinanderfolgender Einzelbilder. Bei der automatischen Gestenerkennung wird zwischen diskreten und nicht diskreten Gesten unterschieden. Nicht diskret sind Gesten mit einem nichtdiskreten Wert als wesentliche Eigenschaft wie das Zeigen auf einen variablen Punkt, der sich kontinuierlich verändern kann oder zum Beispiel das 'Größerziehen' eines Bildes auf dem Smartphone. Dagegen nennen Gesten sich diskret, wenn sie zeitlich begrenzt, voneinander abgrenzbar

sind und sie erst bei Beendigung der Ausführung erkannt werden können. Somit wäre beispielsweise ein Kopfnicken eine diskrete Geste.

Während zur Gestenerkennung auch Datenhandschuhe oder andere zum Beispiel mit Beschleunigungssensoren ausgestattete Geräte genutzt werden, beschränkt sich diese Arbeit auf die kamerabasierte Gestenerkennung, wie im Abschnitt 2.2 erläutert.

Allgemein kann das Vorgehen bei kamerabasierter Gestenerkennung in folgende Teilschritte aufgeteilt werden: Die Aufteilung orientiert sich an Folien einer Lehrveranstaltung https://diuf.unifr.ch/main/diva/sites/diuf.unifr.ch.main.diva/files/joomla_presentationCarrino.pdf, 2019.

Zuerst muss das entsprechende Körperteil detektiert werden. Dieses wird in den Folgebildern mit Tracking-Verfahren weiter verfolgt oder fortwährend neu detektiert, während zu jedem Einzelbild Merkmale extrahiert werden. Bezogen auf die Kopfgestenerkennung bilden somit zeitlich aufeinanderfolgende Bildsequenzen die Ausgangsbasis, auf denen beispielsweise mittels einer automatischen Gesichtserkennung die Position des Kopfes lokalisiert wird. In dem lokalisierten Bildbereich werden Merkmale extrahiert, die Information bezüglich der Kopfausrichtung beinhalten. Auf Basis der extrahierten Merkmale lassen sich zum Beispiel mit Verfahren des maschinellen Lernens Funktionen berechnen, die daraus die Kopfausrichtung schätzen. Als dritter Schritt kann die Abfolge derartiger Schätzungen über einen Zeitraum als Kopfgeste untersucht werden und eine Abgrenzung oder Segmentierung einer Geste innerhalb der Bilderfolge vorgenommen werden. Die Zuordnung zu einer bestimmten Geste wird von Klassifikationsverfahren durchgeführt. Zu den verbreitetesten Verfahren gehören hierbei Hidden Markov Models (HMM), Dynamic Time Warping (DTW), Entscheidungsbäume oder Support Vector Machines (SVM).

Im weiteren Verlauf der Arbeit werden insbesondere DTW und SVM/SVR verwendet, weshalb diese sowie weitere Verfahren im Abschnitt 2.7 näher beschrieben werden.

2.2 Sensorik zur Erfassung von Kopfgesten

Bei der Realisierung eines Projektes im Bereich der Kopfgestenerkennung steht man anfangs vor der Wahl einer geeigneten Sensorik. Für die Regis-

trierung und Erfassung von Kopfbewegungen werden verschiedene Sensoren verwendet, die je nach Anwendungsgebiet verschiedene Anforderungen erfüllen müssen.

In den überwiegenden Fällen werden bildbasierte Ansätze verwendet. Dabei kommt meist ein Fotosensor mit zweidimensionaler Bildaufzeichnung zum Einsatz. Dieser ist je nach Qualität sehr günstig und bereits in vielen Alltags-Geräten integriert, sodass für viele Anwendungsszenarien keine zusätzliche sensorische Hardware nötig ist. Beispiele aus dem Konsumenten-Bereich sind Webcams oder Digitalkameras. Neben 2D-Kameras können auch Tiefenbild-Kameras verwendet werden. Diese können die Tiefe als zusätzliche Raumdimension erfassen und somit charakteristische Unebenheiten wie die Nase und die nach hinten verlaufene Wangenregion als Information nutzen. Stereokameras verrechnen dabei zwei Bilder, die aus zwei räumlich versetzten 2D-Kameras aufgenommen wurden, zu einer 3D-Datenwolke. Bei anderen 3D-Systemen wird ein charakteristisches Muster per Infrarot in den Aufnahmebereich projiziert. Aus den perspektivischen Verzerrungen, die aufgrund der Tiefe entstehen, wird das dreidimensionale Bild errechnet. Beispiele sind ASUS Xtion (ASUSTeK (2019)) und Microsoft Kinect 3D Kamera (Microsoft (2019)).

Eine jüngere Technologie zur Aufzeichnung von Tiefenbildern wurde durch TOF (time-of-flight)-Kameras (Iddan and Yahav (2000)) realisiert. Diese Kameras verwenden Photomischdetektoren, auch PMD-Sensoren genannt, die nach dem Laufzeitverfahren funktionieren. Hierbei wird ein Lichtsignal ausgesendet und die Zeit bis zur Registrierung dessen Reflektion gemessen. Durch die Laufzeit des Lichtes kann direkt auf die Entfernung des reflektierten Gegenstandes geschlossen werden.

3D-Kameras schaffen mit Erfassung der zusätzlichen Raumdimension mehr Spielraum für Merkmalsextraktion. Jedoch sind sie noch relativ teuer, weniger verbreitet und die Verarbeitung der Aufnahmedaten oft rechenlastiger als bei 2D-Kameras.

Ein Spezialfall sind Eyetracker-Module. Diese sind darauf spezialisiert, mittels Infrarot-Signal die Reflektion an den Augen zu lokalisieren, um das Blickverhalten zu erfassen. Das Wissen über die Augenpositionen lässt auch Rückschlüsse auf die Kopfbewegungen zu. Jedoch muss dabei beachtet werden, dass sich Kopfbewegungen so nicht von Ganzkörperbewegungen unterscheiden lassen. Ein weiterer Nachteil ist der kleine Sichtkegel und ebenfalls hohe Kosten.

Nicht bildbasierte Ansätze wären zum Beispiel Messgeräte, die direkt am Kopf montiert werden und durch Lage- und Gyroskop-Sensoren die Kopfausrichtung messen. Ein großer Nachteil ist hierbei jedoch, dass derartige Messgeräte aktiv vom Nutzer getragen werden müssen und oft eine Initialisierung benötigen. Beispiele sind das "head position sensor system MINDS"(Advanced Safety Concepts, Inc.) von Heitmann et al. (2001), sowie Head-Mounted Devices (Walsh and Daems (2015)).

2.3 Gesichtserkennung

Techniken zu automatischer Erkennung von Gesichtern auf Bildmaterial sind vor allem in der Vision-basierten Mensch-Maschine-Interaktion unabdingbar. Während bei uns Menschen die Wahrnehmung von Gesichtern schon seit dem Säuglingsalter intuitiv erfolgt [Simion and Di Giorgio (2015)], stellt die Gesichtserkennung ein technisches System vor vielfältige Herausforderungen. So treten menschliche Gesichter mit variabler Gesichtsgröße, Position und Orientierung im Bild auf und ändern ihr Erscheinungsbild gemäß unterschiedlicher Lichtverhältnisse und teilweisen Verdeckungen. Ein Detektionsverfahren muss daher Merkmale finden und verarbeiten, die robust gegen viele Störfaktoren sind. Wichtige Aspekte sind dabei das Erreichen einer hohen Treffsicherheit und aufgrund vieler Echtzeit-Anwendungen eine geringe Rechenlaufzeit. Wird ein Gesicht erstmal lokalisiert, kann der Bildausschnitt eine Grundlage für weitere Verarbeitungsschritte wie Kopfwinkelschätzung, Gesichtsanalyse, Gestenerkennung oder Blickverhalten bieten.

M. Rizvi gliederte Algorithmen zur Gesichtserkennung in vier Kategorien, die im Folgenden kurz beschrieben werden (Rizvi (2011)):

Die sogenannten Wissensbasierten Methoden basieren auf das Festlegen von Regeln. Es wird menschliches Vorwissen über das Erscheinungsbild eines typischen Gesichts verwendet und versucht, diese Gesetzmäßigkeiten und Zusammenhänge möglichst exakt zu definieren.

Merkmalsinvariante Ansätze haben dagegen das Ziel, strukturelle Merkmale im Bild zu finden, die auch dann wiederzufinden sind, wenn sich Ausrichtung und Blickwinkel ändern oder die Lichtverhältnisse variieren. So lassen sich Bildelemente wie Gesichter auch in verschiedenen Erscheinungsformen lokalisieren.

Eine weitere Gruppe von Detektionsansätzen sind Template-Matching-

Verfahren. Dabei wird eine Datenbank mit prototypischen Gesichtern oder Sets von Gesichtsausschnitten angelegt. Ein Suchfenster läuft dann durch ein Bild und gleicht Bildbereiche mit den Templates ab. Ein Ähnlichkeitsmaß zu vorhandenen Templates der Datenbank wird berechnet und eine Entscheidung getroffen, ob es sich um ein Gesicht handelt.

Die letzte der vier Kategorien bilden die Erscheinungsbasierenden Methoden. Sie verfahren ähnlich zu Template-Matching-Verfahren. Jedoch werden die Prototypen oder prototypische Beschreibungen statistisch begründet oder mit Maschinellen Lernverfahren erstellt, um einen repräsentativen Merkmalsraum zu bestimmen. Merkmale von Bildausschnitten werden dann in den Merkmalsraum projeziert, um Gesichter zu detektieren.

Die im Jahre 2001 von Paul Viola und Michael Jones entwickelte Viola-Jones-Methode war das erste effiziente Verfahren, das in Echtzeit demonstriert wurde und weltweite Berühmtheit erlangte (Viola, Paul and Jones (2001)). Dabei wurde eine neue Bildrepräsentation namens Integralbild eingeführt, die schnelle Merkmalsberechnungen ermöglicht. Bei einem Lernalgorithmus der AdaBoost-Variante ist jedes Merkmal ein schwacher Klassifikator. Diese werden in Form einer Kaskade kombiniert, wodurch Nicht-Gesicht-Regionen schnell verworfen werden können und somit die Effizienz gesteigert wird.

Eine besonders effiziente und mittlerweile häufig verwendete Methode wird mit der Verwendung von HOG-Merkmalen realisiert (Rekha and Kurian (2014)). Hierbei werden als Merkmale die Intensitätsverteilung und Anordnung der Kanten genutzt. Dazu werden Orientierungen von zusammengefassten Pixelbereichen gemittelt und in Histogrammen kodiert. Die Histogramme bilden einen Merkmalsraum, in dem versucht wird, einen Unterraum für Gesichter abzugrenzen.

Aktuell gewinnen Deep Learning basierte Ansätze zumehnend an Popularität. Diese kommen ohne eigens definierte Merkmalsberechnung aus. Stattdessen dient das gesamte Bild als Eingangs-Vektor eines neuronalen Netzes, in dem es durch verschiedene spezialisierte Zwischen-Layer durchgereicht wird und jeweils die Dimension reduziert wird. Der letzte Layer liefert beispielsweise nur noch zwei Ausgangswerte, die als ”wahründ ”falsch”mit Beispielen trainiert werden können. Ein Nachteil dieser Verfahren liegt immer noch in dem sehr hohen Ressourcenverbrauch, der durch Nutzung von Grafikkarten und Parallel Computing zuweilen angegangen wird. Weitere Nachteile sind die Notwendigkeit von sehr vielen Trainingsdaten, die komplexe

Justierung der Netz-Parameter sowie die teilweise kaum zu durchschauenden Prozesse innerhalb der Zwischen-Layer.

2.4 Merkmale von Kopf und Gesicht

Durch unsere Kopfhaltung und Gesichtsmimik geben wir viel von unserem emotionalen Zustand preis. Oft ist es einem Menschen direkt anzusehen, ob er beispielsweise Freude, Wut, Gelassenheit oder Angst empfindet. Dafür sorgen eine Vielzahl an Muskeln, die ein Lächeln oder ein Zusammenziehen der Augenbrauen herbeiführen. Wird ein Gesicht erst von einem Gesichtsdetektor erkannt und lokalisiert, bietet dies die Grundlage für eine automatische Gesichtsanalyse. Der naheliegende nächste Schritt wäre somit, weitere Merkmale innerhalb des Gesichts, wie zum Beispiel Mundwinkel, Nase und Augen korrekt zu lokalisieren, um weitere Informationen über die Intention oder den kognitiven Zustand der Person zu erlangen.

Ein Ansatz besteht darin, ein Alignment-Modell, wie in der Abbildung 2.1 dargestellt, für markante Gesichtspunkte zu bestimmen. Dabei werden Landmarken definiert, die üblicherweise jedes menschliche Gesicht aufweist und die in einem geometrischen Zusammenhang zueinander stehen. Ein Alignment-Modell ermöglicht die Annotation von Trainingsbildern für das Training von Klassifikatoren, um entsprechende Landmarken in beliebigen Gesichtern wiederzufinden. Ein sehr schnelles Lösungsverfahren zum Problem des Gesichts-Alignments gelang V. Kazemi und J. Sullivan mithilfe eines Ensembles von Regression Trees (Kazemi and Sullivan (2014)).

Ein System zur Beschreibung von Gesichtsausdrücken publizierten Ekman und Friesen bereits im Jahre 1978 und nannten dieses Gesichtsbewegungs-Kodierungssystem Facial Action Coding System (FACS) (Ekman and Friesen (1978)). Darin sind 44 Bewegungseinheiten beschrieben, denen bei Verwendung eine Ausprägungsstärke zwischen A (schwach) und E (stark) zugewiesen wird. Dieses Kodierungssystem wird häufig bei der Entwicklung von Klassifikatoren zur Erkennung von Gesichtsausdrücken verwendet (Valstar and Pantic (2006), Senechal et al. (2015)).

Die Lokalisierung von Gesichtsmerkmalen kann auch Aufschluss über die Blickrichtung oder Kopfausrichtung geben. Mit der Auswertung über mehrere Folgebilder in einer Videosequenz können ganze Bewegungsabläufe und Gesten untersucht werden.

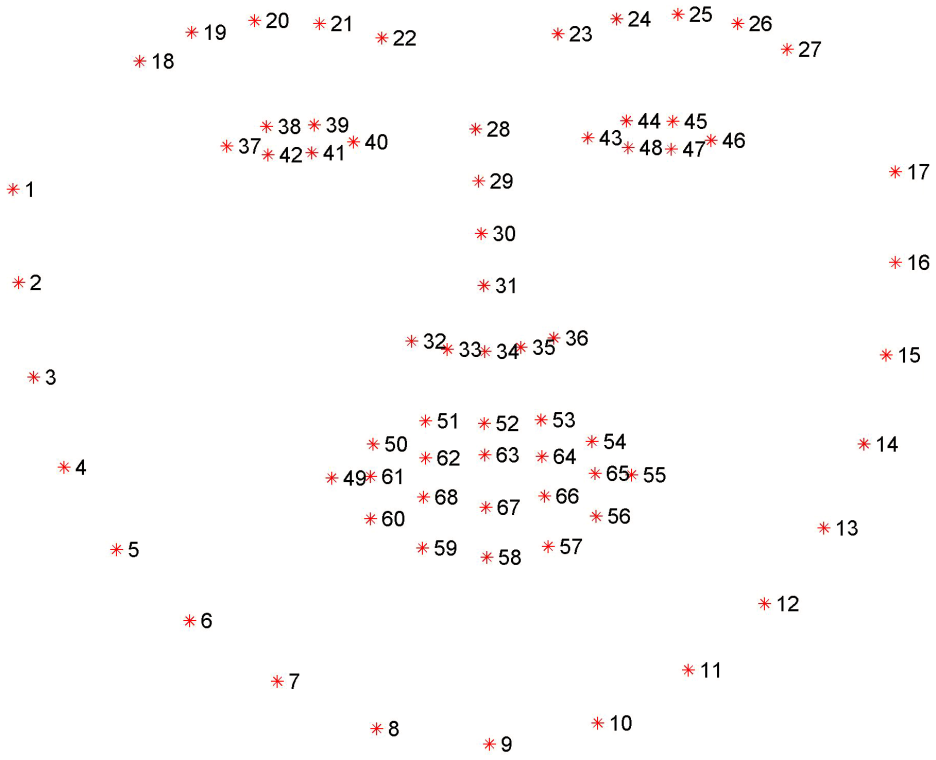


Abbildung 2.1: Ein Alignment-Modell für markante Gesichtspunkte, wie es auch von Kazemi and Sullivan (2014) verwendet wurde. Die Grafik wurde im Jahr 2017 folgender Webseite entnommen: https://ibug.doc.ic.ac.uk/resources/300-W/

2.5 Kopfwinkelschätzung

Die Sinne des Menschen, insbesondere Augen, Mund und Ohren, sind so angelegt, dass mit der Kopfausrichtung eines Menschen auch der Fokus der Aufmerksamkeit mitbestimmt wird. Wendet sich der Gesprächspartner ab und schaut am Sprecher vorbei, kann dies somit als ein Zeichen von Desinteresse oder Ablenkung gewertet werden. Auch die sich verändernde Kopfausrichtung und Bewegungen über die Zeit bilden Trajektorien, die untersucht und als Kopfgesten eingeordnet und verstanden werden können.

Die Kopfausrichtung ist eng gekoppelt mit der Blickrichtung (Johnson and Cuijpers (2013)). So gilt in den meisten Fällen, dass eine bereits bekannte Kopfausrichtung auch eine grobe Schätzung der Blickrichtung zulässt. Schließlich wird bei Bewegung des Kopfes das Koordinatensystem der möglichen Blickrichtungen ebenfalls mitbewegt.

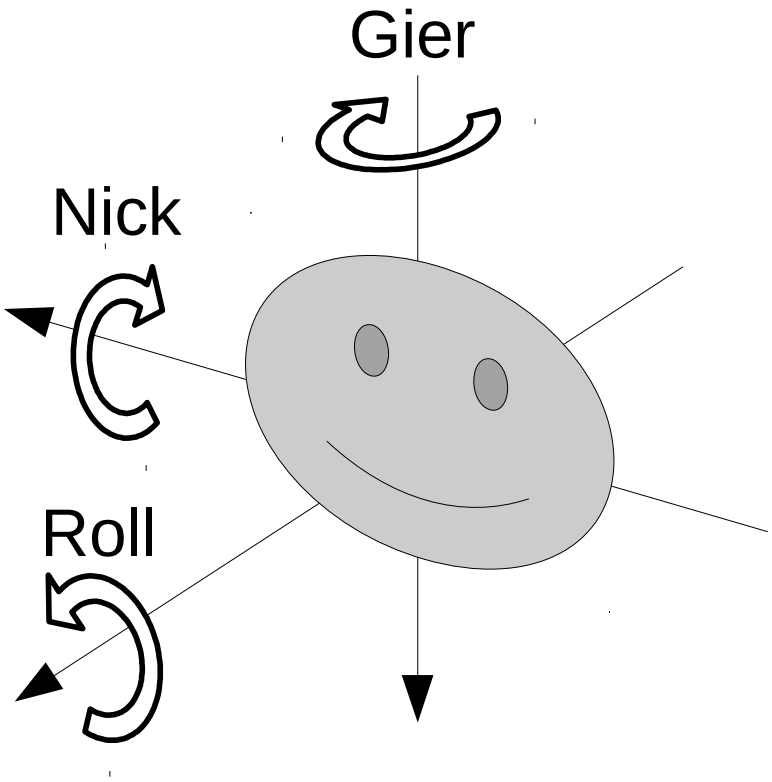


Abbildung 2.2: Die drei möglichen Drehachsen des menschlichen Kopfes: Roll, Nick und Gier.

Die Anatomie der Halswirbelsäule wird im Kapitel 3.1 näher beschrieben und ermöglicht Bewegungen in drei Achsen. Die entsprechenden Kopfwinkel werden im Folgenden gemäß der Abbildung 2.2 Roll, Nick und Gier genannt.

Aufgrund der überwiegend vertikalen Bewegungsabläufe beim Kopfnicken, ist in unserem Fall vor allem der Nick-Winkel relevant.

Ein automatisches Schätzsystem für die Kopfausrichtung besteht in der Regel aus einer Sensorik, die den Kopf bzw. Kopfmerkmale in zeitlicher Abfolge erfasst, sowie aus einer Funktion, die aus den gemessenen Merkmalen Schätzwerte für die Kopfwinkel berechnet.

Zur Umsetzung automatischer Schätzsysteme für die Kopfausrichtung existieren viele Ansätze, von denen einige im Folgenden herausgegriffen werden.

Zu den simpleren Methoden gehört die Verwendung von Optical Flow (Horn and Schnuck (1981)), welches oft mit weiteren Verfahren kombiniert wird (Zhu and Fujimura (2003)). Bei dessen Anwendung handelt es sich nicht direkt um eine Schätzung des Kopfwinkels, sondern um die Bewegungsänderung zwischen aufeinanderfolgenden Bildern. Somit handelt es sich bei diesem Verfahren um eine Schätzung der Änderung des Kopfwinkels innerhalb eines Zeitfensters. Es lassen sich so Bewegungen und Gesten erfassen. Je nach Wahl der Merkmale kann jedoch nicht immer berücksichtigt werden, ob sich der gesamte Körper gleichermaßen mitbewegt hat und der Kopf relativ zum Körper möglicherweise gar keine relevante Bewegungskomponente vorweist. Wird beispielsweise lediglich die Gesichtsposition verfolgt, so wird sowohl bei einer Kniebeuge als auch bei einem Kopfnicken jeweils eine Ab-Auf-Bewegung erfasst. Zudem sind bei Bewegungserfassung aufgrund der Differenzbildung immer mindestens zwei Bilder mit zeitlicher Relation nötig, um eine Schätzung zu berechnen.

Ähnliches gilt für Tracking-Verfahren. So kombinierte W. Yu zur Kopfwinkelschätzung Tracking von Landmarken mit dem Kalman-Filter (Yu and Gang (2011)). Initial ist beim Tracking die Lokalisierung eines zu verfolgenden Bildbereichs nötig. Das kann der gesamte erkannte Gesichtsbereich oder beispielsweise die Nasenspitze sein. Auch hier handelt es sich um eine Schätzung der Änderung inklusive der Nachteile des Optical-Flow Verfahrens. Ein Vorteil ist jedoch, dass die Änderungswerte nahezu pixel-genau präzise und rauscharm berechnet werden können, obgleich auch hier aufgrund der Verarbeitung der Positionsänderungen im 2D-Bild nicht unterschieden werden kann, ob die Person tatsächlich nur den Kopf oder den Körper als Ganzes bewegt.

Mit anderen Methoden wird versucht, einen möglichst anatomisch korrekten Kopfwinkel direkt aus einem Einzelbild zu bestimmen. Da hier eine

Missdeutung von Translation des gesamten Körpers wegfällt, können zuverlässigere Ergebnisse erwartet werden. So können mithilfe von maschinellen Lernverfahren Klassifikatoren trainiert werden: Anfangs wird eine Aufteilung mit Ausrichtungsintervallen vorgenommen, die zusammengefasst werden, um den Zielraum zu diskretisieren (Zabulis et al. (2009)). Jedes Intervall ist nun eine Klasse, für die ein separater Klassifikator trainiert werden kann. Mit der Anzahl der Klassen wird somit durch die Diskretisierung auch die maximal erreichbare Genauigkeit festgelegt.

Diese Einschränkung wird mit Regressionsverfahren aufgehoben, die mit einem kontinuierlichen Zielraum funktionieren und so durch funktionales Mapping die erreichbare Genauigkeit zunächst nicht einschränken (Murphy-Chutorian and Trivedi (2007)). Bei Regressionsmethoden wird eine Funktion berechnet, die eine Abbildung von einem beliebigen Punkt im Merkmalsraum zu einem geschätzten Ausgabewert, in diesem Fall Winkel, darstellen soll. Diese Funktion wird ebenfalls anhand von Beispiel-Bildern mithilfe von bereits bekannten Ausgabewerten trainiert. Im Vergleich zu Klassifikatoren muss hier jedoch der gesamte Zielraum mit nur einem Modell abgebildet werden, während Klassifikatoren auf ihren jeweils definierten Bereich spezialisiert sind.

In einer Untersuchung verschiedener Verfahren zur Kopfwinkelschätzung stellte E. Murphy-Chutorian 2009 (Murphy-Chutorian and Trivedi (2009)) weitere Methoden heraus, wie Erscheinungsbasierte Methoden, die auf Basis von Vergleichen mit Bildern bekannter diskreter Kopfausrichtung arbeiten. Zudem sind auch mit geometrischen Ansätzen, die Abstände zwischen Gesichtsmerkmalen als Merkmalsbasis verwenden, erfolgreiche Publikationen hervorgegangen, wie Canton-Ferrer 2006 (Canton-Ferrer (2006)) bewies.

G. Guo evaluierte 2008 die Leistungsfähigkeit von Klassifikation mittels SVM und Regression mittels SVR für vertikale und horizontale Kopfwinkelschätzung (Guo et al. (2008a)), sowie eine vielversprechende Kombination von SVM und SVR namens LAAR (Guo et al. (2008b)). Während SVR sowohl im vertikalen als auch im horizontalen Bereich gute Ergebnisse lieferte, zeigten SVMs zwar im vertikalen Bereich minimal bessere, im horizontalen Bereich jedoch deutlich schlechtere Ergebnisse. Durch LAAR wurde im Vergleich zur SVR keine nennenswerte Leistungssteigerung erreicht. Anzumerken sei hier jedoch, dass alle Bilder des Evaluationsdatensatzes (Zabulis et al. (2009)) als Ground Truth einer von 9 vertikalen und 13 horizontalen Klassen zugewiesen sind und der Datensatz somit auf Klassifikation speziali-

siert ist. Da mit der SVR eine kontinuierliche Funktion trainiert wird, waren somit für große Datenbereiche keinerlei Trainingsdaten vorhanden, weshalb ich die SVR trotz der knappen Niederlage im vertikalen Bereich dennoch als leistungsstärkeres Verfahren in diesem Anwendungsbereich sehe.

2.6 Relevante Arbeiten zu Kopfnickdetektion

Die Erforschung automatischer Erkennungssysteme für Kopfnicken begann schon vor über 20 Jahren. Da sich bei einer Kopfnick-Bewegung jegliche Landmarken im Gesicht mitbewegen, bilden Tracking-Verfahren von Gesichtsmerkmalen oft die Grundlage zur Kopfnick-Detektion (Kawato and Ohya (2000), Li and Danielsen (2007)). Vor allem in den frühen Arbeiten wurden häufig Optical-Flow-Varianten genutzt, um Kopfbewegungen zu erfassen. Bei der Wahl geeigneter Klassifikations-Methoden für Kopfnicken lassen sich im Laufe der Jahre Trends erkennen, wie in diesem Kapitel deutlich wird. Lange Zeit galten Hidden-Markov-Modelle als Standard-Verfahren, ab etwa 2012 wurden eher Support-Vector-Machines genutzt und aktuell gewinnen neuronale Netze mit Deep-Learning-Methoden zunehmend an Popularität.

Bereits im Jahre 1995 wurden auf der erstmals stattfindenden Konferenz namens *International Workshop on Automatic Face and Gesture Recognition* Systeme vorgestellt, die ein Tracking von Gesichtsmerkmalen mit einer Verarbeitungsgeschwindigkeit von mehreren Frames pro Sekunde leisten konnten (Jacquin et al. (1995)). Im darauf folgenden Jahr veröffentlichte A. Zelinsky auf derselben Konferenz eine Arbeit, die unter anderem einen in 30 Frames pro Sekunde lauffähigen Algorithmus zum Erkennen einer 'Ja'-Geste beinhaltete (Zelinsky and Heinzmann (1996)). Das Kernstück war dabei das MEP Tracking Vision System von FUJITSU (`http://users.cecs.anu.edu.au/~rsl/hrintact/vishard.html`). Diese Hardware leistete Objekt-Tracking mittels Template Korrelation bei 30Hz und einer NTSC-Auflösung. Zur zuverlässigen Anwendung für Gesichts-Tracking wurden Kalman Filter (Kalman (1960)) verwendet, um das Tracking-Ergebnis mit einem geometrischen Gesichtsmodell zu kombinieren. Eine Geste wurde als Aneinanderreihung von atomaren Bewegungs-Aktionen definiert, die probabilistisch bestimmt wurden. Eine 'Ja'-Geste wurde dabei als Aktionen-Kette wie 'Kopf hoch', 'Stopp', 'Kopf runter' repräsentiert. Die Zuverlässigkeit der Erken-

nungsergebnisse wurde jedoch nicht evaluiert.

Mit zunehmend leistungsfähigeren Computern wurden auch unabhängig von komplexer spezieller Hardware lauffähige Erkennungssysteme möglich. So gelang S. Kawato (Kawato and Ohya (2000)) ein simplerer Ansatz auf Basis eines einzelnen Farbvideo Streams. Es wurde ein Detektor für die Augen-Zwischenregion 'The Between Eyes' auf Basis eines kreisförmigen Frequenz-Filters entwickelt. Nach erfolgreicher Detektion wird der Bildbereich anschließend als extrahiertes Template getrackt, womit ohne spezielle Hardware eine Tracking-Geschwindigkeit von 13 Frames pro Sekunde erreicht wurde. Die Entscheidung, wann Kopfnicken erkannt wird, wird nach einer simplen regelbasierten Abfrage von aufeinanderfolgenden Auf- und Abbewegungen entschieden.

Neben der Augen-Zwischenregion wurden die Augen selbst als Merkmalsbereich wiederholt in Arbeiten aufgegriffen. Zum Beispiel nutzte A. Kapoor (Kapoor and Picard (2001)) den aus der Blitzfotografie bekannten Rote-Augen-Effekt mittels pulsierender IR-LEDs und einer Kamera ohne IR-Filter. Das Differenzbild zwischen den zusätzlich beleuchteten und nicht zusätzlich beleuchteten Bildern zeigte die nunmehr leicht zu detektierenden Pupillen. Diese wurden getrackt und die über die Zeit gesammelten Daten geglättet. Die Bewegungsrichtungen wurden mit einer symbolischen Repräsentation kodiert und auf Basis der aufeinanderfolgenden Daten ein Hidden Markov Model (HMM) trainiert. Dieses Modell erreichte ebenfalls Echtzeit-Geschwindigkeit und wurde mit einem Korpus von 25 Kopfnick-Beispielen trainiert. Diese entstammen jedoch von instruierten Nutzern, die auf Fragen eines virtuellen Agenten mittels Kopfbewegung antworten sollten. Mit der IBM-Pupilcam (Morimoto et al. (2000)) verwendete J. W. Davis (Davis and Vaks (2001)) vergleichbare Hardware. Jedoch kam statt eines HMMs ein Finite State Machine Modell zum Einsatz. Im Rahmen dieser Veröffentlichung wurde auch ein Anwendungsszenario mit einem 'dialog-box agent' vorgestellt, wobei alternativ zum Klicken eines Buttons eine Kopfgeste als Zustimmung oder Ablehnung als Eingabemöglichkeit gegeben war.

Ein ähnlicher Ansatz wird von Tan and Rong (2003) verwendet, wobei hier statt Infrarot eine Kamera mit visuellem Frequenzspektrum genutzt wird. Es werden zuerst das Gesicht und die Augen des Nutzers detektiert. Daraufhin dienen die relativen Änderungen der Augenpositionen als Merkmale für Kopfnicken, mithilfe derer ebenfalls ein HMM trainiert wird. Die Trainingsdaten bestehen aus 37 Kopfnick-Beispielen, die von instruierten

Nutzern gesammelt wurden.

S. Fujie entwickelte ein Erkennungssystem auf einer Roboterplattform eines Konversations-Roboters, um Kopfgesten als para-linguistische Information nutzen zu können (Fujie et al. (2004)). Methodisch kommen ebenso Optical-Flow und HMM zum Einsatz. Eine Besonderheit ist die Nutzung zweier Bereiche für OpticalFlow. So wird zum einen wie bisher üblich die Kopfregion getrackt, aber separat auch unterhalb des Kopfes der Körperbereich. Somit können Bewegungen, die nicht von dem Nutzer, sondern von der Roboterkamera verursacht wurden, herausgerechnet werden. Anstatt sich ausschließlich auf die Augen als Orientierungspunkt zu verlassen, kann Optical Flow verwendet werden, um Gesichtsbewegungen zu erkennen.

In einer Arbeit von C. Lee wurde wieder Infrarot-Hardware zur Pupillendetektion verwendet (Lee et al. (2004)). Jedoch wurden diese durch die Zuhilfenahme einer Stereo-Kamera erweitert, mit welcher mittels eines 'adaptive view-based appearance'-Modells (Morency and Darrell (2003)) ein zusätzliches Kopf-Tracking-Verfahren beigesteuert wurde. Es wurde auch ein HMM mit 11 Leuten trainiert. Das Experiment wurde mit einem Nickverhalten eines Konversations-Roboters ergänzt, um Nicken bei Nutzern zu provozieren.

L. Nguyen setzte bei der Entwicklung eines Nickerkenners auf einen multimodalen Ansatz (Nguyen et al. (2012)). Aufbauend auf Gesichtsdetektion, Tracking des Kopfbereiches und Berechnung der Geschwindigkeitsvektoren gemäß Optical Flow, wurden die Geschwindigkeitswerte einer Fourier-Transformation unterzogen. Auf Basis dieser Merkmale wurde mittels einer SVM ein Nick-Klassifikator trainiert. Zusätzlich wurde die Sprechaktivität des Nutzers binär annotiert und dabei festgestellt, dass 92% des Nickvorkommens keine Sprechaktivität beinhaltet. Wurde statt des Trainings einer einzigen SVM sowohl für Nickbeispiele mit Sprechaktivität als auch ohne Sprechaktivität jeweils eine separate SVM trainiert, stiegen die Erkennungsraten. So konnte der berechnete F1-Score von 0.559 bei ausschließlich visuellen Merkmalen zu 0.628 bei multimodalen Merkmalen signifikant erhöht werden.

S. Chu entwickelte eine Schnittstelle für Digitalkameras, die es ermöglicht, diese zur Selbstportrait-Aufnahme mittels Kopfnicken auszulösen (Chu and Tanaka (2012)). Zur Anwendung kam Template-Matching von Merkmals-Punkten zur Gesichtsdetektion und ein Lucas Canade Optical Flow Ver-

fahren (Lucas and Kanade (1981)). Um Fehlmessungen durch starke, nicht durch Kopfgesten verursachte Bewegungen zu vermeiden, wurde eine Safe-Zone eingeführt, sodass nur ein Bereich um das Gesicht herum berücksichtigt wurde.

H. Wei verwendet in Wei et al. (2013) einen Kinect-Sensor zur Schätzung der Kopfausrichtung. Eine symbolische Bewegungsrichtung wird auf der Grundlage von Neigungs- und Gier-Änderungen berechnet. Mithilfe von 150 Stichproben auf der Grundlage von Richtungssymbolen wird ein HMM zum Training eines Klassifikators für Kopfnicken verwendet.

J. Terven (Terven et al. (2014)) entwickelte ein System zur Erkennung von Kopfnicken mit einer "complete HMM". Dabei handelt es sich um eine Weiterentwicklung des HMMs, in der alle mögliche Zustände jederzeit berücksichtigt werden und somit auch Gesten mit kleinen Ausreißern nicht so schnell ausgeschlossen werden. Die Schätzung der Kopfausrichtung erfolgt mithilfe einer SDM (Xiong and Torre (2013)) mittels Detektion von Gesichts-Landmarken. Auf die Kopfausrichtung wird nun geschlossen, indem die Anordnung der erkannten Landmarken mit den Soll-Werten eines 3D-Modells abgeglichen werden. Zur Evaluation wurde eine Interaktion mit einer Software initiiert, in der die Nutzer Fragen mit Mausklick auf Antwort-Buttons beantworten sollten. Anschließend wurden sie aufgefordert, diese Antwort in Form einer Kopfgeste zum Ausdruck zu bringen, wodurch Erkennungsraten von bis zu 98.5% erreicht wurden.

Ein besonders ausgeklügelter Ansatz wird in Chen et al. (2015) vorgestellt: Ein mit Bildern eines Kinect-Sensors arbeitender 3D-Kopf-Tracker dient zum Schätzen der Rotationsmatrix des Gesichts. Innerhalb eines Zeitfensters werden Frequenz- und Achsen-Merkmale extrahiert, die auf die Änderungen der Rotationsmatrix basieren. Zur Klassifikation von Nicken wird eine Support-Vektor-Maschine verwendet. Das Modell wird mit 543 Nick-Stichproben trainiert, die aus einem Korpus stammen, der Konversationsinteraktionen wie beispielsweise Vorstellungsgespräche enthält. Mit einem F-Score von etwa 70% und einer aussagestarken Evaluation, gehört dieser Nickerkenner wohl zu den leistungsstärksten überhaupt.

K. Gopakumar nutzte ebenfalls eine Support-Vector-Machine, nachdem mit AdaBoost das Gesicht erkannt und die Augen getrackt werden (Gopakumar and Suni (2016)). Als Videostream wurden hochauflösende Aufnahmen von 1280x1024x9fps verwertet. Die sehr hohe Erkennungsrate für Kopfnicken von über 93% entstammen jedoch nicht aus einem natürlichen

Gesprächskontext. Die Testdaten bestanden aus einer Menge von 15 Kopfnickgesten, 15 Kopfschütteln und 15 neutralen Kopfstellungen.

X. Xu veröffentlichte im Jahr 2017 erstmals ein System, das auf die Verwendung von globalen inklusive lokalen Convolutional Neural Network (CNN) Merkmalen zur Schätzung der Kopfausrichtung setzt (Xu and Kakadiaris (2017)). Im selben Jahr entwickelte S. Ota mittels künstlicher neuronaler Netze einen multimodalen Kopfnick-Erkenner auf Basis eines Kinect-3D-Sensors. Ota et al. (2017) Dabei wurde neben der Kopfrichtungs-Schätzung auch für den Sprachrythmus jeweils ein Neuronales Netz trainiert, die beide zu einem Erkennungsergebnis fusioniert werden. Die Evaluation wurde mit je zur Hälfte positiven und negativen Stichproben-Sequenzen durchgeführt, wodurch sich kaum auf die Leistung in einem natürlichen Gesprächsszenario schließen lässt.

Einen Ansatz mittels eines Rekurrenten Neuronalen Netzes (RNN) stellte E. Langholz vor (Langholz and Brasher (2018)). Verarbeitet wurde Bildmaterial einer 3D-Kamera eines Smartphones (IPhone X) und 60 Frames pro Sekunde. Trainingsdaten wurden per App aufgezeichnet, indem die Nutzer zur Ausführung einer Geste aufgefordert wurden. Dabei wurden auch verschiedene Zeitfenster festgelegt, in denen die Nutzer langsame oder schnellere Ausführungen der Geste aufzeichnen konnten. Die Daten wurden auf gleiche Längen und Amplituden gestreckt und neue Daten künstlich erzeugt. Mit diesen Daten wurde ein RNN-GRU trainiert und gezeigt, dass Kopfnickerkennung auch auf Smartphones angemessen funktioniert.

Die meisten der bestehenden Arbeiten verwenden Datenbanken mit instruierten Benutzern. Dies kann zu ausgeprägteren Nickmustern führen, die leicht zu erkennen sind. Darüber hinaus enthalten die Korpora oft eine relativ hohe Häufigkeit von Nicken im Vergleich zu Videomaterial ohne Nicken, was zu einer unnatürlich niedrigen Rate an Falsch-Detektionen führt. Eine ausführliche Behandlung zur Evaluation von Kopfnicken folgt im Kapitel 6.

In Bereichen wie der Gestenerkennung hat sich gezeigt, dass DTW im Vergleich zu HMMs bessere Ergebnisse erzielt (Carmona and Climent (2012)). Dennoch wurde in keiner der vorgestellten Arbeiten der Einsatz von DTW beschrieben oder gar empfohlen. Deshalb soll die Anwendung von DTW für die Kopfgestenerkennung in dieser Arbeit untersucht werden. Als alternativer Ansatz soll eine SVM als bewährtes Standardverfahren ebenfalls auf dieses Problem angewendet werden. Die zu verwendenden Verfahren werden im folgenden Abschnitt 2.7 erläutert.

2.7 Detektionsmethoden

Wie aus den vorherigen Abschnitten wiederholt hervorging, werden zur Abgrenzung und Segmentierung einer Geste innerhalb von Bilderfolgen verschiedene Klassifikationsverfahren eingesetzt. Einige wichtige Verfahren wie HMM und Entscheidungsbäume sollen im Folgenden kurz erläutert werden. Methoden, die im weiteren Verlauf der Arbeit Verwendung finden, werden umfangreicher erklärt. Dazu gehören insbesondere DTW und SVM/SVR.

Eine der einfachsten Verfahren zur Klassifikation stellen Entscheidungsbäume dar (Quinlan (1986)). Dabei wird ein gerichtetes Baumdiagramm erstellt mit einer Menge von Entscheidungen, die hierarchisch gegliedert sind. Die Entscheidungen werden anhand von festgelegten Regeln oder Wahrscheinlichkeiten getroffen. Diese können manuell festgelegt oder mit Verfahren des maschinellen Lernens hervorgebracht worden sein. Die unterste Entscheidungsebene kann somit als Schritt zur Klassifikation dienen.

Ein Hidden Markov Model (HMM) kann als eine stochastische Variante eines endlichen Automaten beschrieben werden (R. Rabiner (1989)). Der Begriff kommt von der Annahme einer Menge verdeckter Zustände, die nicht direkt beobachtbar sind. Diese gehen nach bestimmten Gesetzmäßigkeiten ineinander über. Beobachtbar sind jedoch lediglich deren Auswirkungen, oft durch extrahierte Merkmale beschrieben, aus denen man Übergangswahrscheinlichkeiten zwischen Zustands-Wechsel ableiten kann. Sind eine Reihe von Beobachtungen mit festgelegten Wahrscheinlichkeiten bekannt, lassen sich Wahrscheinlichkeiten für den tatsächlichen aktuellen Zustand bestimmen.

2.7.1 Dynamic Time Warping

Dynamic Time Warping ist ein Verfahren zur dynamischen Zeitnormierung (Sakoe and S. Chiba (1971)). Dabei wird ein Algorithmus beschrieben, mit dessen Hilfe Zeitserien unterschiedlicher Länge aufeinander abgebildet und somit vergleichbar gemacht werden können.

Dieser basiert auf der Grundidee einer nicht proportionalen Stauchung oder Dehnung zweier Zeitserien. Hierfür wird mit zwei Zeitserien so durch Pausieren oder Fortfahren einer der Zeitserien verfahren, dass die Summe der paarweisen Distanzen zwischen den beiden minimiert wird.

Die berechnete Distanz kann als Ähnlichkeitsmaß dienen. Dafür ist eine Distanz- oder Kostenfunktion nötig. Für zwei Zeitserien A und B mit den Längen M and N, beschreibt die Distanzfunktion $d(i, j)$ die Distanz zwischen $A(i)$ and $B(j)$.

Das Distanzmaß kann beliebig gewählt werden. Üblicherweise kommt die Euklidische Distanz oder die Mahalanobis-Distanz (Mahalanobis (1936)) zum Einsatz.

Es soll nun ein 'warping path' berechnet werden. Der 'warping path' ist der Weg mit der geringsten Gesamt-Distanz vom Anfang bis zum Ende beider Zeitserien.

Als Erstes werden alle möglichen Teil-Distanzen $d(i, j)$ in einer Kosten-Matrix C zusammengefasst.

Daraufhin werden alle Pfade von $C(0, 0)$ bis $C(M, N)$ evaluiert, um den optimalen Abgleich zwischen A and B zu bestimmen. Optimal bedeutet hier, dass die Summe der einzelnen Elemente minimal ist.

Mithilfe von dynamischer Programmierung (Bellman and Kalaba (1957)) kann sowohl die Berechnung der Distanz, als auch die Evaluation der Pfade effizient in einem einzelnen Schritt zusammengefasst werden.

Dies geschieht mittels einer akkumulierten Kostenmatrix K_A, die in der Gleichung 2.1 dargestellt wird.

$$C_A(i,j) = \begin{cases} d(0,0), & i,j = 0 \\ d(0,j) + C_L, & i = 0, j > 0 \\ d(i,0) + C_T, & i > 0, j = 0 \\ d(i,j) + min(C_L, C_{LT}, C_T), & i,j > 0 \end{cases} \tag{2.1}$$

wobei

$$\begin{aligned} C_L &= C_A(i, j-1), \\ C_{LT} &= C_A(i-1, j-1), \\ C_T &= C_A(i-1, j). \end{aligned}$$

Hier wird durch jede berechnete Distanz $C_A(i, j)$ gleichzeitig die Gesamt-Distanz des optimalen 'warping path's zu $C_A(0, 0)$ bereitgestellt.

Schließlich entspricht $C_A(M, N)$ der Distanz zwischen den beiden Zeitserien A und B.

Mit einem festgelegten Schwellwert zu einer prototypischen Zeitserie kann der DTW-Algorithmus somit zur Klassifikation verwendet werden.

Die Trainingsphase für eine bestimmte Anwendung von DTW zeichnet sich durch das Finden einer passenden prototypischen Zeitserie aus, sowie durch das Festlegen eines sinnvollen Schwellwertes zur Klassifikation.

Um geeignete Prototypen zu finden, bieten sich verschiedene Verfahren an. Oft lässt sich eine für eine bestimmte Klasse durchschnittliche Zeitserie synthetisch berechnen oder diejenige Zeitserie ausmachen, dessen Summe an Distanzen zu klasseneigenen Prototypen minimal ist. Besitzen die Zeitserien jedoch variable Längen, sind sie oft nur bedingt miteinander vergleichbar. Ein aufwendiges, aber ziemlich sicheres Verfahren ist das systematische Evaluieren einer favorisierten Gruppe oder aller potentieller Prototypen mithilfe eines jeweils ausgegrenzten Test-Datensatzes.

2.7.2 Support Vector Machine und Regression

Die Support Vector Machine ist ein mächtiger und bewährter Algorithmus im Bereich des Maschinellen Lernens um Abhängigkeiten zwischen Eingabe- und Ausgabedaten zu schätzen. Sie wurde von Vapnik und Chervonenkis als Lösungsverfahren für Mustererkennung entwickelt (Vapnik and Chervonenkis (1974)). Dabei widmet sich der Algorithmus folgender Problemstellung:

Es ist eine Datenmenge gegeben mit einer fixen Anzahl gemeinsamer Attribute. Jeder Datenpunkt ist einer von zwei Klassen zugeordnet. Die Annahme besteht nun darin, dass zwischen den Daten und deren Klassenzugehörigkeit ein Muster besteht, welches Aufschluss über die systematischen Relationen dieser Art von Eingabedaten liefert. Die Klassenzugehörigkeit ist im Allgemeinen nur für eine kleine Menge von Eingabedaten verfügbar bzw. bekannt. Das Ziel ist, diese Information in eine Funktion zu integrieren, die zu gleichartigen Eingabedaten eine identifizierende Ausgabe liefert, die generalisierende Eigenschaften aufweist.

Meist wird als Funktion eine Trennebene gesucht, die die beiden Klassen im Raum der Attribute optimal voneinander trennt. Optimal heißt in diesem Fall, dass der Abstand zwischen den nächsten zur Trennebene liegenden Daten im Raum der Attribute maximiert und somit die Anfälligkeit für Rauschen minimiert wird. Dadurch wird versucht, eine hohe Generalisierbarkeit zu erreichen, um auch für Datenpunkte mit unbekannter Klassenzugehörigkeit eine fundierte Abschätzung der Klassenzugehörigkeit zu liefern.

Linearer Fall

Zunächst wird angenommen, die Eingabedaten seien gemäß ihrer Klassenzugehörigkeit linear trennbar. Die Klassen werden mit positiv ("+") und negativ ("−") voneinander unterschieden. Es wird nun eine Trennebene gesucht, die positive von negativen Daten trennt.

Die Problemstellung entspricht nun einer linearen Klassifikation. Eine lineare Funktion lässt sich durch folgende Ebenengleichung beschreiben:

$$w^T * x + b = 0 \tag{2.2}$$

Dabei sind mit x beliebige Eingabedaten des Attributenraumes gemeint, b ist die Verschiebung der Ebene vom Ursprung und w beschreibt die Steigung der Ebene. Des Weiteren werden zwei Ebenen parallel zu der Trennebene definiert mit den Bedingungen:

$$\begin{aligned} w^T * x_i + b \geq +1, y_i = +1 \\ w^T * x_i + b \leq -1, y_i = -1 \end{aligned} \tag{2.3}$$

Die Klassenzugehörigkeit wird mit y angegeben.

Die kürzeste orthogonal zur Trennebene bestehende Entfernung zwischen positiven und negativen Daten mit minimaler Entfernung zur Trennebene wird Margin (Rand) genannt.

Der Margin lässt sich berechnen mit $\frac{2}{||w||}$. Um den Margin zu maximieren, genügt auch das Minimieren von $||w||$, wodurch man ein Optimierungsproblem erhält.

Duales Optimierungsproblem

w kann als eine Linearkombination von einer kleinen Untermenge an Datenbeispielen ausgedrückt werden, wie in 2.4 ersichtlich, sodass x_i die minimale Distanz zur Ebene hält. Dies führt zu der Lagrange-Gleichung, in der jedes x_i einer Datenmenge mit einem Parameter α_i multipliziert wird:

$$L(w, b, \alpha) = \frac{1}{2}|w|^2 - \sum_{i=1}^{n} \alpha_i * y_i * w * x_i + \alpha_i * y_i * b - \alpha_i \tag{2.4}$$

Die Minimierung von:

$$min_\alpha \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i * x_j) \tag{2.5}$$

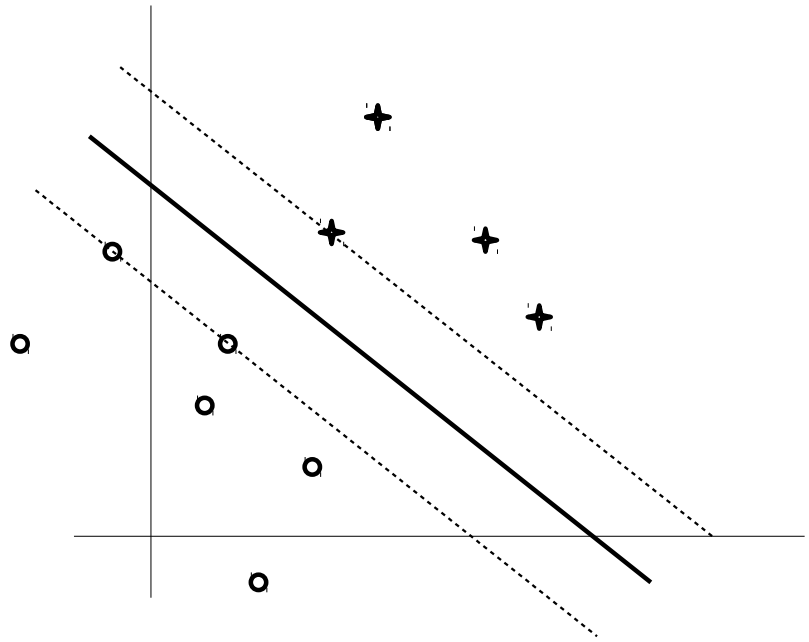


Abbildung 2.3: Die Hyperebene trennt die Daten in zwei Klassen. Die Entfernung beider zueinander parallelen Ebenen (gestrichelt) zur Trennebene werden von den zur Trennebene nächstliegenden Daten bestimmt.

unter den Nebenbedingungen:

$$\begin{aligned} w &= \sum_{i=1}^{n} \alpha_i * y_i * x_i \\ 0 &= \sum_{i=1}^{n} \alpha_i * y_i \end{aligned} \tag{2.6}$$

liefert Werte für alle α, womit w einfach ausgerechnet werden kann. Für die meisten Trainingsdaten ist $\alpha = 0$. Diejenigen extremen bzw. nahe zur Trennebene gelegenen Punkte mit $\alpha > 0$ werden Stützvektoren genannt.

Nur die Stützvektoren werden bei der Optimierung der Ebene verwendet und davon hängt somit die ganze Klassifikation ab. Stützvektoren sind in der Regel die Extremfälle, die schwierig zu klassifizieren sind. Wenn diese

richtig getrennt werden, wird davon ausgegangen, dass die eindeutigeren Daten ebenfalls korrekt klassifiziert werden.

Nichtlineare Trennbarkeit

Jedoch kann es vorkommen, dass zwei Klassen einer Datenmenge nicht linear separierbar sind. Dies wird anhand des XOR-Problems mithilfe eines grafischen Beispiels in Abbildung 2.4 veranschaulicht. So lässt sich eine lineare Trennung dennoch oft mithilfe des sogenannten Kernel-Tricks ermöglichen. Dabei wird der ursprüngliche Merkmalsraum mit einer geeigneten Transformation in einen höherdimensionalen Merkmalsraum überführt, in dem die Datenpunkte linear trennbar oder besser zu trennen sind.

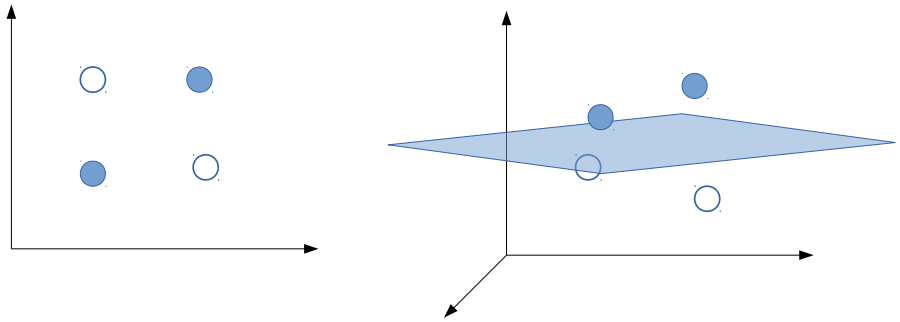


Abbildung 2.4: Hier wird das Prinzip des 'Kernel-Tricks' veranschaulicht. Links befindet sich eine zweidimensionale Datenmenge, dessen Klassen offensichtlich nicht durch eine lineare Funktion trennbar sind. Werden jedoch in einer zusätzlichen Dimension durch eine geschickt gewählte Funktion zusätzliche Werte generiert, so ist eine lineare Trennbarkeit unter Umständen wieder möglich, wie in der rechten Abbildung dargestellt.

Die Funktion zur Transformation des Merkmalsraums wird Kernel-Funktion genannt. Typische Kernel-Funktionen sind der Polynom-Kernel, der Sigmoid-Kernel und der Gauß-Kernel, die je nach Charakteristik der Datenstruktur gezielt eingesetzt werden können.

Regression

Die Support Vector Machine kann auch als Regressions-Verfahren verwendet werden, indem eine lineare Funktion gesucht wird, die die Werte eines Datensatzes in einem gemeinsamen Merkmalsraum am besten interpoliert. Somit wird zwar auch wie bei SVM eine Hyperebene durch die Datenmenge gespannt, jedoch sollen damit nicht zwei Klassen voneinander getrennt werden, sondern die Datenverteilung selbst repräsentiert werden. Durch die erwähnten Kernel-Funktionen können auch hier nichtlineare Strukturen erfasst werden. Da es selten sinnvoll oder auch möglich ist, mit der Funktion jeden Punkt exakt zu treffen, wird eine Slack-Variable für jeden Punkt definiert, die einen gewissen Grad an Fehlern zulässt. Bei der Bestimmung dieser Funktion ergibt sich mit den möglichst geringen Abständen der Punkte zur Funktion ein Optimierungsproblem. Dieses kann beispielsweise mit der Methode der kleinsten Quadrate von Gauss gelöst werden (Stigler (1981)).

Vor- und Nachteile von SVM

Vor- und Nachteile von SVM wurden zum Beispiel von Raghavendra zusammengetragen (Raghavendra and Deka (2014)). So liegt ein großer Vorteil in der Effektivität bei bereits relativ kleinen Datensätzen. Der Algorithmus kann verhältnismäßig gut mit hochdimensionalen Daten verfahren, weil nur Punkte nahe des Margins betroffen sind. Durch austauschbare Kernel-Funktionen wird eine hohe Vielseitigkeit gewahrt. Das gesamte Klassifikationssystem ist nur von relativ wenigen Stützvektoren abhängig und ist somit ein sehr kompaktes Modell mit wenig Speicherbedarf. Wurde ein Modell einmal trainiert und optimiert, lassen sich die Klassifikations-Schritte sehr schnell berechnen, wodurch sich das Verfahren sehr gut für Echtzeit-Anwendungen eignet.

Ein negativer Aspekt ist der oft hohe Trainings-Aufwand. In der Regel ist eine aufwendige Kombination von systematischem Absuchen des Parameter-Raumes und Kreuzvalidierung nötig, um gute Ergebnisse zu erzielen. Wenig geeignet ist das Verfahren auch bei sehr großen Datenmengen und stark verrauschten Daten. Zudem bieten Klassifikationsergebnisse keine direkte probabilistische Aussage, obwohl zumindest eine Schätzung möglich ist.

3 Grundlagen des menschlichen Kopfnickens

Das Kopfnicken ist eine weit verbreitete Kopfgeste, die durch abwechselndes Senken und Heben des Kopfes gekennzeichnet ist. Eine der wohl treffendsten technischen Definitionen für Kopfnicken lieferte W. McGrew 1972 (McGrew (1972)): *'The head is moved forward and backward on the condyles resting on the atlas vertebra, resulting in the face moving down and up. The down-up sequence may be exhibited once or repeated.'* (p. 57).

Einleitend wird hier ein Einblick in die Anatomie von Kopfnicken gegeben, die die Grundlage der physikalischen Kopfbewegung darstellt. Es schließt sich eine umfangreiche Erörterung zu der Bedeutung des Kopfnickens als kommunikatives Signal an, die zahlreiche aktuelle Forschungsergebnisse aufzeigt. Ein weiterer Abschnitt untersucht auch neuropsychologische Erkenntnisse, die über den kommunikativen Zweck von Kopfnicken hinausgehen. Es zeigt sich, dass kulturelle Unterschiede deutlich werden und eine Einordnung dieser Arbeit nötig wird. Zudem werden Ansätze zur Simulation menschlichen Kopfnickverhaltens vorgestellt. Abschließend wird ein Ausblick zur automatischen Interpretation von Kopfnicken gegeben.

3.1 Anatomie der Kopfgelenke

Kopfnicken sowie weitere Kopfbewegungen werden anatomisch durch die Beschaffenheit der Kopfgelenke ermöglicht (Graumann and Sasse (2003)).

Die in der Abbildung 3.1 dargestellten Kopfgelenke bestehen aus den ersten beiden Halswirbeln der Halswirbelsäule. Der erste Halswirbel wird Atlas (klinisch C1) genannt und der zweite Halswirbel Axis (C2).

Der Atlas stützt die gesamte Schädelbasis und wird von einer Gelenkkapsel mit ventralen (vorne) und dorsalen (hinten) Verbindungen stabilisiert. Die Gelenkflächen des Atlas bilden zusammen mit den unteren Flächen der

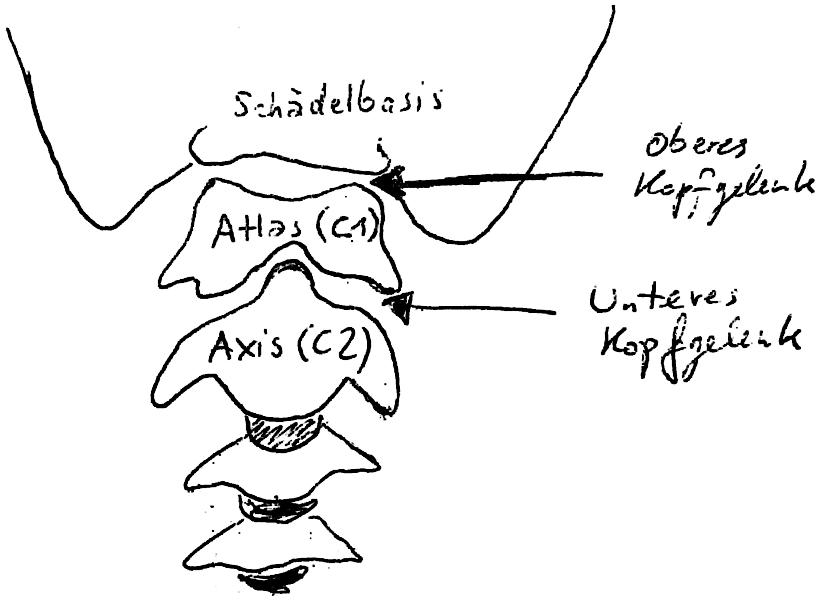


Abbildung 3.1: Durch die Kopfgelenke, bestehend aus der Schädelbasis und den Halswirbeln Atlas (C1) und Axis (C2) werden Kopfbewegungen in allen Raumdimensionen ermöglicht.

Schädelbasis das obere Kopfgelenk und ermöglichen als Ellipsoidgelenk Nickbewegungen.

Das untere Kopfgelenk wird durch das Gelenk zwischen dem Atlas und dem darunterliegenden Wirbel Axis gebildet. Der Axis ermöglicht die Drehung und somit Schüttelbewegungen des Kopfes.

Zusammen ermöglichen die Kopfgelenke somit Bewegungen in drei Raumebenen: Transversal ("drehen"), koronal ("neigen") und sagittal ("nicken").

3.2 Kopfnicken als kommunikatives Signal

Bereits C. Darwin untersuchte das Vorkommen und die Bedeutung menschlichen Kopfnickens in vielen Teilen der Erde und stellte fest, dass es sich

weltweit nahezu einheitlich um einen Ausdruck von Zustimmung handelt (Darwin (1872)). Er erklärte die Entstehung dieser Geste mit dem frühkindlichen Verhalten bei der Nahrungsaufnahme: So drehen Säuglinge an der mütterlichen Brust den Kopf zur Seite, wenn sie angebotene Nahrung ablehnen und nicken nach vorn, um sie zu sich zu nehmen. Dieses prägende Verhalten sei ein naheliegender Grund dafür, dass die meisten Menschen auch im Erwachsenenalter ein Kopfschütteln als Ablehnung und ein Kopfnicken als Zustimmung verstehen und einsetzen. Als alternativen Erklärungsansatz sah er im Kopfnicken eine abgeschwächte und nur angedeutete Form einer Verbeugung, die als eine Form der Respekt-Bezeugung eine mit Zustimmung verwandte Bedeutung assoziiert.

Wenn auch nicht vollständig verstanden, so wurde das menschliche Nickverhalten in den vergangenen Jahrzehnten intensiv erforscht, wie in den folgenden Ausarbeitungen ersichtlich wird. Dabei wurde immer wieder deutlich, dass sich das Bedeutungsfeld des Nickverhaltens längst nicht nur auf bloße Äußerung von Zustimmung, oder selten als Ablehnung wie in Abschnitt 3.4 erwähnt, beschränkt.

So untersuchte A.T. Dittmann im Jahr 1968 das Rückmelde-Verhalten der zuhörenden Person gegenüber der sprechenden Person in Gesprächssituationen (T. Dittmann and G. Llewellyn (1968)). Dabei zeigte sich, dass sich der Sprachfluss des Sprechers in rhythmische Gesprächseinheiten einteilen lässt. Auf Beendigung dieser sprachlichen Einheiten hin sendete der Zuhörer meist Rückmelde-Signale aus. Diese traten oft in Form von Vokalisierungen wie 'M-hmm', 'Verstehe' oder 'Okay' auf und wurden häufig von Kopfnicken begleitet. Wurde während einer derartigen Vokalisierung genickt, kam das Nicken der Vokalisierung etwa 175 Millisekunden zuvor. Dies liege darin begründet, dass eine verbale Äußerung den Sprecher womöglich unterbrechen würde, während eine Kopfgeste als weniger störend empfunden wird. So kann die Kopfgeste eine frühe Rückmeldung an den Sprecher senden, ohne ihn in seinem Redefluss durch eine verbale Äußerung zu stören.

Wurde das Nickverhalten isoliert betrachtet, so ließen sich 70% des Nickvorkommens in zwei Muster aufteilen: Entweder kündigte ein Zuhörer-Nicken eine kurz bevorstehende Anmerkung oder Frage an, oder es handelte sich um eine Reaktion auf eine direkte Frage oder einer beiläufigen 'Weißt du?'-Phrase des Sprechers. Somit liegt nach Dittmann die Vermutung nahe, diese 70% der Kopfnicken des Zuhörers seien entweder mit dem Wunsch

zu sprechen verknüpft oder eine Reaktion auf eine Feedback-Anfrage des Sprechers. Die restlichen 30% der beobachteten Kopfnicken wurden als 'Anhaltende Signale von Aufmerksamkeit' betrachtet.

Zwei Jahre später prägte Birdwhistell 1970 den Begriff der 'cinesic units' (Birdwhistell (1970)). Dazu zählte er körperliche Bewegungsabläufe, Gesten und Gesichtsausdrücke, die als nonverbale kommunikative Ausdrücke verstanden und in eindeutige, klar definierte Einheiten zusammengefasst werden können. Er schätzte, dass diese nonverbalen Informationen innerhalb einer Konversation etwa 70% der übermittelten Information ausmachen. Dabei zählte er auch Kopfnicken zu einer eindeutigen 'cinesic unit' mit klarer Definition. Er untersuchte sie an Amerikanern und bestimmte für eine vertikale Kopfbewegungs-Trajektorie einen Geschwindigkeitsbereich im Wortlaut zwischen 0.8 und 3 Grad pro 1/24 Sekunden (was etwa 19 bzw. 72 Grad pro Sekunde entspricht) über einen Winkel von 5 bis 15 Grad, um diese der 'cinesic unit' namens 'Kopfnicken' zuzuordnen. Variationen in der Ausführung unterschieden sich in den Attributen Geschwindigkeit, Amplitude und Zyklen-Frequenz. Diese traten insbesondere je nach Funktion der Nickgeste in unterschiedlichen Ausprägungen auf. So sei bei einer einfachen Zustimmung eher eine zyklische Ausführung mit durchschnittlicher Geschwindigkeit zu erwarten, während schnelle, scharfe Nicken eher mit dem Attribut 'Ungeduld' oder 'Stress' assoziiert werden könnten. Eine große Aussagekraft habe jedoch bereits der Kontext, in dem genickt wird. So könne ein Nicken als Reaktion auf eine Frage sehr wahrscheinlich Zustimmung signalisieren, aber ein Zuhörer-Nicken während einer aktiven Sprechphase des Gesprächsteilnehmers eher Ungeduld.

Yngve untersuchte Backchannel-Signale sowie Verhaltensmuster für Turn-Taking (Yngve (1970), Duncan (1972)). Als Backchannel gilt dabei jegliche verbale oder gestische Rückmeldung des Zuhörers, die nicht auf eine Regulation von Turn-Taking abzielt. Backchannels signalisieren dem Sprecher die aktive Zuhörerschaft des Hörers ohne den Sprecher zu unterbrechen.

Turn-Taking ist ein Begriff aus der Analyse der Gesprächsorganisation und beschreibt die Dynamik des Sprecherwechsels. Nach Sacks (1992) sei Turn-Taking eine kollaborativ organisierte Errungenschaft, die eine komplexe 'soziale Maschinerie' in ihrer lokalen Leistung durch Sprecher und Hörer ausnutzt. Darunter fällt die an seinen Zuhörer gerichtete Aufforderung eines Sprechers, das Wort zu ergreifen. Aber auch die Wortergreifung durch Eigeninitiative. Yngve stellte fest, dass Kopfnicken sowohl als Backchannel

eingesetzt, aber auch als Signal für einen initiierten Sprecherwechsel genutzt wird und somit auch eine Funktion als Turn-Taking übernehmen kann.

Nach A. Kendon sind Kopfnicken entscheidender Bestandteil in der Produktion, in der Aufrechterhaltung und in dem Management von Konversationen (Kendon (1972)). So beobachtete er, dass Kopfbewegungen beim Sprechen dazu eingesetzt werden, damit verschiedene Spracheinheiten, genannt 'locutions', besser voneinander unterschieden werden können. Ein von einem Sprecher vor oder bei Sprechbeginn ausgeführtes Nicken erklärte er als ein Signal für 'floor-apportionment', also als eine Bekräftigung der eigenen Wortergreifung.

Brown schrieb bereits 1986, dass sich verbale und nonverbale Backchannel-Signale gegenseitig ergänzen und regulieren (Brown (1986)). Dennoch können sie auch vereinzelt ohne gegenseitige Regulation auftreten oder verschiedene Gesten ausgeführt werden, ohne dass sich die transportierte Aussage ändert. So gäbe es zu einer bestimmten Äußerung nicht immer eine bestimmte passende Geste, sondern es könnten oft verschiedene gestische Backchannel-Signale dieselbe Bedeutung vermitteln.

B. Giges arbeitete 1975 verschiedene funktionelle Klassen für unbewusst eingesetztes Kopfnicken heraus (Giges (1975)). Er ging davon aus, dass nicht bewusst eingesetztes Kopfnicken zielführende oder vermeidende Verhaltensmotivationen beinhaltet. Der eigentliche Zweck der Geste liege darin, ein Signal auszusenden, um etwas für sich wieder zu empfangen.

Er unterschied zwischen angefordertes und nicht angefordertes Kopfnicken. Nicht angefordertes Kopfnicken wird oft unbewusst eingesetzt, während angefordertes Nicken als Reaktion auf eine Anfrage des Gesprächspartners als bewusstes Information-sendendes Signal getätigt wird. Wiederholendes Nicken des Zuhörers weise oft auf dessen mangelndes Interesse oder mangelnde Aufmerksamkeit hin. Einige seiner Klassendefinitionen waren 'rescue nodder', 'control nodder' oder 'put-down nodder'. 'Rescue nodder' sind demnach nickende Zuhörer, die meinen, der Sprechende fühle sich nicht wohl, wenn jetzt nicht genickt wird oder könne damit nicht umgehen. Ein 'Control nodder' ist entweder ein Ungeduld zeigender Zuhörer oder ein Sprecher, der mit Nick-Gesten versucht, seinen Gesprächspartner 'zurückzudrängen'. Erwartet die zuhörende Person, dass ein Fehler eingeräumt wird, so zeigt sich bei erfüllender Erwartung ein genugtuendes 'put-down'-Nicken.

Neben Birdwhistell stützte auch U. Hadar 1985 die Annahme, dass sich die

Ausführung einer Kopfgeste mit unterschiedlicher kommunikativer Funktion auch in den physikalischen Eigenschaften unterscheidet (Hadar et al. (1985)). Er unterteilte Kopfgesten in die Funktions-Kategorien Ja/Nein, Synchronisation, Antizipation und Weitere. Als Kopfnicken betrachtete er sämtliche Gesten mit vertikaler Kopfbewegung und unterschied so zwischen zyklischen Bewegungen und linearen Bewegungen. Er zählte in seinen Untersuchungen den größten Teil des Nickvorkommens zu der Kategorie 'Ja-Nicken'. Davon enthielten die meisten eine begleitete verbale Rückmeldung. Die Ausführungen waren recht gleichmäßig, mit relativ hoher Zahl an hintereinanderfolgenden abklingenden (um die drei) Zyklen und durchschnittlich stark ausgeprägter Amplitude.

Die nach den Ja-Nicken am zweithäufigsten beobachtete Kategorie war Synchronisations-Nicken. Diese zeigten sich meist während oder kurz nach Unstetigkeiten innerhalb des Sprecher-Redeflusses, traten innerhalb von etwa 0.2 Sekunden nach der Unstetigkeit auf und waren nicht von begleiteter verbaler Äußerung gekennzeichnet. Dabei war das Nicken häufiger linear als zyklisch, von kürzerer Dauer und die Bewegung tendenziell nach oben gerichtet. Die Amplituden blieben eher schwach ausgeprägt.

Antizipations-Nicken war linear oder mit nur einem Zyklus zu beobachten und wies besonders ausgeprägte Amplituden auf. Typische Situationen, in denen Antizipations-Nicken auftrat, waren kurz vor Beginn von Sprachausführungen.

Insbesondere zeigte sich, dass sich diese Eigenschaften zwischen den Versuchspersonen nicht signifikant unterschieden. Allerdings basieren diese Untersuchungen auf Ergebnissen mit lediglich sechs Probanden und einer Gesamtzahl an Kopfnicken von unter 100.

Eine Tabelle der Verteilung der Funktionskategorien für Nickbewegungen ist in der Abbildung 3.2 dargestellt.

S. Maynard untersuchte Dialoge zwischen 12 japanischen Studierenden (Maynard (1987)). Obwohl die meisten Nicken in Form von Zuhörer-Nicken als Backchannel auftraten, wurden auch viele Sprecher-Nicken beobachtet. Die japanischen Probanden nickten dabei fast ausschließlich entweder bei Silbenendungen am Ende von Sprechphrasen oder während möglicher Turn-Taking Situationen. Die Sprecher-Nicken kennzeichneten oft ein Satz-Ende oder das Ende von Sprechabschnitten. Das Nicken am Ende von Sprechabschnitten leitete oft einen Turn-Taking Vorgang ein. Des Weiteren wurden damit Zustimmung und Betonung assoziiert. Oft wurde auch genickt, bevor

TABLE I
HEAD NODS

Differential distribution of linear and cyclic head nods between the various functions (number of cycles per bout of cyclic movement and mean amplitude in each function).

	Functions						
	Yes	No	Synchrony	Antici-pation	Other	Total	Significance
Total number of movements	40	0	23	15	15	93	G(4)=49.9
Linear movements	1	0	16	9	10	36	Corrected for continuity
Cyclic movements	39	0	7	6	5	57	p<0.01
Mean cycles per bout	2.9	-	2.0	1.2	1 3		
Mean amplitude (deg)	13.3	-	11.4	18.3	26.6		F(3,89)=16.8
SD amplitude	9.4	-	6.6	7.3	12.1		p<0.05

Abbildung 3.2: Die Abbildung zeigt funktionelle Klassen für Kopfnicken nach Hadar et al. (1985). Auf der linken Seite sind die beobachteten physikalischen Merkmale aufgelistet und werden oben den Funktionsklassen gegenübergestellt.

das Wort ergriffen und somit ein eigener Sprachbeitrag eingeleitet wurde. Somit hatte Kopfnicken sowohl semantische, syntaktische, also auch interaktionale Funktion.

Während die japanischen Probanden im Durchschnitt alle 5.75 Sekunden nickten, wurde auf eine Studie von Clancy (1982) verwiesen, nach der bei Amerikanern in vergleichbarem Kontext zwischen zwei Nicken durchschnittlich 22.5 Sekunden lagen.

E. McClave untersuchte Kopfbewegungen im Gesprächskontext auf linguistische Funktionen hin (McClave (2000)). Eine der Hauptaussagen war Folgende: Setzt ein Sprecher Kopfnicken ein, so fungiert dieses als Backchannel-Anfrage, für die Zuhörer besonders sensibilisiert sind. Auf Sprechernicken folgte innerhalb von 0.7 Sekunden oft ein Hörer-Nicken oder ein

'um hum'. Dies lässt die Annahme zu, dass das Zuhörer-Nicken als Backchannel durch das Sprecher-Nicken getriggert wurde. Sprecher senden somit Anfragen aus, um Feedback zu erhalten. Diese interaktive Funktion des Sprecher-Nickens sei jedoch kulturspezifisch.

Wilbur untersuchte die Amerikanische Gebärdensprache, auch ASL (American Sign Language) genannt, auf die beinhaltenden Kopfnick-Gesten (Wilbur (2000)).

Dabei wurden drei mögliche Funktionen für Kopfnicken wahrgenommen. Einmal als Beendigungszeichen einer Spracheinheit, dann als langsames abwegendes Kopfnicken als Fokus-Markierung und drittens als wiederholendes Kopfnicken, dass bei großer langsamer Ausführung eine starke Behauptung kennzeichnet oder als schnelles Nicken mit schwächerer Amplitude der Absicherung dient.

Cappar stellte im Jahr 2000 fest, dass bei Japanern nicht zu unterscheiden ist, ob mit einer Nickausführung lediglich Verständnis und Evidenz für aktives Zuhören gemeint ist, oder ob es sich dabei um ein Indiz für Zustimmung handelt (Capper (2000)). Die schwierige Unterscheidbarkeit führte im Dialog mit Nicht-Japanern zu Verständigungs-Problemen.

Auch Nunn ging im Jahr 2003 der Frage nach, inwiefern sich Nicken innerhalb von Dialogen zwischen Menschen unterschiedlicher Kultur auswirkt und ob es die Kommunikation verbessert oder verschlechtert (Nunn and Maya (2003)). Im Rahmen seiner Untersuchungen nahm er, wie auch der Tabelle 3.3 zu entnehmen, folgende Kategorisierung für Kopfnicken vor: Zunächst unterschied er zwischen Sprecher-Nicken und Zuhörer-Nicken. Sprecher-Nicken unterteilte er in 'Zustimmungs-Anfrage an Zuhörer' und in 'Funktion zur Regulation von Turn-Taking'.

Das Hörer-Nicken dagegen wies weit mehr verschiedene Klassen auf. Die am einfachsten beobachtbare Kategorie war Zustimmung, Genehmigung, oft auch mit einer zustimmenden Verbalisierung begleitet. Des Weiteren wurde eine Kategorie für 'Ermutigung zum Weitersprechen' festgelegt. Diese tritt vor allem in Sprechpausen auf. Manchmal könne dies auch als Weigerung des Zuhörers gedeutet werden, das Wort zu ergreifen. Dieser Effekt wurde von Knapp und Hall 1997 intensiv untersucht (Hall et al. (1997)).

Die dritte Zuhörer-Kategorie beinhaltet Nicken als Willensbekundung oder Anfrage, das Wort zu ergreifen. Eine zutreffende Intention wäre etwa *'Jetzt lass' mich mal zu Wort kommen, ich will etwas sagen.'*.

Die vierte Kategorie beschreibt eine Bekräftigung, dass das Gehörte ver-

Nods when holding the floor	**Nods when another speaker has the floor (Back-channel nods)**
(1) Nods asking for agreement. (2) Nods regulating turn taking.	(1) Nods of agreement/approval accompanying 'Yes', 'Yeah' or non-verbalized with the same meaning[3]. (2) Nods to encourage the speaker to continue (possibly, turn refusing[4]). (3) Nods to request a turn, around the end of turns. (4) Nods acknowledging understanding. (5) Nods that occur for no apparent reason, as they are not visible to the floor holder.

Abbildung 3.3: Verschiedene Typen für Kopfnicken nach Nunn and Maya (2003). Die Hauptkategorisierung erfolgte gemäß Zuhörer- (rechts) und Sprecher-Nicken (links).

standen wurde.

Unscheinbare Nick-Gesten, die keiner Kategorie zugeordnet werden konnten, wurden zu einer Kategorie zusammengefasst, die für den Gesprächspartner in der Regel nicht sichtbar ist. Für diese Kategorie wird davon ausgegangen, dass sie keine Signal-Funktion besitzt.

Neben dieser funktionalen Einteilung sei auch eine Gliederung von Kopfnicken in zwei folgende Kategorien möglich: Erstens mit der Absicht zu zeigen, dass der Gesprächsinhalt verstanden wurde und zweitens Kopfnicken mit lediglich solidarischer Funktion.

Besonders auffällig war, dass japanische Studierende in der Sprecher-Rolle etwa drei bis vier Mal häufiger nickten, als Studierende aus anderen Ländern. Bezüglich der Frage, inwiefern das Nickverhalten interkulturelle Kommunikation verbessert oder verschlechtert, gab es zwei wesentliche Beobachtungen: Einerseits schien extensives Nicken eine solidarität-schaffende Funktion inne zu haben, andererseits schien es auch eine negative Auswirkung auf die

Effektivität der Kommunikation zu haben. Dies liege darin begründet, dass Nicken fälschlicherweise als Verstehen interpretiert wurde, obwohl es lediglich solidarisch gemeint war.

S. Adolphs untersuchte 2007 ebenfalls Kopfnick-Verhalten innerhalb von Gesprächskontexten (Adolphs and Carter (2007)). Sie sah Kopfnicken als eine der markantesten Gesten in der Kommunikation, die vor allem als Backchannel-Signal wichtig ist. Adolphs empfahl, bei der Exploration von Kopfnicken innerhalb der zu untersuchenden Datenbasis Funktion, Timing, Signifikanz und (mögliche) Antwort bei der Nick-Ausführung zusätzlich aufzuzeichnen und zu untersuchen.

T. Stivers untersuchte 2008 das Nickverhalten von Zuhörern, wenn ihnen eine Geschichte erzählt wird (Stivers (2008)). Das Nicken eines Zuhörers sei in diesem Kontext ein Indiz dafür, dass dieser beansprucht, einen Zugang zu der dargestellten erzählten Situation gefunden zu haben und daran teil hat.

Petukhova stellte 2009 die These auf, Kopfnick-Gesten, die sich bezüglich Ausführungsgeschwindigkeit, Dauer, Timing und Intensität unterscheiden, können durchaus unterschiedliche Bedeutungen vermitteln. Jedoch würden sie isoliert betrachtet keine angemessene Interpretation erlauben, inwiefern die ausführende Person eine Information tatsächlich so verstanden hat, wie es die beobachtende Person durch das erhaltende Kopfnick-Signal übermittelt bekommt (Petukhova and Bunt (2009)).

Kopfnicken mit verbaler Zustimmungs-Äußerung wurde von Annotatoren meistens übereinstimmend als 'belief adoption' und somit sozusagen als Übernahme der Meinung des Gegenübers in einem bestimmten Aspekt gedeutet. Dagegen waren Annotatoren sich oft uneinig über die Bedeutung, wenn das Kopfnicken lediglich von einer schwachen Vokalisierung wie 'uh-uhu' begleitet war.

S. Moubayed zeigte durch Studien mit einem virtuellen Avatar, dass die Sprachverständlichkeit bei den Probanden stieg, wenn der virtuelle Agent prominente Wörter mit Kopfnicken oder Augenbrauenanheben akzentuierte (Moubayed et al. (2009)). Die Ausführung eines Kopfnickens dauerte etwa 350 Millisekunden, was der Länge nach einer betonten Silbe im Schwedischen entspricht. Durchschnittlich waren 1-3 Prominenz-Marker pro Satz vermerkt.

I. Poggi nahm im Jahr 2010 eine intensive Untersuchung von 100 Nick-Beispielen vor und versuchte, daraus eine Kategorisierung der verschiedenen

möglichen Bedeutungen von Kopfnicken auszuarbeiten (Poggi et al. (2010)). Dabei prüfte sie diese auf charakteristische physikalische Merkmale, auf beiläufig auftretende Nebeneffekte in anderen Modalitäten und untersuchte den entsprechenden Gesprächskontext. Sie unterschied zunächst, welche Rolle die nickende Person im Gesprächskontext einnehmen kann. Die Kategorien 'Gesprächspartner oder Zuhörer', 'Dritter Zuhörer' und 'Sprecher' reduziere ich im Folgenden auf den Dialog-Kontext auf 'Zuhörer' und 'Sprecher'.

Ein Zuhörer-Nicken, während der Sprecher spricht, wird als Backchannel-Signal interpretiert. Nickt ein Zuhörer nachdem der Sprecher gesprochen hat, sollte es im Kontext der Aussage des Sprechers betrachtet werden. Zunächst wird das Kopfnicken der zuhörenden Person beschrieben.

In der Rolle des passiven Gesprächspartners wird auftretendes Kopfnicken entweder als *Eingeforderte Bestätigung* (Englisch: requested confirmation) eingeordnet oder als *Spontane Bestätigung* (Englisch: spontaneous confirmation). Die Eingeforderte Bestätigung ist meist eine Antwort auf eine Ja-Nein-Frage. Die Spontane Bestätigung erfolgt als Reaktion auf die Nennung von angenommenen Fakten. Beides bringt zum Ausdruck, man teile dieselben ausgesprochenen Annahmen mit dem Sprecher.

Bezieht sich ein Nicken auf eine wertende oder bewertende Aussage, handelt es sich um ein Zustimmendes Nicken (Agreement). Ein Agreement-Nicken ist normalerweise etwas langsamer und ausgeprägter als ein Bestätigungs-Nicken und wird manchmal von einem leichten, fast hochmütig wirkenden, Schließen der Augen begleitet.

Unterbreitet der Sprecher gerade einen Vorschlag, könnte Nicken Einverständnis vermitteln.

Weitere Anwendungsfälle von Kopfnicken zeigen sich als Geste zur Begrüßung und als Zeichen des Bedankens.

Drei wichtige Nicktypen mit entsprechenden physikalischen Attributen sind:
(1) 'backchannel of confirmation' mit der Aussage: 'Ich bestätige, dass ich folge - ich verstehe was du sagst.' Dabei kann die erste Nickausführung sehr ausgeprägt sein, dann folgen oft zwei oder mehr kurz und rhythmisch wiederholte Nicken.

(2) 'taking note' mit der Aussage: 'Ich nehme zur Kenntnis was du

sagst.' Hier ist die Nickausführung eher sehr kurz und schnell.

(3) 'backchannel nods of agreement' mit der Aussage: 'Ich habe dieselbe Meinung wie du.' Diese Agreement-Nicken sind meist vereinzelt, ausgeprägt und deutlich.

Seltener tritt die Kategorie 'ironic agreement' auf, sowie 'back-agreement'. Back-Agreement besagt, dass der Hörer nickt, weil der Sprecher etwas sagt, was der Hörer bereits gesagt oder gedacht hat. Dieses ist manchmal von Lächeln oder Seufzen begleitet und mit Gedanken wie 'Na also, endlich siehst du ein, dass ich Recht habe.'

Eine weitere wichtige Zuhörer-Kategorie ist der 'Processing Nod'. Dieses Nicken kommt zustande, wenn gedanklich tiefergehende Denkprozesse ausgeführt werden, die einen größeren Teil der eigenen Aufmerksamkeit fordern. Folglich wird in dem Fall mit einem Kopfnicken eine Art Selbstbestätigung ausgedrückt, sobald die Denkprozesse zu einer besseren Einordnung oder zu besserem Verstehen der verarbeiteten Information geführt haben. Beim Processing Nod handelt es sich somit nicht um ein Signal an den Sprecher, sondern um eine Art Selbstreflexion. Daher ist dieses Nicken auch oft mit einer kurzen geistigen Abwesenheit begleitet. Oft lässt sich beobachten, dass man konzentriert die Stirn runzelt und den Gesprächspartner dabei nicht anschaut, sondern auf den Boden, zur Seite oder nach oben schaut.

Das Sprecher-Nicken ist etwas weniger häufig und wurde in etwas weniger Kategorien aufgeteilt. Signale in Form von Kopfnicken, die von dem aktiven sprechenden Gesprächspartner als Sprecher-Nicken ausgesendet werden, erstrecken sich grob in zwei Dimensionen: Wichtigkeit und Bestätigung. Kategorien zu Sprecher-Nicken mit dem Attribut von Wichtigkeit sind:

(1) 'Betonung': Hierbei wird Ausgesprochenes betont und hervorgehoben. Der Kopf wandert eher nach vorn, sodass eine vorn-runter Bewegung im Beiklang mit betonter Silbe und Blick auf den Gesprächspartner gerichtet stattfindet. Ein Fokussieren der Augen kann hier ebenso als Betonung fungieren.

(2) 'Batonic': Diese Kategorie wird verwendet, wenn nicht nur wichtige Wörter betont werden, sondern jede Silbe im Satz mit Nicken artikuliert wird. Das kann auch bedeuten, dass der Sprecher selber im Rhythmus

bleiben möchte und sich dabei unterstützt.

Zur Kategorie der Bestätigung gehört vor allem 'Fragendes Nicken'. Dabei wird Blickkontakt gehalten, oft mit Stirnrunzeln oder leicht seitlicher Kopfhaltung. Manchmal handelt es sich dabei lediglich um eine Backchannel-Anfrage, ob die ausführende Person sich der Aufmerksamkeit des Gesprächspartners noch sicher sein kann oder ob dieser abgelenkt ist.

Eine Tabelle gemäß der Kategorisierungen von I. Poggi findet sich in der Abbildung 3.4.

Im Jahr 2010 untersuchte Boholm die zyklischen Eigenschaften von Kopfnicken schwedischer Versuchsteilnehmer und kam zu dem Schluss, dass Nicken mit mehreren wiederholten Zyklen immer eine Bestärkung des Ausdrucks oder des Inhalts als Funktion innehat (Boholm and Allwood (2010)). Abgesehen von Feedback seien weitere Funktionen die Folgenden: Erstens die Bestärkung der aussprechenden Wörter, wenn es gemeinsam mit Sprache ausgeführt wird. Zweitens ein Ausdruck von Selbstbestätigung, meist nachdem man gesprochen hat. Drittens das Einläuten einer Umstellung zwischen direkter und indirekter Rede, was auch eine Art von zitierter Geste sein könnte und viertens ein sogenanntes 'own communication management' wie etwa Wortfindungs-Prozesse.

F. Nori (Nori et al. (2011)) definierte ein Kodier-Schema für Nicken wie folgt: Nicken beginnt, wenn das Kinn anfängt, sich nach oben oder unten zu bewegen. Es hört dann wieder auf, wenn die Bewegung endet. Dabei werden drei Typen unterschieden: Die Person nickt mit einem Nick-Zyklus, mit zwei Zyklen oder mit mindestens drei Zyklen. Nach diesem Schema wurde das Nickverhalten von deutschen mit japanischen Versuchsteilnehmern verglichen. Dabei zeigte sich ein mehr als doppelt so hohes Nick-Vorkommen bei den japanischen Nutzern im Vergleich zu den deutschen Nutzern. Des Weiteren zeigte sich, dass japanische Nutzer verhältnismäßig häufiger einzelne Nick-Zyklen ausführen, während deutsche Nutzer häufiger Kopfnicken einsetzen, die mindestens drei Zyklen beinhalten. Das zeigte sich besonders darin, dass beim Zuhören die japanische Gruppe häufige einzelne Nicken verwendete, die deutsche Gruppe dagegen zwar weniger häufige, dafür aber die wenigen mehr Zyklen aufwiesen. Beides lässt die Annahme zu, dass derartiges Feedback in beiden Kulturen solidarisch eingesetzt wird und so zum Spannungsabbau zwischen Dialogpartnern beiträgt.

C. Navarretta untersuchte das Feedback-Verhalten von dänischen, finni-

schen und schwedischen Versuchsteilnehmern während sie einander zum ersten Mal begegneten (Navarretta et al. (2012)). Tatsächlich war Kopfnicken das am häufigsten beobachtete Feedback-Signal. Es wurden einige Unterschiede bezüglich des Nickverhaltens zwischen den drei Nationalitäten festgestellt. So beinhaltete das Nicken der Dänen außerordentlich häufiger 'down'-Bewegungen, während Schweden stattdessen am häufigsten 'up'-Bewegungen einsetzten. Außerdem zeigten die Finnen wesentlich öfter einzelne Nick-Zyklen in Situationen, in denen die beiden anderen Nationalitäten oft viele aufeinanderfolgende Zyklen aufwiesen.

Weitere Erkenntnisse mit finnischen Versuchsteilnehmern fand E. Toivio 2012 mittels Studien zur Korrelation von Gesten und Sprache im Finnischen (Toivio and Jokinen (2012)); So signalisierten 'Down'-Nicken Backchannel, während 'Up'-Nicken Reaktionen auf unerwartete Information darstellten.

Das Timing von Kopfnicken beziehungsweise die Extremwerte derer Trajektorien korrelierten oft mit der prosodischen Struktur, wie E. Asor 2014 anmerkte (Asor (2014)). Das heißt, dass das Maximum der Nickbewegung gleichzeitig mit dem Maximum der Betonung eines Schlüsselwortes auftrat.

M. Fusaro untersuchte Nickverhalten von Müttern beim Spielen mit ihren Kindern (Fusaro et al. (2014)). Dabei zeigten sich durchschnittlich 4.0 bis 5.8 Kopfnicken pro 10 Minuten, also etwa 2-2.5 Kopfnicken pro Minute. 97.3% der Nicken traten zusammen mit verbaler Äußerung auf.

In einer Veröffentlichung von K. Svinhufvud im Jahr 2016 wird das Anfertigen von Notizen innerhalb von Dialogen zwischen Studienberater und Studierenden beschrieben (Svinhufvud (2016)). Studierende wurden bei einem Beratungsgespräch beobachtet. Dabei war auffällig, dass Studierende den Übergang von Zuhören zu Notizen-Machen oft mit einem Nicken begleiten. Es wird davon ausgegangen, dass somit eine Übergangsgeste zwischen zwei Aktionen ausgeführt wird. Eine weitere Beobachtung war das Nickverhalten während des Schreibens. Dieses kann als ein Signal dafür gedeutet werden, dass trotz des Schreibens gleichzeitig zugehört wird und der Sprecher zum Weiterreden motiviert werden soll.

Bamoallem implementierte in einer Studie kommunikative Kopfgesten in eine Telepräsenz-Plattform (Bamoallem et al. (2016)). Dabei stellte er fest, Echtzeit-Kommunikation benötige mehr als verbale Kommunikation, Gesichtsausdrücke und Kopfnicken. So sollte weitere Gestik wie beispielsweise die Körperhaltung mit berücksichtigt werden.

Wie die vielen in diesem Kapitel genannten Klassifikations-Versuche, zeigt

dies auch nochmal die Komplexität des Themas um Kopfnicken und Kommunikation auf. Dies wird auch im folgenden Abschnitt deutlich, in dem über des kommunikativen Aspekts des Kopfnickens hinausgehende Aspekte angesprochen werden.

3.3 Soziologische und neuropsychologische Erkenntnisse um Kopfnicken

Während der vorige Abschnitt Erkenntnisse im Bereich der kommunikativen Aspekte um Kopfnicken beleuchtet, soll dieser Abschnitt Effekte schildern, die über die Betrachtung von Kopfnicken als kommunikatives Signal hinausgehen.

So zeigte Wells 1980 in einem Experiment, dass die bloße vertikale Nickbewegung ohne kommunikative Intention unbewusst mit Zustimmung gekoppelt ist und sich Nickbewegungen und Zustimmung sogar gegenseitig aktivieren können (Wells et al. (1980)). Er ließ Versuchspersonen Kopfhörer testen mit der Anordnung, zur Überprüfung des Tragekomforts wiederholt bestimmte Kopfbewegungen auszuführen. Die Versuchsgruppen 'Vertikale Bewegungen', 'Horizontale Bewegungen' und 'keine Vorgabe' hörten dieselbe simulierte Radiosendung. Dabei zeigte die Gruppe mit der Anweisung für vertikale Bewegungen signifikant mehr Zustimmung zu den Inhalten als die beiden anderen Gruppen. Andererseits hatten Personen, die trotz bestehender ablehnender Haltung zu den Inhalten nicken mussten, Probleme bei der Ausführung von vertikalen Bewegungen. So war ihr Bewegungsmuster weniger stark ausgeprägt als in den anderen beiden Gruppen. Dies lässt eine Hemmung des für Nicken zuständigen Bewegungsapparates bei empfundener Ablehnung vermuten.

A. Mehrabian stellte in seinem im Jahre 1972 erschienenen Buch 'Nonverbal Communication' fest, dass Menschen bei bewusst unehrlichen Aussagen unter anderem auch weniger Kopfnicken zeigen: *"when people are being dishonest, they tend to nod and gesture less while moving their legs and feet less, speaking less and more slowly with more errors in their speech, and smiling more frequently."* (Mehrabian (1972)). Auch stellte er fest, dass Attribute wie empfundene Unterwürfigkeit oder geringes Selbstbewusstsein im Bezug zum Gesprächspartner sich durch besonders ausgeprägtes Nickverhalten zeigen können.

Veröffentlichte Untersuchungen von J. Förster und F. Strack im Jahr 1996 zeigten, dass positiv assoziierte Wörter besser auswendig gelernt werden konnten, wenn während der Übungseinheiten Nickbewegungen ausgeführt werden sollten (Förster and Strack (1996)). Dasselbe galt auch für den umgekehrten Fall mit negativ assoziierten Wörtern und Kopfschütteln.

Durch ein Experiment zeigte Brinol 2003, dass die Überzeugungskraft einer Mitteilung gesteigert werden kann, wenn der Adressat während des Mitteilungsprozesses der Nachricht Nickbewegungen ausführen soll (Brinol and Petty (2003)).

M. Helweg-Larsen untersuchte die Nickhäufigkeiten innerhalb von Konversationen unter Berücksichtigung von Status und Geschlecht (Helweg-Larsen et al. (2004)). Signifikante Beobachtungen waren dabei, dass weibliche Studierende generell häufiger nicken, als männliche Studierende. Zudem nicken Studierende öfter, wenn sie sich mit einem Professor unterhalten, als in Gesprächen untereinander.

J. Xu zeigte mit seinen Arbeiten, dass Gehirnbereiche, die als zentrale Einheit für die Dekodierung von gesprochenen und geschriebenen Wörtern zuständig sind, auch bei der Interpretation von Gesten aktiv sind (Xu et al. (2009)).

Beschrieben als Mimikry oder Synchronisation konnte gezeigt werden, dass miteinander interagierende Menschen dazu tendieren, zueinander kohärente Bewegungsmuster auszuführen (Bernieri and Rosenthal (1991)). Entsprechend untersuchte Hale (Hale et al. (2018)) die Kohärenz der vertikalen Kopfbewegungen zwischen zwei Gesprächspartnern. Hohe Kohärenz mit einer Zeitverzögerung von etwa 0.6 Sekunden wurde bei Frequenzen von vertikalen Kopfbewegungen zwischen 0.2 bis 1.1 Hz gefunden, während bei sehr schnellem Nickverhalten eine signifikante Anti-Kohärenz gemessen wurde. Schnelles Nicken könnte demnach zwei wesentliche Bedeutungen enthalten:

Erstens könnte es sich um einen körperlichen Ausdruck von Einstimmung in den Sprachrhythmus des Sprechers handeln. Möglicherweise, um das gegenseitige Verständnis zu unterstützen, wie es bereits von anderen Arbeiten beschrieben wurde.

Zweitens könnte es sich um ein Backchannel, speziell um ein Signal der Aufmerksamkeit und des Engagements an den Sprecher, handeln.

Eine weitere Erkenntnis war, dass bei Beobachtung von sehr schnellem kurzen Nicken mit sehr hoher Wahrscheinlichkeit davon auszugehen ist, dass der Gesprächspartner der entsprechenden Person gerade spricht.

S. Moretti gelang 2018 die Erkenntnis, dass die Verarbeitung von wahrer Information automatisch die Simulation von vertikalen Kopfbewegungen aktiviert, die gewöhnlich bei positiven, zustimmenden Antworten durchgeführt wird (Moretti and Greco (2018)). Die Ergebnisse aus ihren Experimenten zeigten, dass Wahrheitsbeurteilung mit Kopfbewegungen interagiert, die typischerweise bei Nick- oder Schüttelbewegungen ausgeführt werden.

3.4 Kulturelle Unterschiede und Einordnung

Lassen sich, insbesondere im amerikanischen und mitteleuropäischen Raum, ähnliche Verhaltensmuster bezüglich Kopfnicken feststellen, so existieren auch Kulturen, in denen teilweise sogar gegenläufige Interpretationen üblich sind. So wird beispielsweise in Bulgarien Kopfnicken als Ablehnung gedeutet (Hadar et al. (1985)).

In der Arbeit von Navaretta (Navarretta et al. (2012)) wurde gezeigt, dass sich in den skandinavischen Kulturen sogar innerhalb der verschiedenen Nationalitäten unterschiedliche Nickmuster zeigen. Obwohl Dänemark, Schweden und Finnland über weite Strecken eine gemeinsame Kultur teilen, zeigen sich dennoch Unterschiede im Kopfnickverhalten. So nicken die Dänen eher mit einer betonten 'Nach-Vorn-Unten'-Bewegung, während die Schweden eher eine betonte 'Nach-Oben'-Bewegung zeigen. In Finnland tritt besonders häufig vereinzeltes Nicken mit nur einem Zyklus auf.

Mehrfach belegt wurden auch wesentliche Eigenarten in der japanischen Nick-Kultur (Nunn and Maya (2003), Nori et al. (2011)). Im Vergleich zu Englisch-sprachigen Kulturen nicken Japaner häufiger, kürzer und öfter solidarisch.

Auffällig ist auch, dass trotz teils erheblicher kultureller Unterschiede in der Bedeutung des Kopfnickens in den bisher genannten Publikationen diese Geste scheinbar kulturunabhängig als eine Geste anerkannt ist, der eine Information zugesprochen wird.

Wenn nicht anders beschrieben, beziehe ich mich in dieser Ausarbeitung auf das Kopfnicken, das im überwiegenden westlichen Kulturkreis des englischsprachigen sowie des mitteleuropäischen Raumes als solches verwendet und verstanden wird.

3.5 Ansätze zur Simulation menschlichen Kopfnickverhaltens

Wiederholt wurde versucht, einem Roboter oder virtuellen Agenten ein möglichst natürliches Nickverhalten zu implementieren. Bamoallem stellte die Hypothese auf, Telepräsenz-Roboter würden durch Nick-Simulation eine höhere Gesprächs-Beteiligung wie verstärktes Blickverhalten, Lächeln oder Anzahl der Fragen und Antworten hervorrufen (Bamoallem et al. (2016)). Er konnte darin jedoch keine signifikanten Unterschiede zur Vergleichsgruppe feststellen. Mögliche Ursachen lägen jedoch an zu kurze Interaktions-Szenarien, an Sprachproblemen internationaler Teilnehmer, sowie an schlechtem Bild-Signal der Videoübertragung.

J. Lee arbeitete an einem Modell für Sprecher-Nicken (Lee et al. (2009b), Lee et al. (2009a)). Sie annotierte den linguistischen sowie den emotionalen Dialog-Kontext zu Video-Szenen mit Sprecher-Kopfnicken und trainierte mit Hilfe dieser Daten HMMs zur Vorhersage von Kopfnicken in Abhängigkeit auftretender Kontext-Informationen. Das Modell erzielte Erkennungsraten im Bereich von 80-90% und stellt eine vergleichsweise einfache Möglichkeit dar, beispielsweise ein Kopfnick-Verhalten für virtuelle Agenten zu generieren, ohne ausgefeilte Regelwerke festlegen zu müssen.

C. Liu stellte im Jahr 2012 ein Modell zur Generierung von Nicken vor (Liu et al. (2012)). Dabei wird Nicken vorwiegend während der letzten Silbe von Sprachabschnitten ausgeführt, sowie im Zustand des Zuhörens in Form von Backchannel-Signalen. Ein generiertes Nicken hat dabei eine Länge von etwa 0.4 bis 0.7 Sekunden.

H. Kihara veröffentlichte 2016 ausführliche Analysen und Statistiken zum Nickverhalten japanischer Studierender (Kihara et al. (2016)). Dabei wurde das Auftreten von Zuhörer-Kopfnicken während einer Gruppenarbeit beobachtet, um Daten für das Design eines simulierten Nickverhaltens für einen Roboter zu sammeln. Eine mittlere Nick-Periode dauerte demnach 0.27 Sekunden, wobei 96% aller Kopfnicken innerhalb des Bereiches zwischen 0.17-0.57 Sekunden lagen. Die mittlere Verzögerungszeit zwischen der Beendigung der Sprecher-Äußerung und der Nick-Reaktion des Zuhörers wurde mit 0.3 Sekunden beziffert, wobei 95% aller Zeitabstände zwischen -0.78 und 1.4 Sekunden lagen. 55% der Nicken bestanden aus einem einzelnen Zyklus. Zwei Zyklen kamen zu 24% vor, drei zu 12%, vier zu 3%

und fünf zu 2.1%. Lediglich 3% waren Nicken mit noch größerer Anzahl an aufeinanderfolgenden Zyklen.

3.6 Ausblick zur automatischen Interpretation

Nach umfassender Betrachtung relevanter Forschungsarbeiten zu Aspekten des menschlichen Kopfnickens lassen sich einige Schlüsse für eine automatische Verarbeitung schließen.

Zum einen stellen einige Arbeiten die großen Limitationen bei der Interpretation von Kopfnicken heraus. So schrieb Bamoallem et al. (2016), es sei schwierig, signifikante Schlüsse aus vereinzelt beobachteten Verhaltensmustern wie verbale Kommunikation, Gesichtsausdrücken oder Kopfnicken zu ziehen. Man müsse diese mit weiteren Modalitäten gemeinsam betrachten und beispielsweise die Körperhaltung miteinbeziehen. Auch Petukhova and Bunt (2009) waren zwar davon überzeugt, dass verschiedene Arten von Kopfnicken auch verschiedene Bedeutungen vermitteln können, betonten jedoch, diese ließen nur bedingt auf die tatsächliche Intention der aussendenden Person schließen.

Andererseits kann es auch dann nützlich sein, die innewohnende Funktion eines Kopfnickens zu verstehen, wenn sie nicht mit dem Grounding des Senders übereinstimmt, indem das interpretierende System sich anschließend womöglich genauso verhält, wie es der Sender mit dem 'falschen' Nicken beabsichtigte. Zudem stellen mehrere Publikationen wie die Untersuchungen von Hadar et al. (1985) und Birdwhistell (1970) heraus, die physikalischen Eigenschaften einer ausgeführten Kopfnick-Geste korreliere in signifikant vielen Fällen auch mit dessen kommunikativer Funktion.

Insbesondere Giges, Hadar und Nunn lieferten ausgearbeitete Kategorisierungen mit detaillierten, auch physikalischen Beschreibungen verschiedener Funktionsklassen von Kopfnicken. Auf dieser Basis werden im Kapitel 7 einzelne Aspekte herausgegriffen und auf die Möglichkeiten zur automatischen Interpretation hin untersucht.

Table 2 Interlocutor's nods

	1. Previous turn	2. Type	3. Meaning	4. Signal features
Speaker finished	Yes/no question	1.A1. Requested Confirmation	I confirm that what you hypothesize is true	Single nod, head first goes upward and then downward
	Information	1.A2. Spontaneous Confirmation	I confirm that what you say is true	Head movement downward
	Assessment	1.A3. Agreement	I agree with your judgement	Single nod downward, with head movement of high amplitude and tension. Gaze generally directed to Speaker, sometimes with a slow closing of the eyelids
	Proposal	1.A4. Approval	I approve	Nod downward single or repeated, with head movement of high amplitude. Gaze possibly directed to Speaker, often with eyebrow frowning
	Permission request	1.A5. Permission	I allow you to do this, I confirm that you may do this	
	Dominant request	1.A6 Submission	Yes, sir, I submit to you	
	Prosocial (communicative) action	1.A7. Greetings	I take a bow to you	Slow, generally accompanied by a smile and a closing of the eyes
		1.A8. Thanks	I thank you	Generally accompanied by a smile
Speaker continued	Backchannel	1.B1. Backchannel "I understand"	I confirm I am following you	Brief fast repeated downward movement
		1.B2. Backchannel "I take note"	I record your communicative act consider it relevant for our social relationship	short, repeated nods, gaze to speaker
		1.B3. Backchannel "I agree"	I confirm I agree	Brief fast repeated downward movement, frowning and gazing to Speaker, possibly smile
	Disagreement	1.B4. Ironic backchannel	I do not agree at all	Possibly asymmetrical (i.e. ironic) smile
	Self-agreement	1.B5. Back-agreement	I agree with you (but just) because you are repeating my previous statements. Do you acknowledge that I was right?	Possibly ironic smile, sometimes a sigh. Gaze to Speaker
		1.B6. Processing Nod	I am reasoning on what the speaker means and what he is aiming at, or, I am planning my response, and I approve of my reasoning	Repeated brief and slow downward movement, generally accompanied by frowning and possibly by a smile. Not gazing to Speaker

Table 4. Speaker's nods

1. Type	2. Goal or meaning	3. Signal features
3.1. Emphasis	This (part of my) sentence (discourse) is important	Head moves forward-downward over one stressed syllable Gaze to Interlocutor
3.2. Batonic	I stress syllables to help myself keeping rhythm	Repeated head movements downward in correspondence of more than one stressed syllable
3.3. List	This (part of my) sentence (discourse) is important because here starts an item of my list	Stresses the items in a list, often parallel with enumerating gestures
3.4. Interrogative nod	I ask you if you confirm or not my hypothesis	Gazes at Int. with oblique head, slightly tilted sideway. Or Eyebrow frowning like in interrogative sentences
3.5. Rhetorical interrogative nod	Isn't it so? I want you to confirm	
3.6. Backchannel request	I ask you if you confirm or not that you understand what I mean	(sometimes) accompanied by open hand palm up gesture

Abbildung 3.4: Verschiedene Typen für Kopfnicken nach Poggi et al. (2010).

4 Ein Assistenzsystem als Begleiter für Menschen mit Unterstützungsbedarf

Im Rahmen des Forschungsprojektes KOMPASS (Sozial kooperative virtuelle Assistenten als Tagesbegleiter für Menschen mit Unterstützungsbedarf) wurde ein Assistenzsystem mit dem Ziel entwickelt, hilfsbedürftigen Menschen im Alltag kognitiv und sozio-emotional unterstützen zu können (Kopp et al. (2018)). Die Basis bildet eine Umgebung für eine Dialog-basierte Mensch-Agenten-Interaktion. Diese besteht aus einem technischen System mit Bildschirm, Lautsprechern, Mikrofon und Kameras. Als virtueller Agent wird 'Billie' eingesetzt, der im Abschnitt 4.2 näher beschrieben wird.

4.1 Anwendungsszenario

Ein erprobtes Anwendungsszenario ist die interaktive Kalenderdomäne. Die Annahme ist, dass viele Menschen, vor allem in hohem Alter, in der Verwaltung ihrer Termine an ihre Grenzen stoßen, sich diese oft nicht mehr merken und überschauen können. Hierbei soll das Assistenzsystem unterstützen, indem es den entsprechenden Klienten interaktiv durch einen digitalen Kalender führt, Fragen stellt, zuhört und mit Hilfe des Klienten den Kalender mit dessen persönlichen Terminen korrekt ergänzt. Gelingt dies erfolgreich, kann das Assistenzsystem wiederum den Klienten rechtzeitig an die besprochenen Termine erinnern. Dabei erfolgt die Interaktion mithilfe von Sprache, aber auch mittels Kameras zur Gesichts- sowie Gestenerkennung.

Im Projektrahmen wurden zwei Studiendurchgänge durchgeführt.

In einer ersten Studie (WOZ1-Studie) wurde ein Wizard-of-Oz Szenario erstellt und der Nutzer durch eine festgelegte Dialog-Struktur geführt. Sowohl das Dialogmanagement, als auch das Vorschlagen und Eingeben von Terminen erfolgte automatisiert. Menschlich gesteuert wurde lediglich das

Starten und Beenden der Interaktion, das Wechseln zwischen verschiedenen Eingabe-Modi für die Nutzer, sowie das Auslösen von automatisch generierten Terminvorschlägen. So konnten gezielt Reparatur-Strategien und Feedback der Versuchsteilnehmer insbesondere bei für die Nutzer unerwartetem Verhalten des Systems und zeitlichen Verzögerungen untersucht werden. Die Versuchspersonen bestanden aus folgenden drei Kategorien: Studierende, Senioren, sowie kognitiv eingeschränkte Personen. Das Setup wird in der Abbildung 4.1 dargestellt.

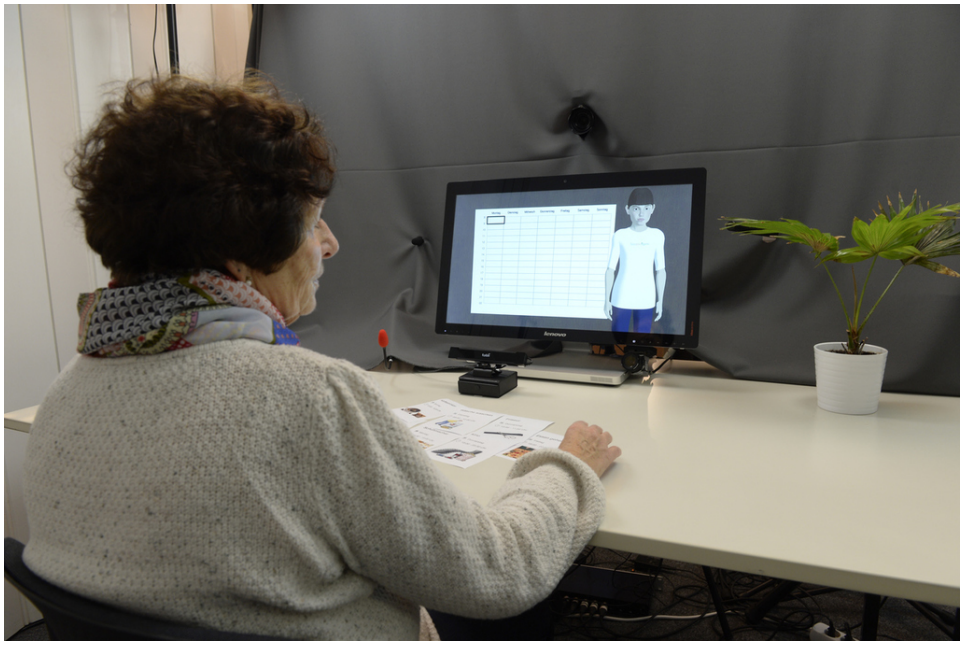


Abbildung 4.1: Versuchsperson der Wizard-of-Oz Studie (WOZ1) interagiert mit virtuellem Assistenzsystem.

Eine weitere Studie namens EVAL1 erfolgte mit einem vollautomatischen System. Hier wurden Erkenntnisse aus der ersten Studie umgesetzt, um Wechsel zwischen verschiedenen Eingabe-Modi autonom durchzuführen. Nutzer konnten eigenständig beliebige Termine in beliebiger Zeit in Kooperation mit dem Agenten eintragen und es gelang auch, den Dialog selbstständig zu beenden. Wie bei WOZ1 bestand die Menge der am Versuch Teilnehmenden wiederum aus Studierenden, Senioren und kognitiv ein-

geschränkten Personen. Die Abbildung 4.2 zeigt den Versuchsaufbau von EVAL1 aus der Perspektive von drei Kameras, sowie den Bildschirminhalt mit Billie.

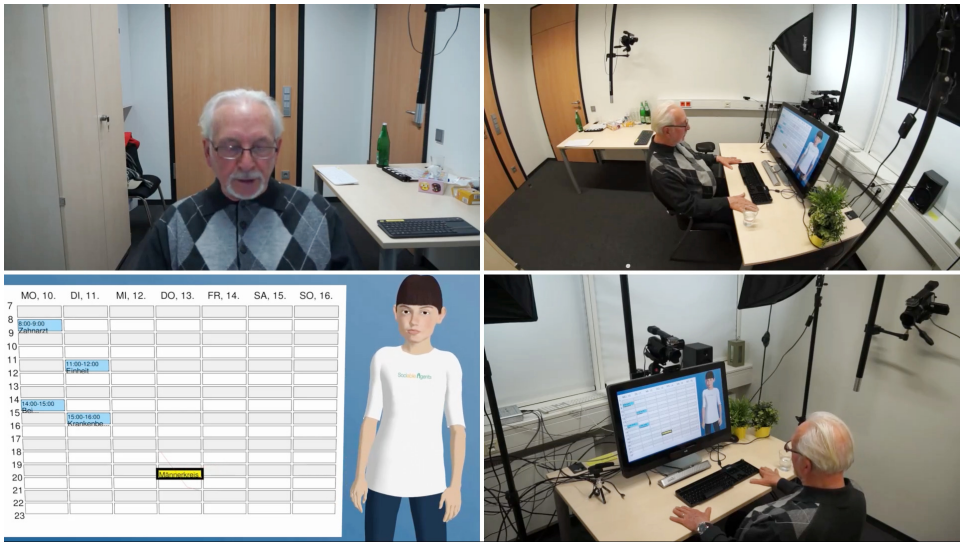


Abbildung 4.2: Perspektiven von drei Kameras zeigen das Setup der Nutzerstudie mit vollautomatischem Dialogsystem. Links unten befindet sich eine Bildschirmaufzeichnung mit einem interaktiven Kalender.

Insgesamt wurden über 100 Interaktionen mit Billie aufgezeichnet und umfangreich annotiert und untersucht. Besonders hervorzuheben ist hierbei, dass die Versuchspersonen zwar über die Videoaufzeichnung in Kenntnis gesetzt wurden, jedoch keinerlei Information dazu bekamen, ob und inwiefern ihre Gestik, insbesondere Kopfnickverhalten, vom System erkannt wird. Demnach ist sämtliches Auftreten von Kopfgestik als natürlich und ungezwungen anzusehen.

4.2 Agent Billie

Der virtuelle Agent Billie wurde im Jahr 2011 als Prototyp vorgestellt und von der AG Kognitive Systeme und soziale Interaktion in der Universität

Bielefeld entwickelt (Buschmeier and Kopp (2011)). Das Besondere an Billie ist der Anspruch eines 'attentive speakers', der kontinuierlich auf kommunikatives Nutzer-Feedback wartet, um darauf sensibel zu reagieren und das eigene kommunikative Verhalten so anzupassen, wie es der Nutzer mit hoher Wahrscheinlichkeit wünscht. Es handelt sich um einen gerenderten Agenten, der sowohl Hand-Gesten, als auch verschiedene Posen zeigen kann. Des Weiteren sind auch Mimiken implementiert. Auch Kopfbewegungen wie Nicken und Augenbewegungen wie das Anvisieren von Zielpunkten sind integrierte Funktionalitäten. Er spricht und wird oft als neutral weder männlich noch weiblich bestimmt.

5 Ein System zur Detektion von Kopfnicken

In diesem Kapitel wird das Konzept und die technische Umsetzung des Systems zur Erkennung von Kopfnicken beschrieben.

Zunächst werden die Anforderungen und Grundannahmen formuliert, die das System erfüllen soll. Daran schließt sich ein Überblick über die einzelnen Systemkomponenten sowie deren Zusammenhänge an. In den weiteren Abschnitten werden nun die einzelnen Komponenten näher ausgeführt. Zuerst wird die Wahl der Hardware im Bezug auf Sensorik und Recheneinheit beschrieben. Als erster Schritt nach Eingang des Bildsignals wird die Erkennung von Gesicht sowie die Extraktion von Gesichtsmerkmalen erläutert. Es schließt sich die verwendete Methodik zur Schätzung der Kopfausrichtung mittels Support Vector Regression (SVR) an, die die Grundlage für die benötigten Merkmalsvektoren schafft. Mit dessen Hilfe werden zwei Verfahren zur Klassifikation von Kopfnicken aus Zeitserien von Kopfwinkelschätzungen erarbeitet. Als erstes Verfahren wird eine Eigenentwicklung einer DTW-Variante ausgearbeitet. Als Gegenentwurf wird anschließend die Anwendung einer SVM zur Klassifikation im letzten Abschnitt beschrieben.

5.1 Systemanforderungen

Wie bereits in der Zielsetzung im Kapitel 1.2 beschrieben, ist es nicht das Ziel der Arbeit, mit leistungsstärkster Hardware maximale Leistungsfähigkeit zu erzielen, sondern mit Hilfe eines einfachen effizienten Konzepts eine solide Leistungsfähigkeit auch mit marktüblicher Hardware zu erreichen.

Als Grundvoraussetzung für solide Leistungsfähigkeit soll Kopfnicken zur Laufzeit innerhalb so kurzer Zeit erkannt werden, dass ein Kontext-Bezug potentieller weiterverarbeitenden Komponenten ermöglicht wird.

Es sollte keine Initialisierung mit der interagierenden Person erforderlich sein, da es die Interaktion für beliebige Nutzer erschweren würde.

Als Software soll aus Gründen der Transparenz, der freien Verfügbarkeit und der Unabhängigkeit überwiegend auf Open Source-Lösungen gesetzt werden. Zum Einsatz kommen unter anderem die Software-Bibliothek für Bildverarbeitung OpenCV (Bradski (2000)) und Dlib (King (2009)) für Maschinelles Lernen.

5.2 Systemüberblick

Das System lässt sich modular als mehrere hintereinander geschaltete Verarbeitungseinheiten beschreiben. In der Abbildung 5.1 findet sich ein Systemüberblick mit den entsprechenden Komponenten.

Zuerst werden mithilfe einer Kamera sukzessive Videosignale aufgezeichnet. Jedes Mal wird das entsprechende 2D Bild verarbeitet und auf das Vorhandensein eines menschlichen Gesichts hin überprüft. Im Erfolgsfall werden in dem lokalisierten Bildausschnitt Gesichts-Landmarken ausfindig gemacht und dessen Positionen als Merkmalsvektor an eine Support Vector Regression weitergereicht. Diese ist darauf spezialisiert, von den Merkmalen auf die Winkel der Kopfausrichtung zu schließen. Nun unterscheiden sich die Verarbeitungsschritte je nach gewähltem Verfahren zur Kopfnickdetektion. Wird DTW angewendet, so werden als Merkmale zusätzlich Ableitungen bestimmt. Ausgehend von dem aktuellen Wert werden Abfolgen zeitlich zurückliegender Werte mit einem prototypischen Nick-Beispiel abgeglichen und eine DTW-Distanz berechnet. Unterschreitet der Differenzwert einen Schwellwert, so wird Kopfnicken detektiert.

Wird jedoch eine SVM verwendet, so wird sukzessive ein Sliding-Window mit fester Fensterbreite über eine Anzahl der letzten extrahierten Merkmale gefahren. Die Werte innerhalb des Fensters werden normalisiert und bilden den Merkmalsvektor für eine SVM. Diese bestimmt die Distanz zu einer trennenden Hyperebene, wodurch ebenfalls eine Entscheidung für oder gegen eine Kopfnick-Detektion getroffen wird.

5.3 Hardware und Sensorik

Als erforderliche Eigenschaft der Bildquelle soll ein Qualitätsstandard eingehalten werden, der von dem Großteil aller marktüblichen Verbraucher-Kamerasysteme bewältigbar ist. Im Detail sind das aktuell beispielsweise

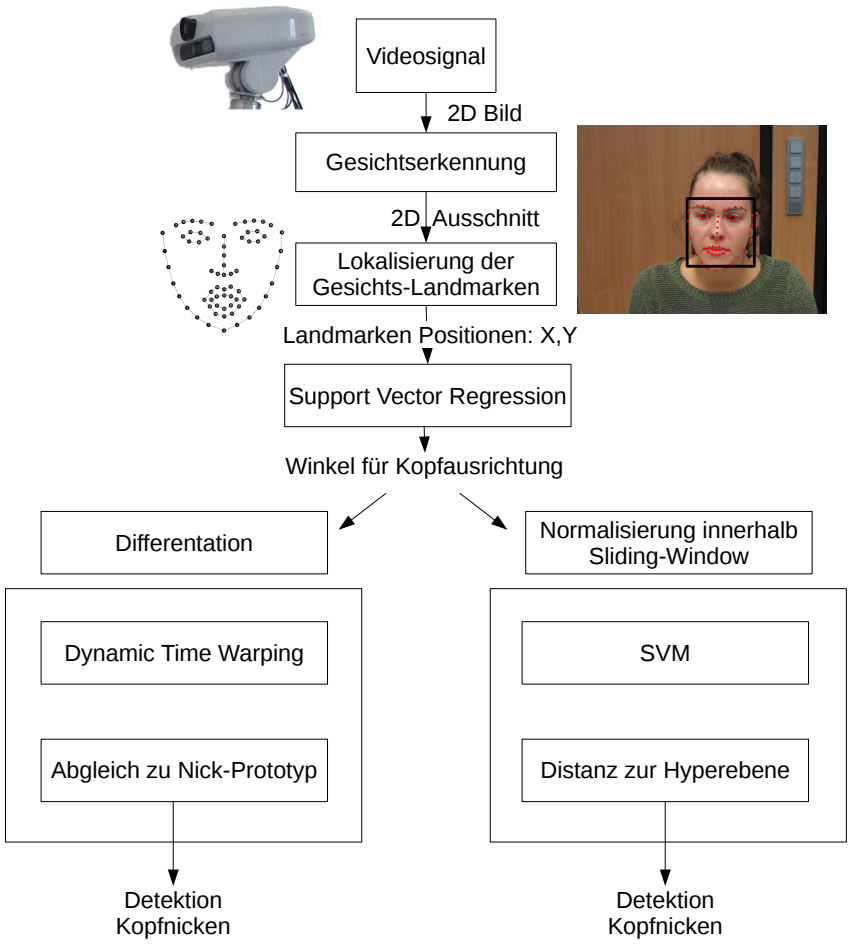


Abbildung 5.1: Ein Systemüberblick: Ausgehend vom Videosignal, Gesichtserkennung und Lokalisierung der Gesichts-Landmarken werden Kopf-Winkel als Merkmale extrahiert und die entstehende Zeitserie zur Kopfnick-Detektion entweder an eine DTW oder SVM weitergereicht.

einfache Webcams mit einer Auflösung von 640x480 (VGA) und einer Framerate von 30 fps. So wurden auch die Videos zur Evaluation des Detektions-Systems für Kopfnicken überwiegend mit der Webcam Logitech C920/C922 frontal aufgezeichnet.

Die notwendige Rechenleistung soll von marktüblichen Rechnern und Notebooks ohne GPU-Unterstützung aufzubringen sein.

Als repräsentatives Test-System kam ein Notebook mit einem Core i7 Prozessor sowie 16 GB RAM Arbeitsspeicher zum Einsatz.

5.4 Gesichtsmerkmale

Einer der ersten Verarbeitungsschritte der von der Kamera gelieferten Einzelbilder ist die Gesichtserkennung. Das Erkennen eines Gesichts und die Lokalisierung dessen Bildbereiches ist Grundvoraussetzung aller weiteren beschriebenen Verarbeitungsschritte.

Über gängige Verfahren zur Gesichtserkennung wurde im Kapitel 2.3 ein Überblick gegeben. Aus Gründen der Effizienz wurde ein Algorithmus auf Basis von HOG-Merkmalen verwendet. Im Speziellen wurde dabei auf den frei verfügbaren HOG Face Detector der Dlib-Bibliothek zurückgegriffen, der den verbreitetesten OpenCV-Erkenner vor allem in der Geschwindigkeit übertrifft, sowie für den Einsatz bei Gesichtern in der Auflösung über 70x70 Pixel empfohlen wird *(https://www.learnopencv.com/face-detection-opencv-dlib-and-deep-learning-c-python)* (2019).

Der HOG Face Detector kann bei Bedarf parametrisch angepasst werden. So kann beispielsweise das zu untersuchende Bild durch Upsampling größer skaliert werden, um auch sehr kleine Gesichter zu erfassen. Im Fall einer erfolgreichen Detektion wird die Gesichtsposition in Form eines Rechtecks wie in Abbildung 5.2 zurückgegeben.

Innerhalb dieses Rechtecks sollen Merkmale extrahiert werden, die einen Mehrgewinn an Information über die Kopfausrichtung liefern. Dafür wurde ein Alignment-Modell mit einem Set von 68 definierten Gesichts-Landmarken verwendet, wie es im Kapitel 2.4 beschrieben wurde. Der Grundgedanke dabei ist, dass die Distanzverhältnisse einzelner Gesichtspunkte im Bezug zu einer frontal positionierten Kamera von der Kopfausrichtung abhängig sind und diese somit ein Mehrgewinn an Information über die Kopfausrichtung liefern. Dazu wurde ebenfalls eine Dlib-Implementation

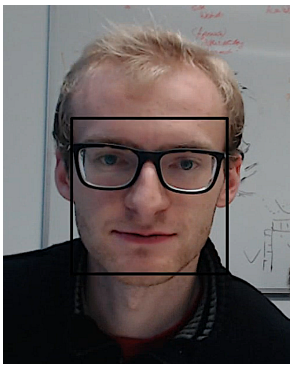


Abbildung 5.2: Der Erkennung von Gesichtsmerkmalen geht eine Eingrenzung des Bildbereichs mittels eines Gesichtserkenners voraus. Hier wurde der Gesichtsbereich als schwarzes Rechteck eingerahmt.

des im Kapitel 2.4 beschriebenen Alignment-Modells von Kazemi and Sullivan (2014) verwendet, das aus einem Set von 68 Landmarken besteht. Eine Beispielaufnahme aus der Anwendung zeigt die Abbildung 5.3.

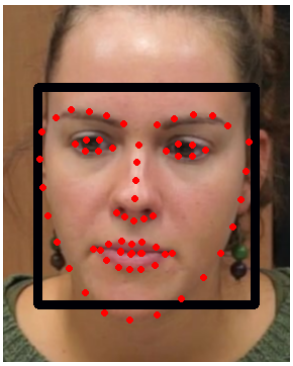


Abbildung 5.3: Hier zeigt sich das Alignment-Modell für 68 Landmarken in der Anwendung

5.5 Kopfwinkelschätzung mittels SVR

Auf der Basis der Gesichtsmerkmale werden nun Informationen bezüglich der Kopfausrichtung gewonnen. Da Kopfbewegungs-Trajektorien durch eine Serie von geschätzten Kopfwinkeln repräsentiert werden, sind diese Merkmale besonders wichtig. Im Kapitel 5.5 wurden gängige Verfahren zur Kopfwinkelschätzung vorgestellt.

Da es mir wichtig ist, bevorzugt die anatomisch korrekte Form des Nickens im Sinne der Nick-Drehachse gemäß der Abbildung 2.2 zu erfassen und von anderen vertikal anmutenden Bewegungen des Kopfes wie Translation unterscheiden zu können, soll auf Verfahren mittels Optical Flow und Tracking verzichtet werden.

Auch halte ich hier den Einsatz eines Regressions-Verfahrens für sinnvoller als eine Menge von Klassifikatoren, die die erreichbare Genauigkeit einschränken, wie im Kapitel 5.5 erläutert wurde.

Insbesondere da bei Kopfgesten im Endeffekt nicht nur bestimmte Kopfausrichtungen, sondern ganze Trajektorien auch in feinen Bewegungsabläufen mit hoher Genauigkeit zu untersuchen sind, soll somit auf ein Regressions-Verfahren zurückgegriffen werden.

Auch wenn mit Zuhilfenahme von Tiefeninformation einer Stereo-Kamera zwar die bisher besten Ergebnisse zu erwarten sind, so sollten die Ergebnisse mit den gegebenen Merkmalen der 2D-Kamera zumindest annähernd im Bereich des aktuellen Forschungsstandes zu verorten sein.

Als Datenbasis für das Training einer Support Vector Regression (SVR) wurde auf die freie Datenbank von Kinect-Aufnahmen mit annotierter Ground Truth inklusive 2D-Aufnahmen verwendet (Fanelli et al. (2011)). Dabei handelt es sich um 24 Sequenzen von 20 verschiedenen Personen, die einen Meter entfernt vor einer 3D-Kamera sitzen und ihren Kopf in alle möglichen Bewegungsrichtungen wenden sollten. Die Abbildung 5.4 soll durch einen Ausschnitt an Beispielbildern in etwa die Abstufungen der extrahierten Beispiele zeigen. Durch ein offline angewendetes Tracking-Verfahren namens ICP (Morency and Darell (2002)) wurde eine Kopfausrichtung bestimmt.

Die gesamte Datenbasis umfasst etwa 15000 Einzelbilder mit Annotationen der Kopfausrichtung in Form von Euler-Matrizen, aus denen alle drei Kopfwinkel extrahiert werden können. Als Vereinfachung sollen für Kopfnicken hauptsächlich der Nick- und teilweise der Gier-Winkel (Vertikale und

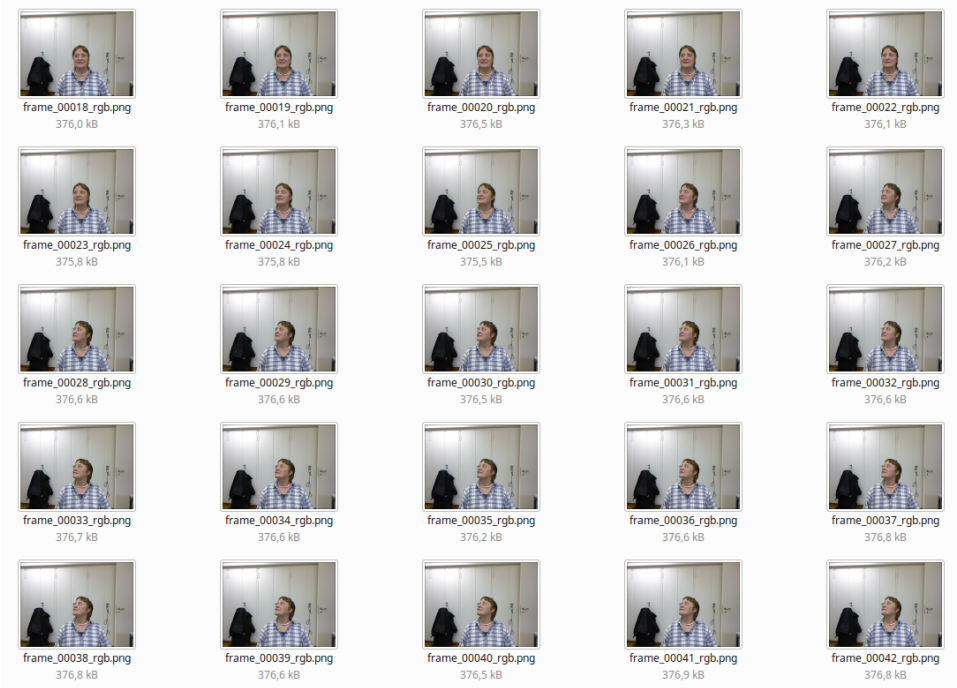


Abbildung 5.4: Diese Abbildung zeigt einen kleinen Ausschnitt der extrahierten Beispielbilder des 3D-Kinect Datensatzes von Fanelli et al. (2011)

.

Horizontale) herangezogen werden. Die Datenbasis hält dafür Annotationen in einem Bereich um ±75° für die Horizontale und um ±60° für die Vertikale bereit.

Obwohl die Daten mit einer Tiefenbild-Kamera aufgenommen wurden, so zeichnete die Kamera zusätzlich auch 2D-Bilder in VGA-Auflösung auf. Um die aus den 3D-Daten gewonnenen Winkeln auch ausschließlich aus den 2D-Daten annähernd zu reproduzieren, werden die entsprechenden annotierten 3D-Winkel mit den 2D-Daten gekoppelt. Da diese Daten sehr genau sind, bieten sie eine gute Grundlage für das Training einer SVR.

Bei der Auswahl der Trainingsdaten gibt es viele Herausforderungen zu bewältigen.

Um sicherzustellen, dass jedes Bild auch als Trainingsmaterial geeignet ist, wurden diejenigen Einzelbilder aussortiert, auf denen von dem verwendeten Gesichtsdetektor kein Gesicht erkannt wurde. Für jedes Bild, auf dem ein Gesicht erkannt wird, wird auch ein Alignment-Modell berechnet, welches gerade in Grenzfällen wie beispielsweise stark seitwärts gewandtem Gesicht teilweise sehr mangelhafte Ergebnisse liefern kann.

Es ist zumindest fragwürdig, ob man nur Trainingsdaten nehmen sollte, mit denen das Alignment augenscheinlich gut funktioniert. Nach einigen Versuchen hat sich herausgestellt, dass sich keine signifikante Änderung der Leistungsfähigkeit zeigte, wenn man mangelhafte Alignments aussortiert. Dies kann darin begründet sein, dass einerseits durch das Beibehalten auch von mangelhaften Alignments für das Training auch bei der Erkennung die Toleranz gegenüber schlecht matchenden Alignments steigt und somit die Robustheit. Andererseits kann die gewonnene Toleranz auf Kosten der Genauigkeit in den 'sicheren' Bereichen gehen, wodurch sich die Effekte in etwa ausgleichen.

Deshalb wurden diejenigen Einzelbilder herangezogen, auf denen auch die im Kapitel 2.4 beschriebenen Gesichtsmerkmale augenscheinlich angemessen extrahiert werden konnten.

Die somit ausgewählten Trainingsdaten sind bezüglich ihres Ground Truth Winkels in der Regel nicht gleichverteilt, wodurch zu erwarten ist, dass Wertebereiche, die überdurchschnittlich häufig im Datensatz vertreten sind, durch das Trainingsverfahren stärker gewichtet und andere Bereiche umso schlechter erkannt werden. Jedoch sollte das Ziel sein, möglichst eine konstante Leistungsfähigkeit über einen größtmöglichen Definitionsbereich zu erreichen, der für den speziellen Anwendungsfall sinnvoll ist.

Ein Blick auf die Häufigkeitsverteilung der Winkelannotationen, zu sehen in der Abbildung 5.5, zeigt, dass sich sowohl bei der vertikalen als auch bei der horizontalen Bewegungsrichtung in der Mitte um 0 Grad das höchste Datenaufkommen befindet. Dieses nimmt zu größeren Winkelwerten hin sowohl im positiven als auch im negativen Bereich kontinuierlich ab.

Man könnte einfach mit allen Daten trainieren und argumentieren, dass extreme Kopfwinkel seltener vorkommen und die Verteilung somit die Realität widerspiegelt, jedoch hat sich bei ersten Tests gezeigt, dass gerade bei Gesten die seltener vorkommenden extremeren Kopfwinkel wichtig sind und deren Genauigkeit so im Training leidet.

Möchte man also eine Teilmenge entnehmen, die annähernd einer Gleich-

verteilung entspricht, tritt folgendes Problem auf: Wenn man extremere Datenbereiche berücksichtigen will, die selten vorkommen, so können auch von den häufig zur Verfügung stehenden Daten sehr wenige verwendet werden. Je nach Verteilung der Gesamtmenge muss erörtert werden, ob auf die Berücksichtigung seltener vorkommender Werte zugunsten einer größeren Datenmenge verzichtet werden sollte oder nicht. Die Darstellung 5.5 zeigt eine für diesen Anwendungsfall als sinnvoll erachtete Kompromiss-Lösung.

Ein weiterer zu beachtender Effekt des Trainings mittels SVR ist die Auswirkung der Reihenfolge der Trainingsdaten. So werden im Training nicht alle Trainingsdaten gleichzeitig verarbeitet, sondern die Regressionsfunktion mit jedem Trainings-Sample in Abhängigkeit aller bisher verarbeiteten Samples sukzessive angepasst. Deshalb ist eine Randomisierung der Reihenfolge der Trainingsdaten wichtig, um Spezialisierungseffekte in einem bestimmten Datenbereich zu vermeiden.

Nach der Vorverarbeitung der Trainingsdaten blieben etwa 5600 Bilder für die horizontale Drehbewegung und etwa 1600 Bilder für die vertikale Drehbewegung übrig. Mit diesen Trainingsdaten wurden zwei voneinander unabhängige SVR's für Gier- und Nick-Winkel trainiert.

Mithilfe des 'Grid-Search'-Verfahrens und einer dreifachen Kreuzvalidierung wurde eine Parameter-Optimierung durchgeführt.

Der Regressionsfehler des besten Modells lag in der Nick-Richtung bei einer Standardabweichung von 6.3° und in der Gier-Richtung bei einer Standardabweichung von 6.2°.

Verglichen mit den Ergebnissen relevanter Publikationen sind dies annehmbare Werte. So liegen die Fehler-Werte der in dem Survey von Murphy-Chutorian and Trivedi (2009) gesammelten Publikationen meist darüber. Einige jedoch mit wesentlich besseren Pitch-Werten wie Wang and Sung (2007) mit 2.56° oder Oka (2005) mit 2.0° wurden mit Spezial-Hardware wie Lagesensoren oder Stereo-Kameras erreicht.

5.6 Dynamic Time Warping zur Kopfnickdetektion

Im Kapitel 2.7.1 wurde der grundlegende Algorithmus für Dynamic Time Warping beschrieben. Zur Anwendung für die Kopfnickerkennung in Echtzeit sind einige Anpassungen erforderlich, die im Folgenden detailliert beschrieben werden. Dazu gehört die Verarbeitung der Daten in einem kontinu-

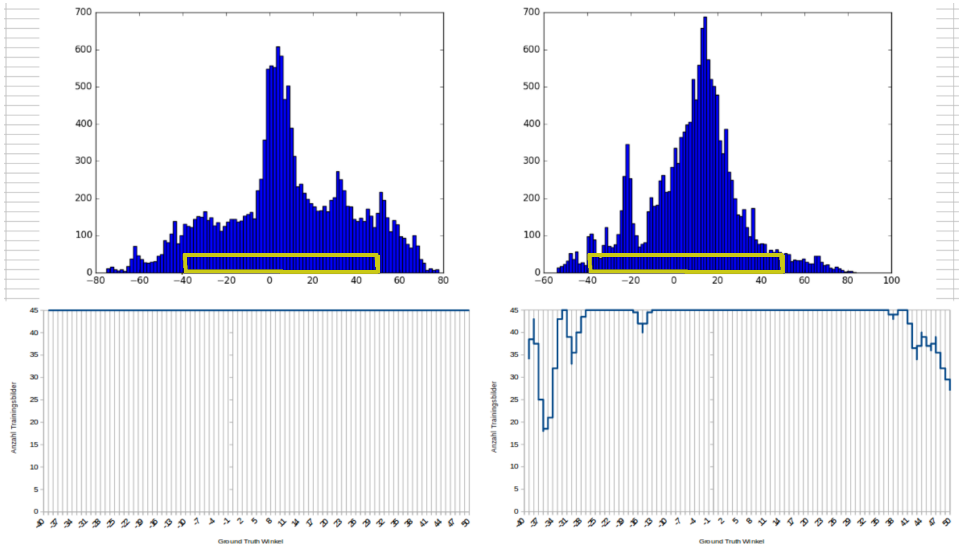


Abbildung 5.5: Jeweils markiert in einem schmalen Rechteck befinden sich links die Daten der horizontalen Kopfwinkel und rechts die vertikalen annotierten Kopfwinkel der verfügbaren Datenmenge. Die umrahmten Bereiche zeigen eine gemäß der Häufigkeit der Winkelintervalle gleichverteilte Teilmenge, die für das Training aus der Gesamtverteilung extrahiert wurde.

ierlichen Datenstrom, Verwendung von Slope Constraints, die Konfiguration der Kostenfunktion, sowie die Vorverarbeitung der Merkmale. Grundzüge der Bearbeitung des Themas DTW zur Kopfnickerkennung habe ich bereits veröffentlicht (Wall et al. (2017)).

5.6.1 Online Dynamic Time Warping

Der Begriff 'online' meint im gegebenen Kontext die Eigenschaft, dass die Erkennung auch auf kontinuierlich wachsenden Zeitserien vonstatten gehen kann. Innerhalb der Zeitserien sind keine definierten Zeitpunkte bekannt, an denen ein Event wie eine Geste beginnt oder endet.

Wie jedoch bereits beschrieben, sollte bei der Evaluation der Distanz-

Pfade zweier zu untersuchender Zeitserien insbesondere deren Endzeitpunkt bekannt sein.

Dies ist bei kontinuierlich wachsenden Zeitserien nicht gegeben, weshalb einige Anpassungen erforderlich sind.

Bei kontinuierlich wachsenden Sequenzen entstehen stets neue mögliche Teilsequenzen, deren Längen von dem aktuellen Warping-Pfad abhängen. Um diese inkrementellen Aktualisierungen der zu vergleichenden Sequenz zu handhaben, kommt eine Modifikation von DTW zum Einsatz, die in ähnlicher Form in der Literatur *'subsequence dynamic time warping'* genannt wird (Müller (2007)).

Zu jedem neuen Zeitpunkt wird die Kostenmatrix C_A um eine neue Spalte ergänzt.

Diese wird gemäß der Gleichung 2.1 mit den akkumulierten Distanzen für den aktuellen Zeitpunkt ausgefüllt.

Für eine zu dem aktuellen Zeitpunkt j endende Teilfolge steht die minimale Distanz $D_{M,j}$ in der untersten Zeile.

Im klassischen Fall von Offline-DTW beginnt der Warping-Pfad bei dem fixen Punkt $C_A(0,0)$. Für diesen Zielpunkt muss ein dynamischer Ersatz definiert werden. Dieser ist innerhalb der kontinuierlich wachsenden Zeitserie B zu verorten. Dafür wird der Fall $(i = 0, j > 0)$ in der Gleichung 2.1 aus dem Kapitel 2 zu $d(0,j)$ neu definiert. Daraus folgt die Gleichung 5.1.

$$C_A(i,j) = \begin{cases} d(0,j), & i = 0 \\ d(i,0) + C_T, & i > 0, j = 0 \\ d(i,j) + min(C_L, C_{LT}, C_T), & i > 0, j > 0 \end{cases} \tag{5.1}$$

Nun befindet sich in der letzten Zeile von C_A für jeden möglichen Vergleich zwischen den Zeitserien $A = [a_0, ..., a_M]$ and $B_k = [b_k, ..., b_N]$ die jeweilige minimale Distanz $D_{M,j}$. Schließlich wird ein Kopfnicken genau dann erkannt, wenn $D_{M,j}$ einen bestimmten Distanz-Schwellwert unterschreitet. Das Zeitintervall für das erkannte Nicken wird für den Zeitraum zwischen den Punkten k und N definiert, was den Start- und Endpunkt des entsprechenden minimalen Warping-Pfades festlegt. Dies kann auch in der Abbildung 5.7 nachvollzogen werden.

Um Überlappungen mit möglicherweise kurz hintereinander erkannten Kopfnicken zu vermeiden, wird nach einer Detektion die Spalte N mit dem

Wert *Unendlich* neu initialisiert, sodass erst ab der folgenden Spalte der Beginn einer neuen Nick-Geste erkannt werden kann.

5.6.2 Merkmale und deren Vorverarbeitung für DTW

Für jedes Einzelbild, auf dem auch ein Gesicht erkannt wird, wird von der SVR eine neue Kopfwinkelschätzung geliefert. Daher besteht eine grundlegende Zeitserie unseres Anwendungsszenarios aus einer Abfolge von Kopfwinkel-Schätzungen in der Grad-Einheit. Die wohl wichtigsten Werte sind dabei die vertikalen Kopfwinkel, da hier für Kopfnicken die größte Informationsdichte zu erwarten ist.

Je nach Grundstellung der Kopfausrichtung vor und nach dem Kopfnicken, nehmen die Schätzwerte für die vertikalen Winkel eher höhere oder niedrigere Werte an. Diese Variation des Versatzes in der Menge der Kopfnick-Trajektorien, auch Offset genannt, stellt eine Schwierigkeit für DTW dar. Werden nämlich zwei an sich identische Trajektorien ausgehend von verschiedenen Grundstellungen des Kopfes ausgeführt, so ergibt ein Vergleich mittels DTW dennoch eine große Differenz.

Aufgrund des dynamischen Charakters des Online DTW lässt sich diese Offset-Variation auch nicht durch Normalisierung eliminieren.

Eine Möglichkeit, die durch die verschiedenen Kopf-Grundstellungen bedingte Variation in den Daten zu umgehen, ist die Verwendung von Differenzwerten der berechneten Winkel als Merkmale anstelle der Winkel selbst.

Jedoch geht mit der Differentation auch ein gewisser Informationsverlust einher. Bereits bei der Kopfwinkel-Schätzung ist aufgrund der unvermeidbaren Messungenauigkeiten ein mehr oder weniger starkes Rauschverhalten zu erwarten. Daher ist oft der Einsatz eines Glättungsfilters wie ein gleitender Mittelwert oder Gauss-Filter zu empfehlen. Insbesondere bei weiterer Verarbeitung der Daten zu Ableitungen verstärkt sich das Rauschverhalten enorm.

Dem soll mit einem speziellen Filter entgegengewirkt werden. Ein Glättungsfilter, der die Glättung der Daten mit deren Ableitung in einem Schritt vereint, ist der Filter von Savitzky und Golay(Savitzky and Golay (1964)). Gegenüber einem gleitenden Mittelwert und dem Gauss-Filter hat dieser den besonderen Vorteil, die Spitzen und lokalen Extrema inklusive derer Charakteristiken besser zu erhalten. Die allgemeine Formel des Savitzky-Golay-Filters ist in der Gleichung 5.2 beschrieben, in der n die

Länge des Filterfensters kodiert:

$$y_t = \frac{1}{h} \sum_{i=-\frac{n-1}{2}}^{\frac{n-1}{2}} a_i x_{t+i} \tag{5.2}$$

Hierbei bezieht sich x_t direkt auf die unverarbeiteten Kopfwinkel-Schätzwerte. Um die geglättete Ableitung zu bestimmen, muss a_i gemäß Savitzky-Golay auf i gesetzt werden. In der folgenden Tabelle sind die entsprechenden Normalisierungs-Faktoren h gelistet:

Tabelle 5.1: Normalisierungsfaktoren für den Savitzky and Golay Filter

filter length	h
5	10
7	28
9	60
11	110

Von empirischen Tests ausgehend wurden die Koeffizienten auf $n = 9$ und $h = 60$ gesetzt.

In der Abbildung 5.6 werden vertikale Kopfwinkel und deren Ableitungen von zwei hintereinander ausgeführten Kopfnickgesten grafisch dargestellt.

Eine schematische Abbildung von Online Dynamic Time Warping wird in der Abbildung 5.7 gezeigt.

5.6.3 Slope Constraints in Online Dynamic Time Warping

Gemäß bisheriger Definitionen wurden für den Warping-Pfad zwar Konditionen für den Start- und Endpunkt beschrieben, jedoch für den Pfad dazwischen keine Einschränkungen festgelegt.

Dies hat möglicherweise die Auswirkung, dass bei der Bestimmung der DTW Distanz an einer Stelle einer Zeitserie sehr lange pausiert wird. Da-

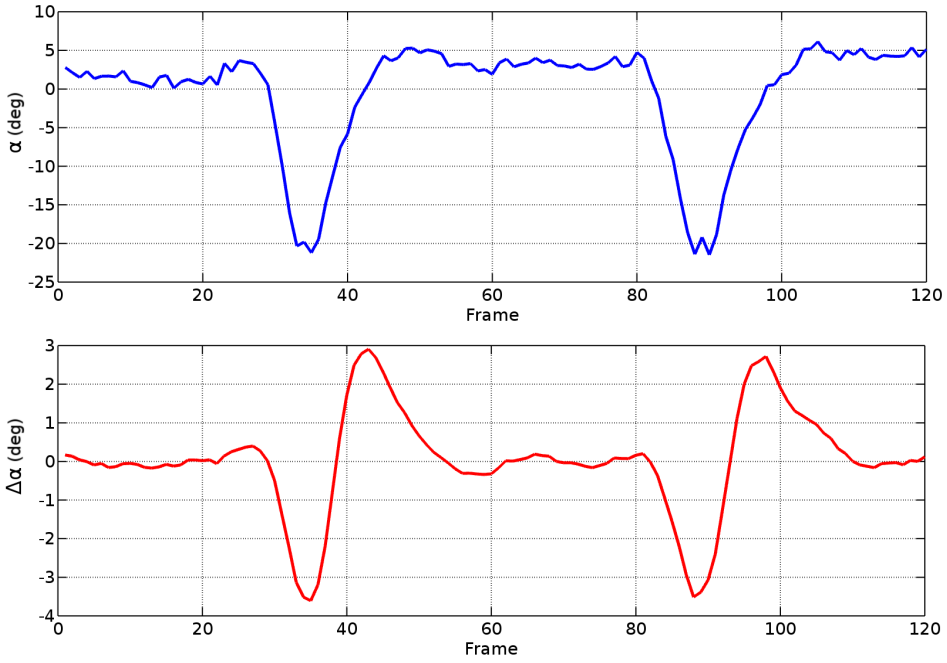


Abbildung 5.6: Vertikaler Kopfwinkel (blau) und deren geglättete Ableitungen (rot) zweier typischer Kopfnick-Gesten.

durch würden kurze mit sehr langen Zeitserien miteinander verglichen, was in unserem Fall nicht erwünscht ist.

Wir wollen daher dafür sorgen, dass der Warping-Pfad innerhalb der DTW Distanz Matrix nicht zu stark von der Diagonalen abweicht. So kann der Effekt verhindert werden, indem der Verlauf des Warping-Pfades insofern eingeschränkt wird, dass dieser nur eine begrenzte Anzahl dieselbe direkt aufeinanderfolgende Richtung (horizontal oder vertikal) annehmen darf.

Nach empirischem Testen hat sich dabei ein Maximum von 2 Schritten bewährt.

Der Raum möglicher Pfade mit der Beschränkung auf zwei Schritte gleicher Richtung wird in der Abbildung 5.8 schemenhaft dargestellt.

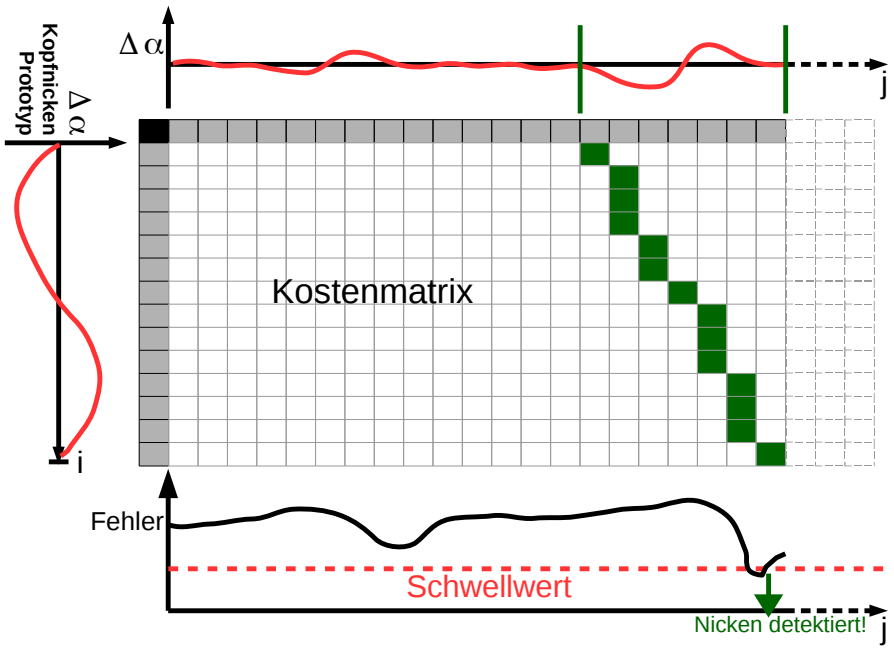


Abbildung 5.7: Detektion von Kopfnicken mittels Online Subsequence Dynamic Time Warping. Links zu sehen ist die prototypische Nickbewegung dargestellt als zeitliche Abfolge von vertikalen Geschwindigkeitswerten. Oben befindet sich die kontinuierlich wachsende Zeitserie der von der Kamera in Echtzeit erfassten Geschwindigkeitswerte. Unten wird die Kostenfunktion dargestellt, die bei Unterschreitung eines Schwellwertes eine Detektion auslöst. Vom Zeitpunkt der ausgelöster Detektion ausgehend, lässt sich anhand der Kostenmatrix der Warping-Pfad zurückverfolgen, um das Intervall des Detektionsfensters zu bestimmen.

5.6.4 Normalisierung der Kostenfunktion

Wie bereits beschrieben, wurden Ableitungs-Merkmale integriert, um den Ruhe-Offset zwischen verschiedenen Nickmustern zu kompensieren. Neben den Ruhe-Offsets und der zeitlichen Variationen, die durch DTW gut er-

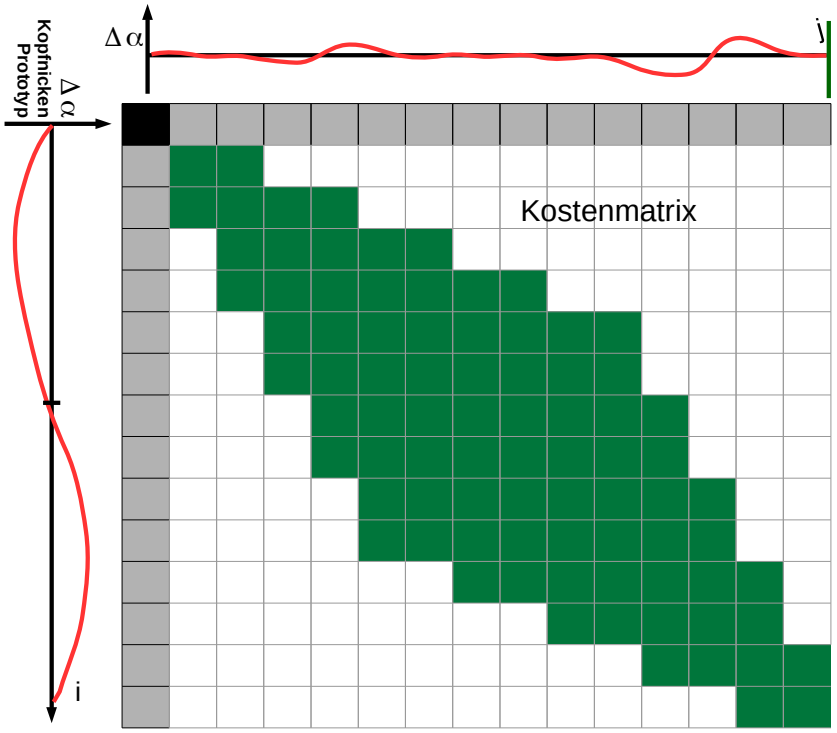


Abbildung 5.8: Diese Abbildung zeigt den Raum der möglichen Warping-Pfade mit dem festgelegten Maximum von 2 aufeinanderfolgenden Schritten gleicher Richtung.

fasst werden können, ist ein weiteres Kriterium die unterschiedlich stark ausgeprägte Amplitude eines Nick-Zyklus'. Dies sollte jedoch berücksichtigt werden, da Kopfnicken durchaus mit unterschiedlichen Geschwindigkeiten ausgeführt werden kann. Allgemein wurde dabei folgende bemerkenswerte Beobachtung deutlich: Vergleicht man zwei sehr ähnliche Zeitserien, die beide große Amplituden aufweisen, so ist deren Differenzwert (DTW-Distanz) größer als die DTW-Distanz zwischen zwei ebenfalls ähnlichen Zeitserien mit geringer Amplitude. Würde man zum Beispiel die Merkmale zweier Zeitserien mit demselben Faktor größer skalieren, so bleibt die wahrgenommene Ähnlichkeit zwischen den beiden zwar identisch, die absoluten Distanzwerte

vergrößern sich jedoch ebenfalls, was insgesamt zu einer größeren DTW-Distanz führt.

Bei der Auswahl eines Prototypen als Referenz für die Detektion ist daher der Ansatz, einen Normalisierungs-Faktor bei der Kostenberechnung einzuführen, der diesen Distanzunterschieden entgegenwirkt. Dieser Normalisierungs-Faktor für die Kostenfunktion soll die Amplitude der Teilsequenz berücksichtigen. Dazu wird als Näherungslösung zur Bestimmung der Amplitude die Standardabweichung aller Werte der mit dem Prototyp zu vergleichenden Teilsequenz bestimmt. Dies ist möglich, da diese Standardabweichung mit der Amplitude der Teilsequenz in etwa linear korreliert. Nun wird der summierte Distanzwert in der letzten Zeile der Akkumulierten Kostenmatrix mit der entsprechenden Standardabweichung normalisiert.

5.6.5 Auswahl des repräsentativen Prototypen

Die ein Kopfnicken beinhaltende Zeitserie, die als Referenz zu allen weiteren Sequenzen verwendet wird, sollte sorgfältig ausgewählt werden, weil sie die Gesamtheit der Kopfnick-Gesten repräsentiert. Denkbar wäre eine Mittelung von vielen Kopfnick-Gesten zu einer neu generierten Zeitserie. Bessere Ergebnisse sind zu erwarten, wenn aus einer Auswahl von potentiellen Prototypen jeder einzelne systematisch anhand eines Testdatensatzes evaluiert wird, um herauszufinden, welcher davon am besten generalisiert. Dieser Ansatz, sowie entsprechende Evaluationsmaße und Ergebnisse werden im Kapitel 6 ausführlich beschrieben.

5.7 Support Vector Machine zur Kopfnickdetektion

Als Referenz und Alternative zum DTW Verfahren soll eine Support Vector Machine für den Einsatz zur Detektion von Kopfnicken genutzt werden. Während mit DTW aufgrund spezieller Anpassungen eine vollständige Eigen-Implementation entstand, wird für den Einsatz von SVM die Implementation aus der Software-Bibliothek Dlib verwendet. Dabei liegt der Hauptaugenmerk auf die Zusammensetzung und Verarbeitung geeigneter Merkmalsvektoren, sowie die Realisierung eines angemessenen Training-Verfahrens. Als besonderen Unterschied zur DTW wird zunächst die Notwendigkeit einer fixen Fensterbreite beschrieben und anschließend die Zusammensetzung der Merkmale detailliert beschrieben.

5.7.1 Fensterbreite

Ein wesentlicher Unterschied zum DTW-Verfahren ist die Notwendigkeit einer fixen Dimensionalität der Merkmalsvektoren, da die Bestimmung einer Trennfunktion einen gemeinsamen Merkmalsraum mit den zu trennenden Merkmalsvektoren voraussetzt. Im Falle von Zeitserien werden Merkmalsvektoren daher in der Regel aus Zeitfenstern fixer Länge extrahiert.

Das bedeutet für einen Kopfnick-Erkenner eine Festlegung auf eine fixe Fensterlänge, die darauf untersucht werden soll, ob es sich bei der Sequenz um ein Kopfnicken handelt oder nicht. Da das Zeitintervall, in dem ein Kopfnicken stattfindet, zeitliche Variation aufweist, stellt sich dabei die Frage, wie groß die Fensterbreite gewählt werden soll.

Die Entscheidung für eine Abbildung von dynamischer zu fixer Fensterlänge soll möglichst statistisch begründet werden. Die Herausfoderung besteht darin, einerseits möglichst viel Information auch der längeren Zeitspannen zu erhalten und andererseits so wenig zusätzliches Rauschen und unnötig hohe Dimensionalität durch lange Zeitspannen zu vermeiden.

Als sinnvoll erwieß sich zum Beispiel, das Fenster so groß zu wählen, dass 90 Prozent aller annotierten Kopfnicken das gewählte Zeitintervall in der Länge nicht überschreiten.

5.7.2 Merkmale und deren Vorverarbeitung

Neben der Fensterbreite soll nun die Zusammensetzung geeigneter Merkmalsvektoren festgelegt werden. Wie auch in dem Abschnitt 5.6.2 für DTW beschrieben, sind die vertikalen Kopfwinkel auch die wichtigsten Merkmale für SVM. Der Miteinbezug von Werten für die horizontale Kopfausrichtung könnte dabei helfen, um andere Kopfgesten mit vertikaler Bewegungskomponente nicht fälschlicherweise als Nicken zu interpretieren.

Bei einer Anordnung von vertikalen und/oder horizontalen Kopfwinkeln zu einem Merkmalsvektor ist jedoch zu beachten, dass je nach Ausgangslage der Kopfausrichtung vor und nach dem Nicken, der Wertebereich auch bei scheinbar identischen Nickgesten stark variieren kann und somit ein großer Suchraum gebildet wird. Dies kann kompensiert werden, indem zum Beispiel entweder Differenzwerte beziehungsweise Ableitungen über die Kopfwinkel als Merkmale herangezogen werden, wie es bei der Anwendung von DTW im Abschnitt 5.6.2 durchgeführt wurde.

Bei der Berechnung von Ableitungen kann jedoch viel Information verloren gehen. Während eine Normalisierung der Zeitserie bei DTW aufgrund des dynamischen Zeitfensters nicht möglich war, so ist hier dank der fixen Fensterbreite eine Normalisierung jedes Merkmalsvektors zu der Standardabweichung 0, auch Mittelwertbereinigung genannt, möglich. Hiermit wird der Suchraum stark eingeschränkt und Ableitungen sollten als Merkmale nicht mehr nötig sein.

Merkmalsvektoren für SVM sind, auch wenn die Merkmale in zeitlicher Abfolge geordnet sind, nunmehr nicht als Zeitserie zu verstehen, sondern als Punkte in einem durch die Beschaffenheit der Merkmale festgelegten Merkmalsraum. Daher ist es für SVM wichtig, für die Merkmalsextraktion eine Feinjustierung der annotierten Zeit-Intervalle durchzuführen, wenn dies zum Beispiel durch Vorwissen über die Charakteristik der Klasse bzw. Geste möglich ist. Besitzen nun zwei annotierte Kopfnicken ein identisches, aber zeitlich um einige Einzelbilder versetztes Aktivitätsmaximum, so würde der DTW-Algorithmus eine große Übereinstimmung hervorbringen, im Merkmalsraum für SVM jedoch weit voneinander entfernt liegen. Umso wichtiger ist für die Vorverarbeitung der Merkmalsvektoren eine möglichst genaue Zentrierung des Aktivitätsmaximums innerhalb des Zeitfensters, wie bereits im vorigen Abschnitt erwähnt.

Die Festlegung der Fensterbreite erfolgt entsprechend einer typischen

Länge der Geste, wobei hier ein Kompromiss nötig ist; Wird das Zeitfenster zu klein gewählt, werden womöglich wichtige Ausprägungen längerer Gesten nie im Zusammenhang erfasst. Wird es zu groß festgelegt, so kommen viele für die Geste unbedeutende Merkmale hinzu, die den Merkmalsraum und somit auch den Suchraum unnötig um ein Vielfaches vergrößern.

6 Evaluation des Detektionssystems für Kopfnicken

Klassifikationssysteme sollen Daten automatisch in eine bestimmte Kategorie einteilen. Dies funktioniert in der realen Welt selten perfekt. Somit ist es wichtig, die Leistungsfähigkeit von Klassifikatoren messen zu können, um so einen bestmöglichen Klassifikator für ein spezifisches Anwendungsfeld zu finden. Im Falle eines binären Klassifikators existieren vier Zustände, die Daten nach einer Klassifikation einnehmen können. Die Daten, die detektiert werden sollen, werden entweder korrekt erkannt (Richtig-positiv), oder sie werden nicht erkannt beziehungsweise vom Klassifikator 'übersehen' (Falsch-negativ). Alle restlichen Daten werden entweder fälschlicherweise vom Klassifikator erfasst (Falsch-positiv) oder korrekterweise ausgeschlossen (Richtig-negativ). Die vier Fälle, die bei einer binären Klassifikation vorkommen können, sind noch einmal in der Abbildung 6.1 skizziert.

Aus den vier Fällen werden verschiedene Evaluations-Maße abgeleitet. So bezieht sich die **Sensitivität** (Richtig-positiv-Rate) auf die Daten mit der gesuchten Klasse und gibt den Anteil an, wie viele Samples dieser Klasse auch tatsächlich als diese erkannt wurden. Der restliche unerkannt gebliebene Anteil dieser Klasse wird durch die **Falsch-Negativ-Rate** ausgedrückt.

Dagegen bezieht sich die **Spezifität** (Richtig-negativ-Rate) auf die Daten, die nicht der gesuchten Klasse entsprechen und gibt den Anteil an, wie viele der nicht gesuchten Samples korrekterweise vom Klassifikator nicht detektiert werden. Der restliche fälschlicherweise als gesuchte Klasse detektierte Anteil wird **Falsch-positiv-Rate** genannt.

Bezieht man sich auf die Daten, die positiv klassifiziert wurden, so misst der **Positive Vorhersagewert** den Anteil unter den positiv klassifizierten Samples, der wirklich auch der gesuchten Klasse entspricht. Dagegen nennt sich der Anteil der negativ klassifizierten Samples von der Menge der tatsächlich nicht gesuchten Samples **Negativer Vorhersagewert**.

Bezogen auf die Korrektheit des Klassifikationsergebnisses wird durch die

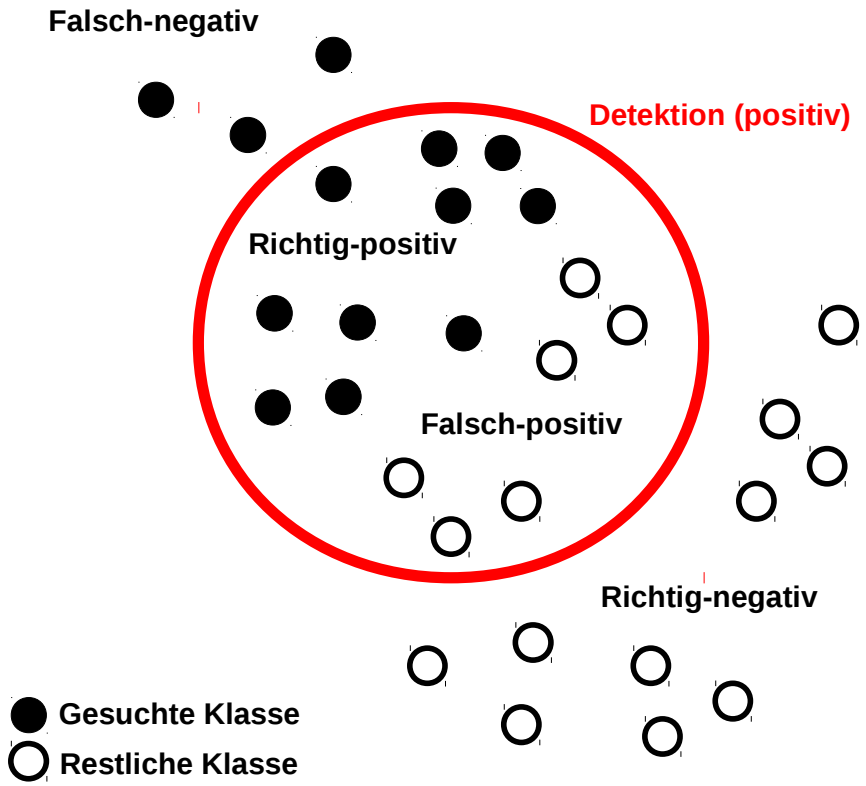


Abbildung 6.1: Die vier möglichen Fälle einer binären Klassifikation. Die schwarzen Punkte stellen Samples der gesuchten Klasse dar, die lediglich umrandeten Punkte sind dagegen alle anderen nicht gesuchten Samples. Der Klassifikator für die gesuchte Klasse bestimmt in dem Merkmalsraum einen Unterraum, innerhalb dessen er alle Samples als die gesuchte Klasse klassifiziert. Der rote Kreis zeigt diesen Unterraum.

Korrektklassifikationsrate der Anteil von allen Samples angegeben, bei dem der Klassifikator eine richtige Entscheidung getroffen hat. Der restliche Anteil beschreibt als **Falschklassifikationsrate** den Anteil unter allen

Samples, bei denen der Klassifikator eine falsche Entscheidung getroffen hat.

Da diese Evaluations-Maße stark korelliert sind, können sie nie komplett unabhängig voneinander optimiert werden. Oft macht es Sinn, je nach Anforderungen des konkreten Anwendungsfalls zwei Gütekriterien auszuwählen, nach denen man die Ergebnisse ausrichtet. Manchmal werden auch mehrere Gütekriterien zu einer gemeinsamen Größe kombiniert. Dazu zählt beispielsweise der sogenannte F-Score. Der F-Score kombiniert die Richtig-positiv-Rate mit dem Positiven Vorhersagewert anhand des harmonischen Mittels zu der Formel:

$F = 2*\text{Sensitivität}*\text{Pos.Vorhersagewert}/(\text{Sensitivität}+\text{Pos.Vorhersagewert})$

Dadurch wird das Gütekriterium mit dem niedrigeren Wert umso stärker gewichtet, je größer die Differenz zu dem Gütekriterium mit dem höheren Wert ausfällt. Somit wird die Balance der Werte der beiden Gütekriterien belohnt und eine starke Abweichung bestraft.

Eine ebenfalls hilfreiche Methode zur Messung der Leistung eines Klassifikators ist die Anfertigung einer sogenannten Receiver-Operating-Characteristic-Kurve (auch ROC-Kurve genannt). Diese dient der Visualisierung der Abhängigkeit zwischen zwei Gütekriterien. Dafür werden beispielsweise die Lern-Parameter eines Klassifikators iterativ verändert, sodass möglichst der gesamte Definitionsbereich zweier ausgewählter Gütekriterien ausgeschöpft wird. Erfolgt die Variation der Parameter in kleinen Schritten, so lässt sich aus den Ergebnissen eine Kurve im zweidimensionalen Raum der beiden Gütekriterien zeichnen. Entlang der Kurve kann man nun entscheiden, welche Kombination man für einen speziellen Anwendungsfall für geeignet hält. Alternativ lässt sich auch der Youden-Index maximieren, der sich mit $J = \text{Sensitivität}+\text{Spezifität}-1$ berechnet lässt, um einen optimalen Schwellwert zu erhalten (Kreienbrock et al. (2012)).

6.1 Frame-basierte und Event-basierte Evaluation

Oft ist es nicht eindeutig, wie die vier grundlegenden Ereignisse Richtig-Positiv, Richtig-Negativ, Falsch-Positiv und Falsch-Negativ in einem speziellen Anwendungsfall definiert werden.

Wesentlich ist dabei die Unterscheidung, ob in der Evaluation eines Datensatzes jedes Einzelbild bewertet wird oder ob lediglich binär abgefragt wird, ob ein Ereignis erkannt wurde oder nicht. Veranschaulicht wird dieser

Unterschied in der Abbildung 6.2. Hier wird ein exemplarischer Ausschnitt mit zwei annotierten Ground Truth Intervallen gezeigt. Die erste Annotation wird von einem Detektionssystem korrekt erkannt, die zweite nicht. Bei der Ereignis- oder Eventbasierten Evaluation wird überprüft, ob sich die Detektion mit der Annotation überlappt. Ist dies der Fall, wird davon ausgegangen, dass das Signal korrekt erkannt wurde und das Ereignis als eine Richtig-Positiv-Bewertung erfasst. In Anwendungsbereichen wie Kopfnicken ist dies besonders sinnvoll, da es oft lediglich darauf ankommt, ob ein Nicken stattgefunden hat oder nicht. Entsprechend wird ein Ground-Truth-Intervall ohne sich überschneidende Detektion als Falsch-Negativ gewertet. Jedoch werden auch die Bereiche zwischen zwei Annotationen jeweils als ein einzelnes Event betrachtet. Herrscht in einem Datensatz eine starke Unausgewogenheit zwischen Positiv und Negativ, so stehen gegebenenfalls sehr kurze positive Ereignisse sehr langen negativen Ereignissen gegenüber und werden gleich stark gewichtet.

Dagegen wird bei der Einzelbild- oder Frame-basierten Evaluation jedes Einzelbild bewertet. Stimmt ein detektiertes Intervall nicht vollständig mit dem annotierten Intervall überein, fließen die nicht überlappenden Einzelbilder als Falsch-Positive bzw. Falsch-Negative mit in die Gesamtwertung mit ein. Ist ein Negativ markierter Bereich doppelt so lang wie ein weiterer, wird der erstere auch doppelt so stark gewichtet, während event-basiert beides gleich behandelt werden würde.

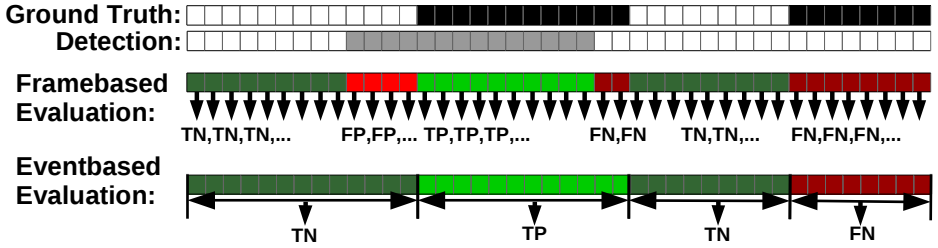


Abbildung 6.2: In einer schematischen Darstellung wird der prinzipielle Unterschied zwischen Einzelbild-basierter (framebased) Evaluation und Event-basierter (eventbased) Evaluation anhand eines Beispieles dargestellt.

Übertragen auf unseren Anwendungsfall des Kopfnickens innerhalb eines

Gesprächskontextes kommt die Event-basierte Variante einem menschlichen Beurteiler recht nahe, da man aus beobachtender Sicht zwar entscheiden kann, ob ein Kopfnicken innerhalb eines kurzen zeitlichen Kontextes stattgefunden hat oder nicht, jedoch Abweichungen von wenigen Einzelbildern nicht als relevant eingestuft werden können. Andererseits liegt in der Regel wesentlich weniger Nicken als Nicht-Nicken vor, wodurch sehr lange Intervalle mit Nicht-Nicken ebenso stark gewichtet werden wie sehr kurze.

6.2 Eingesetzte Evaluationsmethoden in verwandten Arbeiten

Wie in Kapitel 2 aufgeführt, wurden in den letzten Jahren bereits mehrere Systeme zur Detektion von Kopfnicken vorgestellt. Um die im praktischen Einsatz tatsächlich erwartbare Leistungsfähigkeit so eines Systems einschätzen zu können, ist eine aussagekräftige Evaluation essentiell. Dennoch verzichteten bisher leider viele Autoren sogar komplett auf eine Auswertung ihres vorgestellten Systems oder wiesen lediglich darauf hin, dass dieses in der Lage war, Kopfnicken zu erkennen (Zelinsky and Heinzmann (1996),Kawato and Ohya (2000),Davis and Vaks (2001),Lee et al. (2004),Chu and Tanaka (2012)).

Kapoor stellte zur Evaluation einen Datensatz von 62 Samples für Nicken und 48 Samples für Kopfschütteln zusammen (Kapoor and Picard (2001)). Davon trainierte er mit 25 Nicken-Samples und 20 Schütteln-Samples HMMs und separierte die restlichen Samples dieses Datensatzes als Testmenge. Zur Auswertung wurde die Testmenge den trainierten Klassifikatoren übergeben und die Richtig-Positiv-Raten (Sensitivität) bestimmt. Bezüglich Falsch-Positive Erkennungen heißt es lediglich, es seien einige aufgetreten, jedoch keine Zahlen genannt.

Für den von Nguyen vorgestellten Kopfnick-Erkenner mittels SVMs wurde bei Samples zwischen offensichtlichem Nicken, nicht-offensichtlichem Nicken und Nicht-Nicken unterschieden (Nguyen et al. (2012)). Für die Klassen 'offensichtliches Nicken' und 'Nicht-Nicken' wurden jeweils etwa 5000 Samples extrahiert und somit etwa eine gleiche Anzahl an positiver und negativer Samples für das Training verwendet. Um den Sprechstatus als zusätzliches Merkmal zu evaluieren, wurden zwei SVM-Modelle mit dem Merkmal 'sprechend' beziehungsweise 'schweigend' getrennt trainiert. Ein weiteres SVM-

Modell ohne Sprechstatus wurde als Baseline trainiert, um zu prüfen, ob der Sprechstatus als zusätzliches Merkmal hilfreich ist. Während die Evaluation auf Einzelbildern erfolgte (Frame-based Evaluation), kam eine Leave-one-out Kreuzvalidierung auf Sequenzbasis zum Einsatz. Es wurde also beim Training immer eine Videosequenz weggelassen, um darauf anschließend zu testen, bis alle Sequenzen als Test-Sequenz gedient haben. Je nach Schwellwert wurde eine ROC-Kurve Kurve mit den Werten Richtig-Positiv und Falsch-Positiv auf den Achsen erstellt. Zusätzlich wurde der F1-Score angegeben.

Wei definierte die drei Klassen Kopfnicken, Kopfschütteln und Sonstige und stellte dazu jeweils 50 Samples bereit (Wei et al. (2013)). Die Hälfte, die mit 25 Samples je Klasse zufällig ausgewählt wurde, fungierte als Trainingsmenge der HMMs. Die damit trainierten HMMs wurden mit den restlichen Samples getestet und dabei die Trainingset und Testset, lediglich samples verglichen und dabei die Korrektklassifikationsrate (recognition accuracy) bestimmt. Diese erreichte auf den Testdaten einen Wert von 83%.

Bei Terven et al. (2014) bestand der Datensatz aus jeweils 100 Samples für die Klassen Kopfnicken, Kopfschütteln, Links, Rechts, Hoch und Hinunter. Die aufgezeichneten Samples wurden mit Probanden aufgezeichnet, die explizit zu der Ausführung der bestimmten Geste aufgefordert wurden. 70% dieser Daten wurden für das Training von HMMs verwendet. Es wurden ROC-Kurven mit Richtig-Positiv- und mit Falsch-Positiv-Rate erstellt. Dabei wurde eine Erkennungsrate von bis zu 98.5% für Kopfnicken genannt.

Eine der wohl ausgefeiltesten Evaluationen auf dem Gebiet der Kopfnick-Erkennung wurde von Chen publiziert (Chen et al. (2015)). Im Gegensatz zu den meisten anderen Veröffentlichungen testete er nicht mit ähnlicher Anzahl an Sample-Daten pro Klasse, sondern nahm alle extrahierbaren Samples ganzer Interaktionsvideos und somit eine realitätsnahe Verteilung von positiven und negativen Samples. Da der Raum mit negativen Samples (Nicht-Nicken) größer ist, als der Raum mit Nicken, verwendete er für das Training der SVM 3100 positive und 10000 negative Samples aus dem Datensatz 'ubimpressed'. Dieser Datensatz wurde nicht publiziert, lediglich ein kleiner Einblick wird auf der Internetseite https://www.idiap.ch/dataset/ubipose unter der Bezeichnung 'ubipose' gewährt.

Vor und nach annotierten Kopfnicken wurden 7 Frames als Transitions-Frames markiert, die nicht für das Training verwendet wurden, um den Schwierigkeiten von undefinierbaren Zwischenzuständen zwischen Nicken

und Nicht-Nicken zu entgehen. Neben der Einzelbild-basierten Auswertung, beschrieb er zunächst die Evaluationsmaße Precision, Recall und F-Score auch für den Event-basierten Fall. Dazu formulierte er Formeln, die mittels eines wählbaren Schwellenwertes je nach Überlappung der Erkennung mit der Annotation eine Detektion als korrekt oder falsch einstuft. Bei der Klasse 'Kopfnicken' unterschied er wie Nguyen et al. (2012) ebenfalls zwischen offensichtlichen und nicht-offensichtlichen Samples. Zu Trainings-Zwecken wurden nur die Offensichtlichen verwendet. Für jedes Interaktionsvideo des 'ubimpressed'-Datensatzes nahm er als Training eine 'Leave-one-person' Kreuzvalidierung vor und testete die Generalisierungsfähigkeit auf einem anderen Video-Korpus 'KTH-Idiap' Oertel et al. (2014). Beim Training der SVM variierte er zwischen linearem und RBF-Kernel und den Parametern m=1,3,5,7, wobei die besten Ergebnisse bei m=3 und m=5 mit einem F-Score von 0.68 auf Event-Level-Basis lagen.

Ota trainierte ein künstliches Neuronales Netz mit 600 Nick-Sequenzen und ebenfalls 600 Sequenzen mit Nicht-Nicken. Das entwickelte Neuronale Netz ließ er anschließend 100 Nicken und 100 Nicht-Nicken klassifizieren und bestimmte jeweils die Richtig-Positiv-Raten. Zusätzlich wiederholte er die Tests mit Hinzunahme eines Maßes für Sprachrhythmus als weiteres Merkmal und konnte damit seine Ergebnisse leicht verbessern.

Ebenfalls ein Ansatz mittels Neuronalem Netz evaluierte Langholz and Brasher (2018). Er unterschied Kopfnicken, Kopfschütteln und 'Weitere Gesten'. Von dem verwendeten Videomaterial verwendete er 90% für das Training und 10% als Testdaten. Dies resultierte in 172 Nicken im Training und lediglich 15 Nicken im Test-Datensatz. Diese wurden anschließend durch künstliche Variationen auf eine Zahl von 62356 extrapoliert, sodass letztlich mit 62356 Nicken trainiert und mit 5520 Nicken getestet werden konnte. Als Ergebnis wird eine 'accuracy' von knapp 92% angegeben.

6.3 Grundproblematiken verschiedener Evaluationskriterien

Bei der Evaluation eines Klassifikations-Systems zum Erkennen von Kopfnicken gilt es einiges zu beachten. Ziel sollte dabei sein, die tatsächliche Leistungsfähigkeit des Erkenners innerhalb eines Interaktions-Szenarios einschätzen zu können. Je nach Konzeption der Evaluation kann es leicht

vorkommen, dass Detektionssysteme auch mit sehr guten Evaluations-Werten dennoch eine schwache Leistung in der tatsächlichen Anwendung bringen. Deshalb ist es wichtig, die Evaluationsergebnisse mit entsprechenden Evaluations-Maßen zu berechnen, die in der Praxis auch von Belang sind.

Ein Evaluations-Kriterium alleine sagt oft wenig aus. Oft wurden Datensätze zur Evaluation zusammengestellt, die jeweils zur Hälfte aus Nicken und Nicht-Nicken stammen. Dies klingt bei einem binären Klassifikationsproblem zunächst plausibel, jedoch sollte dann bei der Interpretation der Ergebnisse besonders beachtet werden, dass die Klassenverteilung im Anwendungsfall selten bei 50% Nicken liegt. Vielmehr liegt der Anteil in dem bisher betrachteten Videomaterial sogar eher um 5%.

Die Auswirkung soll mit folgendem Beispiel verdeutlicht werden: Angenommen, ein Datensatz mit einer Verteilung von je 50% Nicken und Nicht-Nicken wird von einem Detektionssystem mit einer Richtig-Positiv-Rate von 0.75 bewertet und einer Richtig-Klassifikations-Rate von ebenfalls 0.75. Wird dasselbe Erkennungssystem auf einem Datensatz mit lediglich 5% Nicken getestet, ergibt eine Leistungsfähigkeit von 0.75% auf den verbliebenen 95% Nicht-Nicken plötzlich sehr viele Falsch-Positive im Verhältnis zu den wenigen Richtig-Positiven und schneidet deutlich schlechter ab.

Ein weiteres Beispiel von nicht praxistauglicher Evaluation ist der direkte Vergleich von Kopfnicken mit einer bestimmten anderen Geste. So lässt sich verhältnismäßig leicht Kopfschütteln von Kopfnicken unterscheiden (Gopakumar and Suni (2016)), dazu muss jedoch für die zu untersuchende Sequenz bekannt sein, dass diese entweder Kopfschütteln oder Nicken zeigt, worüber ein Erkennungssystem im praktischen Einsatz üblicherweise keinerlei Information besitzt.

Seien als Beispiel zwei Videoszenen von ähnlicher Dauer mit Frontalaufnahmen einer sich im Gespräch befindenden Person gegeben. Auf dem ersten Video 'A' nickt diese Person über den gesamten Zeitverlauf nur zwei Mal. Auf dem zweiten Video 'B' nickt sie sehr häufig. Ein Detektionssystem erkennt auf A beide Nicken und verzeichnet in dem langen Zeitraum ohne Nicken einige Fehldetektionen. Dasselbe System erkennt auf B ebenfalls die meisten Nicken und führt auch im restlichen Bereich zu einigen Fehldetektionen. Die beiden konstruierten Beispiele werden in der Abbildung 6.3 dargestellt und mit Event-basierten Evaluationsmaßen versehen.

Offenbar erkennt dieses Detektionssystem Nicken sehr zuverlässig, detek-

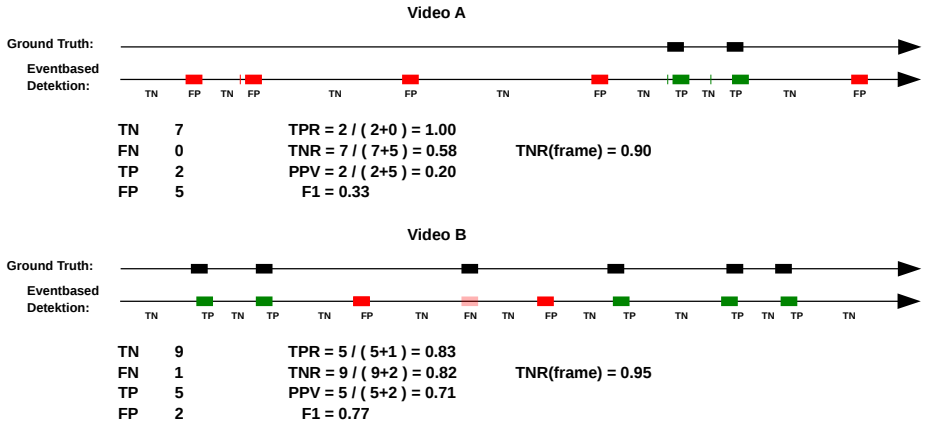


Abbildung 6.3: Dieses Beispiel soll die Auswirkung von stark unterschiedlichen Mengenverhältnissen zwischen zwei Klassen auf die Evaluation verdeutlichen. Dazu werden zwei Videos A und B miteinander verglichen. Der obere Zeitstrahl zeigt annotiertes Kopfnicken (schwarze Markierungen) und der direkt darunterliegende die Ergebnisse eines Detektionssystems (rote Markierungen entsprechen einem FP oder FN, grüne Markierungen einem TP).

tiert jedoch auch oft fälschlicherweise Nicken, obwohl keines vorliegt. Wird in einem Testbeispiel viel genickt (Video B), so kann das System seine Stärke ausspielen und erhält gute Ergebnisse. Wird jedoch selten genickt (Video A), so überwiegt die Eigenschaft der häufigen Falsch-Positive und die Ergebnisse fallen schlechter aus.

Der entscheidende Punkt dabei ist, dass bei langen Negativ-Annotationen (Zeitintervalle ohne Nicken) die Wahrscheinlichkeit für eine Fehl-Detektion größer ist, als bei kurzen Negativ-Annotationen. Die Länge der Zeitintervalle wird jedoch bei Event-basierter Evaluation nicht erfasst. Daher ist in diesem Fall die Frame-basierte Richtig-Negativ-Rate vorteilhafter, da diese mit Anzahl der betroffenen Frames auch die Länge der Zeitintervalle mit berücksichtigt.

Andererseits ist bei der Richtig-Positiv-Rate in unserer Anwendung die Event-Basis vorzuziehen, wie bereits im Abschnitt 6.1 erläutert.

Somit wird die Kombination von Frame-basierter Richtig-Negativ-Rate mit Event-basierter Richtig-Positiv-Rate als gemeinsam zu betrachtende Maßeinheiten als Evaluationskriterium vorgeschlagen, wie es bereits in Wall et al. (2017) vorgenommen wurde. Zum Anderen wird aus der Abbildung 6.3 ersichtlich, dass sich der Präzisions-Wert (PPV) am deutlichsten unterscheidet. Möchte man somit besonderen Wert darauf legen, dass ein Klassifikations-System auch in längeren Zeit-Intervallen ohne Kopfnicken gute Ergebnisse liefert, so ist dieser Wert besonders aussagekräftig.

6.4 Training und Testverfahren für DTW

Wie im Unterkapitel 5.6 beschrieben, wird eine prototypische Zeitserie benötigt, die eine Ähnlichkeit zu möglichst vielen Kopfnick-Zeitserien aufweist. Hierfür wird ein mögliches Vorgehen zur Prototyp-Suche innerhalb eines Trainings-Datensatzes beschrieben, bevor ein Verfahren zur Kreuzvalidierung erläutert wird.

6.4.1 Training in Form von Prototyp-Suche

Soll anhand einer Trainings-Datenmenge von annotierten Zeitserien diejenige gefunden werden, die sich innerhalb dieser Menge am meisten als Prototyp eignet, muss zunächst ein Qualitätskriterium definiert werden. Als Qualitätskriterium soll die Klassifikationsleistung dienen. Das heißt, dass innerhalb eines Datensatzes von allen darin enthaltenen potentiellen Prototypen derjenige gesucht wird, der als Prototyp innerhalb dieser Datenmenge die beste Klassifikationsleistung erzielt.

Gemäß der Grafik 6.4 wird wie folgt verfahren:

Jeder annotierte potentielle Prototyp wird auf dem Trainings-Datensatz getestet und beispielsweise die Richtig-Positiv-, sowie die Richtig-Negativ-Rate gemessen. Dabei wird der Klassifikations-Schwellwert für die DTW-Distanz iterativ abgesenkt, bis ein festgelegter Anteil der annotierten Kopfnicken erkannt wird (Richtig-Positiv-Rate). In unserem Fall wurde der Wert 0.67 gewählt, um sicherzugehen, dass eine deutliche Mehrheit an Kopfnicken erkannt wird. Als Gütekriterium gilt schließlich ein gewünschtes Evaluations-Maß wie der zugehörige Präzisions-Wert (PPV), der sich bei der

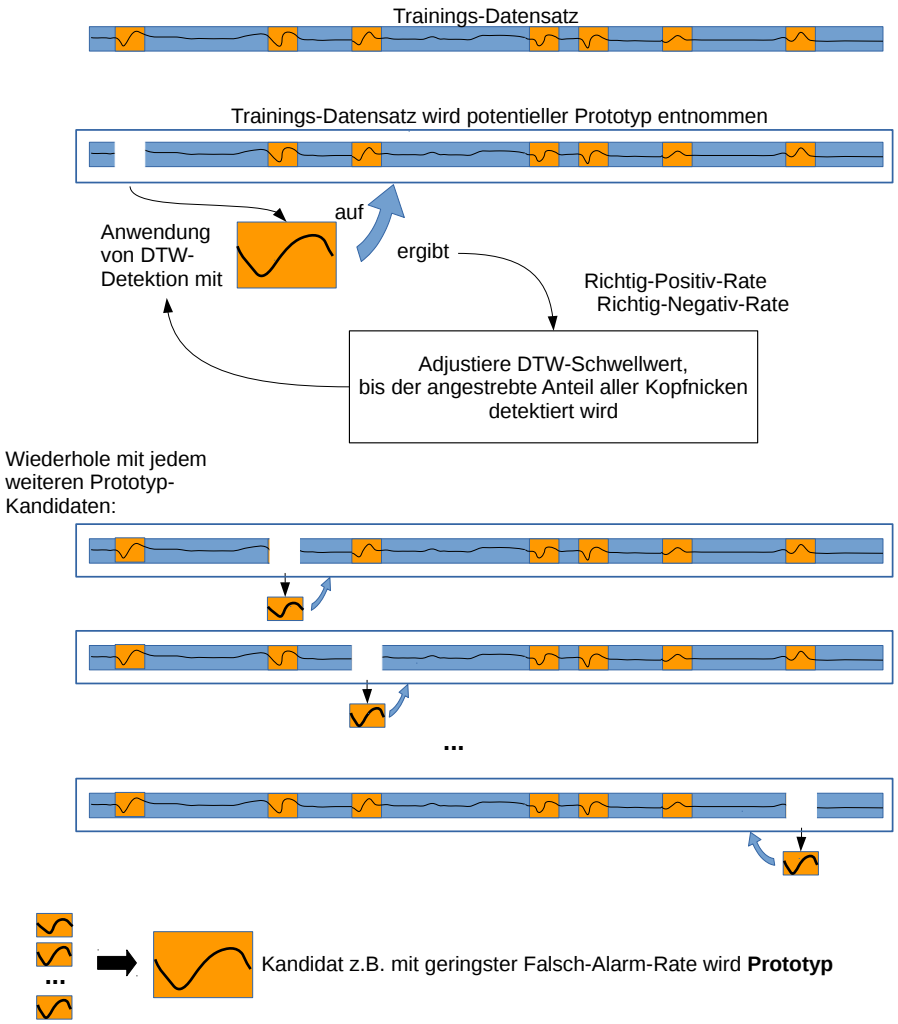


Abbildung 6.4: Hier wird das Auswahlverfahren zur Findung eines geeigneten Prototypen des DTW-Verfahrens aufgezeigt. Dies wird als Trainings-Schritt bezeichnet. Der so ausgewählte Prototyp ist somit auf den Trainingsdaten spezialisiert. Um die Generalisierungsfähigkeit zu testen, wird ein separater Test-Datensatz benötigt, wie im Abschnitt 6.4.2 beschrieben.

festgelegten Richtig-Positiv-Rate ergibt. Unter der Annahme, dass etwa zwei Drittel der Kopfnicken erkannt werden, bewertet der Präzisions-Wert also, wie viele Fehldetektionen (Falsch-Positive) man in Kauf nehmen muss, um diese hohe Richtig-Positiv-Rate zu erreichen. Ebenso kann als Gütekriterium derjenige Prototyp mit der höchsten Richtig-Negativ-Rate oder äquivalent der niedrigsten Falsch-Alarm-Rate herangezogen werden. Der Prototyp mit dem höchsten Gütekriterium kann nun plausibel als der leistungsstärkste Prototyp auf den Trainingsdaten gehandelt werden.

6.4.2 Kreuzvalidierung

Um zu testen, wie gut der im Trainings-Verfahren bestimmte Prototyp auf Daten generalisiert, die nicht im Trainings-Datensatz enthalten sind, muss ein Test-Datensatz vor dem Training ausgeschlossen werden. Die mit dem Prototyp auf dem Test-Datensatz gemessenen Werte zeigen ein realistischeres Ergebnis, wie gut der trainierte Prototyp generalisiert.

Um mit den vorhandenen Daten möglichst viele aussagekräftigen Tests durchzuführen, wird eine Kreuzvalidierung wie folgt durchgeführt: Zunächst wird der Datensatz so aufgeteilt, dass eine oder mehrere Personen als Test-Datensatz ausgeschlossen werden und die übrige Datenmenge jeweils als Trainings-Datensatz fungiert.

Die gemessenen Qualitätskriterien auf den Test-Datensätzen werden anschließend gemittelt. Die Ergebnisse folgen im Abschnitt 6.6 gemeinsam im Vergleich zu den Ergebnissen mit SVM.

6.5 Training für SVM

Im Folgenden wird nun das Training- und Testverfahren der SVM-Variante beschrieben. Zuerst müssen auf geeignete Art und Weise Trainingsvektoren ausgewählt werden. Anschließend folgt die Umsetzung einer Kreuzvalidation auf das SVM-Verfahren und die Bestimmung geeigneter Evaluationskriterien.

6.5.1 Auswahl der Trainingsvektoren

Da bei natürlichen Interaktionen eine stark unausgeglichene Klassenverteilung bezüglich Nicken und Nicht-Nicken zu erwarten ist, sollten die Trai-

ningsdaten besonders sorgsam ausgewählt werden.

Für das Training einer SVM zur Klassifikation von Kopfnicken wird eine Menge an positiven Trainingsdaten (Kopfnicken), sowie eine Menge an negativen Trainingsdaten (Nicht-Kopfnicken) benötigt.

Der positive Anteil sollte alle annotierten Kopfnicken der potentiellen Trainingsmenge beinhalten. Gegebenenfalls ist auch ein Ausschluss von Grenzfällen im Sinne von schwach ausgeprägten Samples oder von Samples mit uneinheitlicher Annotation verschiedener Annotatoren sinnvoll.

Für die Zusammenstellung des negativen Anteils wäre ein einfacher Ansatz, die Anzahl entweder gleichauf mit dem positiven Anteil oder gemäß des Klassenverhältnisses zu wählen.

Allerdings sollte beachtet werden, dass die negativen Samples zu einem sehr großen Anteil aus Interaktions-Szenen stammen, in denen keine erkennbaren Gesten und somit meist auch kaum Kopfbewegungen stattfinden.

Würde man nun die negativen Trainings-Samples zufällig aus den nichtpositiv annotierten Videoszenen ziehen, so bestünde die Gefahr, dass die negativen Samples im Extremfall ausschließlich aus Szenen bestehen, in denen keine wahrnehmbaren Kopfbewegungen stattfinden.

Ziel sollte jedoch sein, die negative Trainingsmenge so zu wählen, dass sie den Raum der Merkmalsvektoren repräsentiert, der nicht als Nicken interpretiert werden soll. Wird jedoch nur ein sehr kleiner konzentrierter Teilbereich aus dem Merkmalsraum, der üblicherweise durch Kopfbewegungen aufgespannt werden kann, als negative Trainingsmenge gewählt, so besteht die Gefahr, dass eine trainierte SVM nicht zwischen Nicken und Nicht-Nicken zu unterscheiden lernt, sondern lediglich zwischen 'keine Kopfbewegung' und 'stärkere Kopfbewegung'.

Um den negativen Raum mit einer Anzahl an negativen Samples besser zu repräsentieren, sollten in der negativen Trainingsmenge auch viele Samples enthalten sein, die stärkere Kopfbewegungen abseits von Kopfnicken zeigen.

Um dies zu erreichen, schlage ich folgendes Auswahlverfahren vor:

Die Standardabweichung aller vertikalen Kopfwinkel innerhalb eines Samples dient als Maßeinheit der Bewegungsstärke.

Es wird ein Histogramm über einen Umfang von Standardabweichungs-Bereichen konzipiert.

Jedes Bin des Histogramms hat eine maximale Füllmenge, sodass die Gesamtzahl an Werten der Gesamtmenge der gesuchten negativen Trainingsmenge entspricht.

Für jedes zufällig gezogene negative Sample wird die Standardabweichung berechnet.

Hat die 'Füllmenge' des entsprechenden Bins des Histogramms noch nicht das Maximum erreicht, wird dieser Bin-Wert inkrementiert und das extrahierte Sample in die negative Trainingsmenge aufgenommen.

Schließlich erhält man so eine Menge von negativen Samples, die die gewünschte Bandbreite an Standardabweichungen enthält und somit den Merkmalsraum der negativen Samples weiträumig repräsentiert.

Als Verhältnis zwischen positiven und negativen Samples erwies sich nach einigen Tests eine leichte Tendenz zur größeren negativen Menge von etwa 60% als sinnvoll.

6.5.2 Kreuzvalidierung mit SVM

Initial wird von einem Gesamt-Datensatz ausgegangen, auf dem die Leistungsfähigkeit der SVM-Klassifikation anhand einer Kreuzvalidierung getestet werden soll. Dabei wird aus dem verfügbaren Datensatz so oft eine oder mehrere Versuchspersonen als Testdatensatz zurückgehalten und mit den verbliebenen (Trainings-)Daten trainiert, bis alle Versuchspersonen einmal als Testdatensatz fungiert haben. Beim Training eines Trainingsdatensatzes wird bezüglich der Auswahl der Trainingsvektoren gemäß des zuvor beschriebenen Abschnitts verfahren. Die auf den Trainingsdaten berechneten Klassifikator-Funktionen werden auf den jeweiligen Testdaten angewendet und die Ergebnisse anschließend zusammengefasst.

Da das Endergebnis stark von den gewählten Parametern der SVM abhängt, wird die gesamte Trainingsprozedur gemäß einer Rastersuche (Grid Search) mit verschiedenen Parameter-Kombinationen ausgeführt, um diejenige Kombination mit dem besten Endergebnis zu finden.

Da in der Regel positive Vektoren mit '+1' und negative Vektoren mit '-1' gelabelt werden, wird die Klassifikations-Schwelle oft bei '0' angenommen. Jedoch kann es aufgrund stark unausgeglichener Klassenverteilung sinnvoll sein, auch andere Schwellwerte zu explorieren, um gewünschte Klassifikationsraten zu erhalten.

Für die Anwendung des Klassifikators für weitere Datensätze würde man nun diejenigen Parameter mit den besten Ergebnissen bei der Kreuzvalidierung heranziehen und damit noch einmal mit dem gesamten Datensatz trainieren, da eine Einschränkung zugunsten von Testdatensätzen nicht mehr

nötig ist. Jedoch ist der Effekt des erneuten Trainingsschritts eher gering, da aufgrund der kleinen Testdatensätze nur eine Person hinzukommt.

6.6 Ergebnisse von DTW und SVM

Als Evaluations-Grundlage wurde der im Kapitel 4 erwähnte WOZ1 Datensatz verwendet. Es wurden Videos ausgeschlossen, bei denen auf weniger als 90% der Frames kein Gesicht erkannt wurde und somit viele Merkmale fehlen. Ebenso ausgeschlossen wurden Videos, in denen fast gar nicht genickt wurde. Somit erstreckte sich das Videomaterial auf etwa 20 Interaktionen mit einer Gesamtlänge von etwa vier Stunden. In der Grafik 6.5 werden die Evaluationsergebnisse von DTW und SVM des WOZ1 Datensatzes aus Kapitel 4 anhand von ROC-Kurven dargestellt. Auf der horizontalen Achse ist die Event-basierte Richtig-Positiv-Rate aufgetragen und somit der Anteil, der von den annotierten Kopfnicken vom System erkannt wird. Die vertikale Achse beschreibt zwei verschiedene Maße; Anhand der durchgehenden Linie wird der Event-basierte Positive Vorhersagewert (PPV) gezeigt, der den Anteil beschreibt, wie viele der ausgelösten Detektionen tatsächlich als Kopfnicken annotiert sind.

Dass die durchgehenden Linien recht nah an der Diagonalen liegen, heißt somit keineswegs, dass die Ergebnisse nur unwesentlich besser als zufällig sind. Vielmehr heißt es zum Beispiel für die SVM Folgendes: Der überwiegende Teil der Kopfnicken werden erkannt, während gleichzeitig auch die meisten Detektionen keine Fehldetektionen waren, obwohl etwa 95% der Daten kein Kopfnicken und somit potentielles Material für Fehldetektionen darstellt.

Außerdem zeigt die gestrichelte Linie die Frame-basierte Richtig-Negativ-Rate und somit den Anteil aus dem Bereich ohne Nicken, der auch korrekterweise keine Detektion ausgelöst hat.

Wie aus der Grafik ersichtlich, wurde mit dem Ansatz mittels SVM besonders im mittleren Bereich bei gleicher Richtig-Positiv-Rate (TPR) wie bei DTW eine über 10% höhere Korrekt-Klassifikationsrate (PPV) erzielt. Somit fällt der Unterschied zwar nicht gravierend, aber doch klar zugungsten des Ansatzes mit SVM aus.

Um die Leistungsfähigkeit besser einordnen zu können, ist neben der Gegenüberstellung der entwickelten Verfahren auch ein Vergleich mit einer

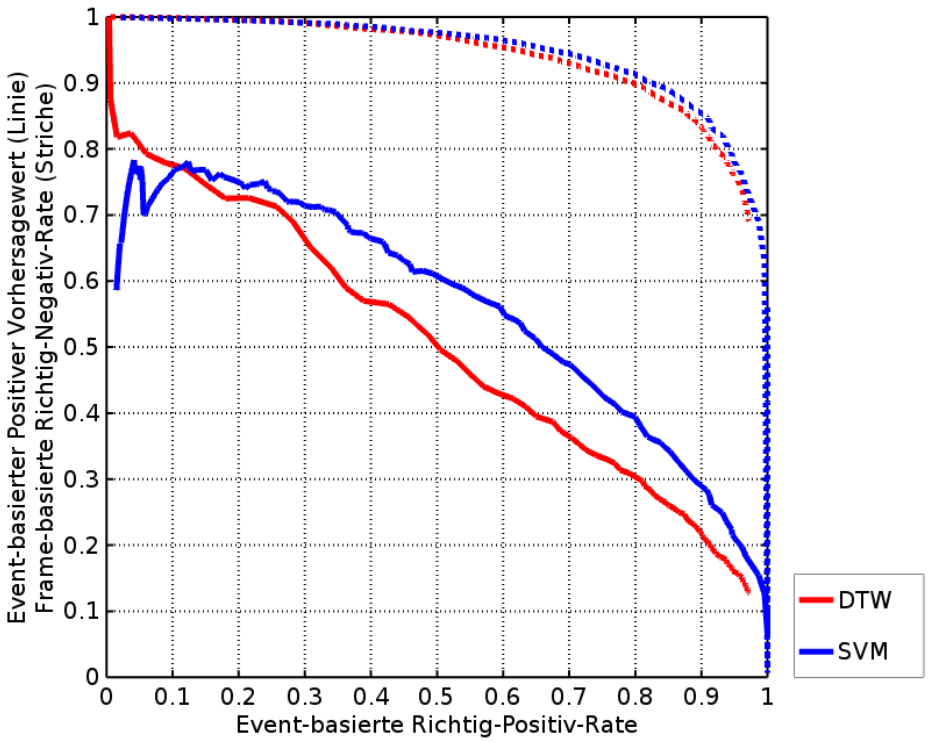


Abbildung 6.5: Hier werden die ROC-Kurven zu DTW und SVM anhand des WOZ1 Datensatzes gegenübergestellt.

State-Of-The-Art-Methode sinnvoll. Im Kapitel 2.6 wurde das Detektionssystem von Chen et al. (2015) als eines der führenden Arbeiten in dem Bereich der Kopfnick-Detektion beschrieben. Dessen Evaluationsansatz mit entsprechenden Testdaten wurde im Abschnitt 6.2 erläutert. Der Datensatz wurde für einen direkten Vergleich nicht herausgegeben. Um dennoch einen Vergleich zu ermöglichen, wurde darauf geachtet, eine Teilmenge des WOZ1 Datensatzes so zu entnehmen, dass das Klassenverhältnis zwischen Nicken und Nicht-Nicken analog zu Wall et al. (2017) in etwa dem der Testdaten von Chen et al. (2015) entspricht. Die bestimmten Evaluationsmaße sind in der Tabelle 6.1 beschrieben. Wenn sich der Positive Vorhersagewert (PPV) bei allen drei Verfahren auf einem einheitlichen Niveau

Tabelle 6.1: Ergebnisse von DTW und SVM im Vergleich zu anderen Methoden bei fixiertem Positiven Vorhersagewert.

Datensatz	Positiver Vorhersagewert (PPV) (event)	Richtig-Positiv-Rate (TPR) (event)	F-Score (event)	Richtig-Negativ-Rate (frame)	#FP /min
WOZ1-DTW	0.60	0.43	0.50	0.97	3.7
WOZ1-SVM	0.60	0.56	0.58	0.97	2.2
Chen-KTH	0.60	0.79	0.68	0.98	1.6

bewegt, so verzeichnet Chen für die Richtig-Positiv-Rate (TPR) und dem F-Score im Vergleich bessere Werte. Immerhin kann zumindest das SVM-Verfahren noch bei der Häufigkeit von Falsch-Positiven (FP) und bei der Richtig-Negativ-Rate (TNR) mithalten. Vor allem da in dieser Arbeit die Merkmale lediglich auf 2D-Daten berechnet wurden, in der Vergleichsarbeit aber ein 3D-Tracking-System zugrunde lag, ist dieses Ergebnis dennoch als sehr angemessen zu betrachten.

6.7 Zusammenfassung der Erkennungsergebnisse

Der auf Dynamic Time Warping basierende Ansatz lieferte beim Vergleichstest des Datensatzes WOZ1 niedrigere Leistungswerte als der Ansatz mittels SVM. Ein Grund könnte in dem Informationsverlust bei der notwendigen Differentation der Merkmale für DTW liegen. Nichtsdestotrotz sollte DTW als Ansatz nicht prinzipiell als schlechter für diese Anwendung angesehen werden. Denn in der Erfassung grundlegender Information wie die Größe des Zeitintervalls von betrachteten Zeitserien und die Berücksichtigung von zeitlicher Reihenfolge der Merkmale bietet DTW Vorteile gegenüber SVM, die beispielsweise durch eine Ausweitung der Prototyp-Strategie stärker ausgespielt werden könnten. Dennoch weist die Nähe der beiden Ergebnisse darauf hin, dass der Flaschenhals weniger das gewählte Verfahren, sondern vielmehr die Grundlagen wie die Merkmalsextraktion durch den Kopfwinkelschätzer und die menschliche Ungenauigkeit im Annotations-Prozess darstellen.

Ein Vergleich mit Ergebnissen anderer Forschungsarbeiten erwies sich als schwierig. Denn wie im Abschnitt 6.2 beschrieben, wurden Rahmenbedingungen einer natürlichen Interaktion meistens nicht eingehalten oder ausführliche Evaluationen nicht durchgeführt.

Im direkten Vergleich mit einem State-Of-The-Art-Verfahren liegen die

Ergebnisse leicht dahinter. Dazu sei allerdings erwähnt, dass das vergleichende Verfahren im Gegensatz zu dieser Arbeit 3D-Information mit spezieller Hardware erfasste (Chen et al. (2015)).

Untersucht man falsch klassifizierte Detektionen, so stellt sich heraus, dass Falsch-Positive vor allem Artefakte aus mangelhaftem Gesichtsalignment sind, die beispielsweise von Brillenträgern verursacht sind. Als eines der größten Schwachpunkte erwies sich, dass stärkere Mundbewegungen beim Sprechen vom Kopfwinkelschätzer oft auch als vertikale Kopfbewegung interpretiert wurden. So führte ein deutliches 'Ja' ohne Kopfbewegung oft zu einer Nick-Detektion, da mit den um den Mund lokalisierten Marker eine 'Runter-Hoch'-Bewegung erfasst wurde. Vielleicht wäre es sinnvoller, in dem Gesichtsmodell die Punkte der unteren Lippe zu entfernen, um durch Sprechen hervorgerufene Artefakte zu vermeiden.

Insgesamt sind die erreichten Ergebnisse jedoch vor allem angesichts der stark unausgeglichenen Klassenverteilung als positiv zu bewerten. Aufgrund der leicht besseren Messwerte für SVM, wird in dem Kapitel 7 auf das SVM-Verfahren gesetzt.

7 Ansätze zur automatischen Interpretation

Im vorigen Kapitel wurde gezeigt, dass Kopfnicken recht zuverlässig erkannt werden kann. Dabei ist zu erwarten, dass die Zuverlässigkeit derartiger Klassifikationssysteme in den nächsten Jahren weiter zunehmen wird. Dafür spricht unter anderem, dass Kamera-Sensoren immer kleiner und hochauflösender wurden, verfügbare Grafikleistung anstieg und somit immer mehr Bildinformationen in kürzerer Zeit als Merkmale verarbeiten werden können. In dem Zuge stellt sich mit steigender Praxistauglichkeit die Frage, wie eine automatische Detektion von Kopfnicken gewinnbringend in der Mensch-Maschine-Interaktion und insbesondere -Kommunikation eingesetzt werden kann.

Im Kapitel 3.2 wurden Arbeiten vorgestellt, die Kopfnicken als kommunikatives Signal beleuchten, Interpretationsansätze liefern und Funktionsklassen ausgearbeitet haben.

In diesem Kapitel werden nun Ansätze herausgegriffen, um einige dieser Erkenntnisse bezüglich des Kopfnickens aus dem zwischenmenschlichen Bereich in die Mensch-Maschine-Interaktion zu übertragen und hier Anwendungsmöglichkeiten zu untersuchen.

So werden in den folgenden beiden Abschnitten zunächst zwei Interpretations-Modelle näher beleuchtet, da diese beiden insbesondere physikalische Unterschiede zwischen Kategorien beschreiben. Zuerst wird das Kopfnickverhalten eines Zuhörers erörtert, bevor versucht wird, die Commitment-Stärke aus der physikalischen Ausführung der Kopfnick-Geste zu extrahieren.

Schließlich wird ein Ausblick über weitere vielversprechende Ansätze gegeben.

7.1 Klassifikation von Zuhörer-Nicken nach Hadar

Im Kapitel 3.2 wurde beschrieben, wie Hadar (Hadar et al. (1985)) das Kopfnickverhalten als Feedback von Zuhörern im Gespräch beobachtete und dabei verschiedene funktionelle Kategorien ausmachte, die auch physikalische Unterschiede aufweisen. In diesem Abschnitt soll nun geprüft werden, inwiefern diese Kategorien zunächst in einem Videokorpus mit Dialog-Kontext wiederzufinden sind und anschließend auch als extrahierte Merkmalsvektoren in einem parametrisierten Merkmalsraum voneinander abgrenzbar sind.

Zuerst werden die Kategorien noch einmal kurz vorgestellt und deren Funktion sowie physikalische Eigenschaften beschrieben.

Anschließend werden diese in einem Videokorpus nachvollzogen und mithilfe Hadars kategorischen Beschreibungen annotiert.

Aus dem annotierten Datensatz werden anschließend Kopfwinkel-Merkmale extrahiert und die Kategorien auf Basis dieser Merkmale analysiert.

Durch das Training einer SVM je Kategorie werden die extrahierten Merkmale abschließend auf die Möglichkeit automatischer Interpretation hin getestet und die Ergebnisse diskutiert.

7.1.1 Die drei Kategorien 'Bestätigung','Antizipation' und 'Synchronisation'

Hadar unterschied bei einem beobachteten Zuhörer-Kopfnicken im Wesentlichen zwischen 'Ja-Nicken' als Bestätigung, Antizipations- und Synchronisations-Nicken. Durch das Erwähnen einer weiteren Kategorie 'Sonstige' wird deutlich, dass mit den drei genannten Kategorien nicht beansprucht wird, ein vollständiges Modell abzubilden. In diesem Kontext soll auf die Kategorie 'Sonstige' jedoch nicht weiter eingegangen werden.

Im Folgenden werden die physikalischen Eigenschaften sowie die Funktion der beobachteten Kategorien für Kopfnicken noch einmal kurz zusammengefasst:

Bestätigung ('hadar-yes') folgt als Antwort auf eine Frage des Sprechers oder signalisiert eine klare Zustimmung zu einer Aussage. Die physikalische Ausführung ist von eher größeren Amplituden der Kopfwinkel-Trajektorie und häufig von mehreren abklingenden Nick-Zyklen gekennzeichnet.

Antizipation ('hadar-anticipation') wird durch eine vertikale Kopfbewegung ausgedrückt, die unter anderem eine Sprachäußerung ankündigen soll. Oft zeigt sich dabei nicht ein kompletter Nick-Zyklus, sondern lediglich eine Auf- oder Abbewegung.

Synchronisation ('hadar-synchronisation') bezeichnet eine Nickbewegung, die durch den Sprachrhythmus des Sprechers hervorgerufen wird. Entweder bei Unstetigkeit des Redeflusses oder begleitend zu betonten Silben des Sprechers. Kopfnicken dieser Klasse sind von kurzer Ausführungsdauer und eher schwächeren Amplituden geprägt.

7.1.2 Datensatz und Annotation

Die vertikalen Kopfbewegungen als entscheidende Merkmale wurden in der Studie von Hadar durch einen speziellen Goniometer durch polarisierendes Licht aufgezeichnet, weshalb auch keine Videoaufzeichnungen mit entsprechenden Annotationen verfügbar sind. Die Daten der ursprünglichen Studie wurden von drei weiblichen und zwei männlichen Versuchspersonen erhoben. Alle waren zwischen 20 und 30 Jahre alt und Studierende. Diese unterhielten sich zur Zeit der Aufzeichnung jeweils für etwa 5 bis 15 Minuten mit einem Interviewer. Insgesamt kamen 40 Exemplare der Kategorie 'hadar-yes' zusammen, 23 von 'hadar-anticipation' und 15 von 'hadar-anticipation'.

Da wir für die Extraktion von Merkmalen auf Videomaterial angewiesen sind, muss ein alternativer Datensatz für unseren Zweck verwendet werden. Hierfür wurden die bereits eingeführten Datensätze WOZ1 und EVAL1 nach Exemplaren der beschriebenen Kategorien durchsucht. Dabei war insbesondere die Kategorie 'hadar-synchronisation' selten anzutreffen. Hier ließe sich vermuten, dass eine maschinelle Stimme mit nur menschenähnlichen Verhaltensmustern weniger Synchronisationsverhalten beim menschlichen Gesprächspartner triggert als in einem rein menschlichen Dialog. Dennoch ließen sich einige Dutzend Beispiele finden. Dabei wurden 34 Exemplare 'hadar-anticipation' zugeordnet, 22 der Kategorie 'hadar-synchronisation' und 35 zu 'hadar-yes'. Insgesamt entstammen die Daten von 14 verschiedenen Studentinnen und Studenten als auch Seniorinnen und Senioren.

7.1.3 Merkmale und Datenanalyse

Als Merkmale wurden Kopfwinkel mithilfe unseres Systems zur Kopfwinkelschätzung extrahiert. Aufgrund der relativ geringen Datenmenge sollte der Merkmalsraum nicht zu groß gewählt werden. Deshalb werden lediglich die vertikalen Kopfwinkel betrachtet und unabhängig von der Länge des annotierten Zeitintervalls eine feste Fensterbreite von 23 Frames pro Kopfnicken extrahiert, was etwas weniger als die Dauer eines durchschnittlichen Kopfnickens ist. Ein Merkmalsvektor für ein Kopfnicken besteht also aus den zeitlich sortierten 23 geschätzten Kopfwinkeln. Die absoluten Werte wurden innerhalb des extrahierten Zeitfensters mittelwertbereinigt. War ein zyklischer Verlauf erkennbar, wurde das Zeitintervall zur Korrektur gegebenenfalls um einige Frames verschoben, sodass sich der lokale Extremwert etwa mittig des Zeitfensters befindet.

Mittelt man nun die einzelnen Kopfwinkel jedes Merkmalsvektors einer Kategorie, erhält man die in der Abbildung 7.1 gezeigten Funktionen.

Folgende bereits von Hadar beobachteten Eigenschaften lassen sich hier wiedererkennen: Ein Bestätigungs-Nicken hat meist größere Amplituden als Synchronisations-Nicken und Antizipations-Nicken besteht oft lediglich aus nur einer linearen Kopfbewegung.

Um einen ersten Eindruck in die Struktur der Daten und dessen Verflechtung im Merkmalsraum zu erhalten, eignet sich der Einsatz einer Hauptkomponentenanalyse (Pearson (1901)). Die Hauptkomponentenanalyse, auch PCA (Principal Component Analysis) genannt, ist ein statistisches Verfahren, welches einen n dimensionalen Datensatz in einen m dimensionalen Raum mit $m <= n$ so transformiert, dass die neu erzeugten Dimensionen (Hauptkomponenten) in Richtung der größten Varianz gelegt werden und somit der Aussagekraft nach sortiert sind. Damit ist eine beliebige Dimensionsreduktion mit minimalem Informationsverlust möglich. Oft beinhalten bereits die ersten beiden Hauptkomponenten einen großen Teil der in dem Datensatz enthaltenen Information und lassen sich im zweidimensionalen Raum gut visualisieren. Die Darstellung der ersten beiden Hauptkomponenten findet sich in der Abbildung 7.2.

Obgleich sich in der Mitte der Darstellung ein Ballungsraum mit verschiedenen Kategorien befindet, stützt sie dennoch die Annahme, dass Antizipation sich besonders gut von den anderen beiden Kategorien abgrenzen lässt. Auch weite Teile der Bestätigung sind gut abgrenzbar. Lediglich die

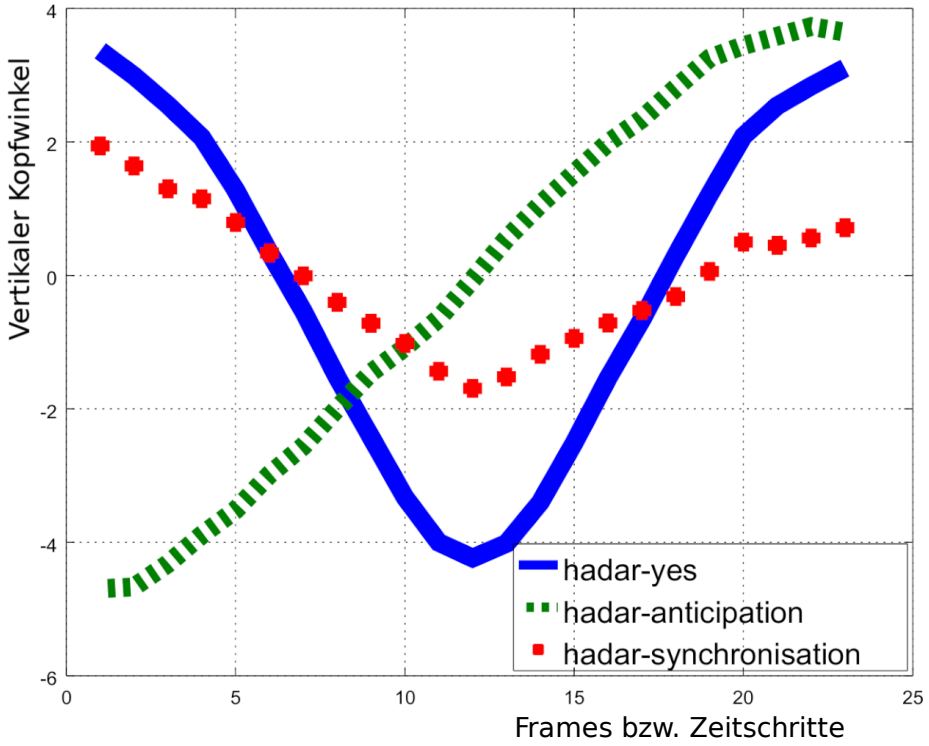


Abbildung 7.1: Diese Abbildung zeigt die gemittelten einzelnen Kopfwinkel jedes Merkmalsvektors der jeweiligen Kategorien 'hadar-yes', 'hadar-anticipation' und 'hadar-synchronisation' je Zeitschritt.

Kategorie 'hadar-synchronisation' scheint sich als eine Art Teilmenge der Kategorie 'hadar-yes' zu erstrecken.

7.1.4 Klassifikation und Ergebnisse

Inwiefern diese dennoch abgrenzbar sind, wird nun in diesem Abschnitt anhand des Trainings und Tests einer Klassifikationsfunktion getestet. Für jede Kategorie soll ein eigener Klassifikator trainiert werden, der sie von den

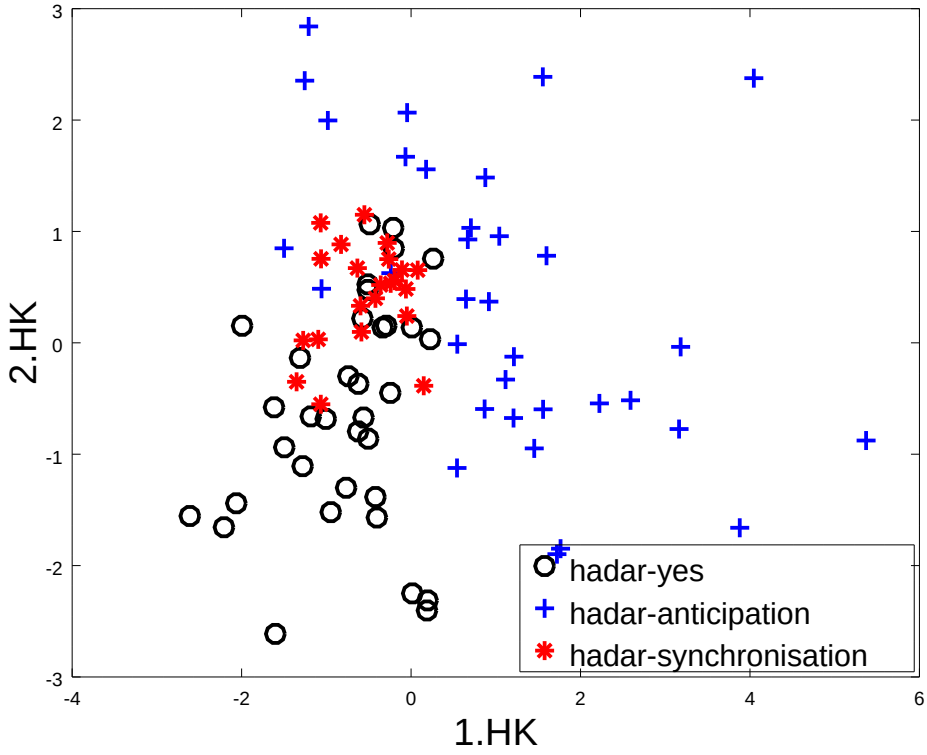


Abbildung 7.2: Diese Abbildung zeigt die Lokalisierung der Merkmalsvektoren aller Datenpunkte im anhand der ersten beiden Hauptkomponenten einer Hauptkomponentenanalyse (PCA). Trotz der Verflechtung aller drei Klassen mittig der Darstellung, lassen sich größere Bereiche insbesondere der Klassen 'hadar-yes' und 'hadar-synchronisation' erkennen.

jeweils restlichen Kategorien abgrenzen soll.

Die Ergebnisse sind in der Tabelle 7.1 wiedergegeben und sind mithilfe eines Grid-Search-Verfahrens und einer 3-fachen Kreuzvalidierung entstanden. Die angegeben Parameter 'gamma' und 'nu' sind dabei die Parameter des SVM-Trainings.

Wie bereits vermutet, lässt sich die Kategorie 'hadar-anticipation' mit

Tabelle 7.1: Ausgewählte Ergebnisse der Kreuzvalidierung der Kategorie-spezifischen SVMs

Kategorie	TPR	TNR	gamma	nu
'hadar-yes'	0.70	0.80	0.16	0.16
'hadar-anticipation'	0.91	0.93	0.03	0.16
'hadar-synchronisation'	0.71	0.83	0.001	0.16

Erkennungsraten über 90% am besten von den anderen Klassen abgrenzen. Aber auch für 'hadar-yes' und 'hadar-synchronisation' ergeben sich mit Erkennungsraten von ca. 70% signifikante Ergebnisse, die die These von Hadar stützen.

7.1.5 Fazit

In diesem Abschnitt wurden die drei Kategorien 'Bestätigung', 'Antizipation' und 'Synchronisation' des Modells für Kopfgesten des Zuhörers nach Hadar im Dialogkontext dargestellt, anhand eines Datensatzes im Kontext von Mensch-Maschine-Dialogen nachvollzogen und auf automatische Klassifizierbarkeit hin untersucht.

Es zeigte sich, dass sich die Kategorien auch in dem Datensatz mit Mensch-Maschine-Dialog wiederfinden ließen. Erwartungsgemäß lassen sich dort die meisten Kopfnick-Gesten der Kategorie 'Bestätigung' einordnen. Jedoch war insbesondere die Kategorie 'Synchronisation' selten anzutreffen. Ein Grund hierfür könnte der maschinelle Redefluss des virtuellen Agenten sein, der sich von einem natürlichen Sprachverhalten merkbar unterscheidet. Für nähere Betrachtungen der Kategorie 'Synchronisation' wären daher weitere Untersuchungen wie der Vergleich mit einem Mensch-Mensch-Dialog interessant. Antizipations-Nicken nimmt dabei eine Sonderstellung ein; Obwohl Hadar zwar jede Geste mit vertikaler Kopfbewegung als Kopfnicken definiert, handelt es sich bei dieser Geste oft um eine lineare Bewegung und nicht um ein Kopfnicken gemäß gängiger Definitionen wie sie in Kapitel 3 erläutert wurden. Daher sind vor allem die beiden anderen Kategorien 'Bestätigung' und 'Synchronisation' von besonderer Relevanz, da sie beide die Definition

für Kopfnicken voll erfüllen.

In der Datenanalyse wurden bereits einige physikalische Unterschiede zwischen den Kategorien deutlich. Während Bestätigungs-Nicken und Synchronisations-Nicken einem sehr ähnlichen gemeinsamen Muster folgen, zeigte sich vor allem in der durchschnittlich größeren Amplitude bei Bestätigung ein deutlicher Unterschied.

Einige physikalische Eigenschaften konnten mit dem getesteten Ansatz nicht angemessen berücksichtigt werden. So wurde die Gesamtlänge einer Kopfgeste durch die notwendigerweise fixe Fensterbreite nicht als Merkmal berücksichtigt. Auch die Betrachtung der Anzahl der Nick-Zyklen konnte aus diesem Grund nicht miteinbezogen werden. Dennoch konnte gezeigt werden, dass die durch die Merkmalsextraktion erfasste Information ausreicht, um signifikante Ergebnisse in der Klassifikation zu erzielen.

Insgesamt konnten die Beobachtungen von Hadar, die aufgezeigten Kopfgesten würden sich nicht nur in der Funktion, sondern auch in den physikalischen Eigenschaften unterscheiden, bestätigt werden.

Für den Anwendungsfall eines interpretierenden Systems könnte es im konkreten Fall Folgendes bedeuten: Ein Nicken der Kategorie Antizipation wurde meistens kurz bevor die Person etwas sagen wollte ausgeführt und kann also als nonverbale Ankündigung einer Äußerung gelten. Oft wurde sie daraufhin vom virtuellen Agenten unterbrochen. Würde Antizipation von einem System automatisiert erkannt werden, könnten derartige Unterbrechungen vermieden werden und das System könnte dem menschlichen Dialogpartner bewusst 'zuhören' und ausreden lassen. Auch könnte ein erkanntes Synchronisations-Nicken als ein Hinweis auf bewusstes Zuhören gedeutet werden und dem System die Rückmeldung eines aktiven Zuhörers geben.

Einschränkend sollte dabei erwähnt werden, dass die Annotation der Nick-Exemplare in dieser Studie nach eigenem Ermessen gemäß der Kategorie-Beschreibung durchgeführt wurde; Für eine stärkere Aussagekraft sollte dieser Schritt jedoch von einem oder mehreren unabhängigen Kommunikations-Experten erfolgen.

7.2 Feedback in Form von Zuhörer-Nicken: P1-P3

In diesem Abschnitt soll der Fokus auf dem Zuhörer-Verhalten liegen. Wie im Kapitel 3.2 beschrieben, wurde in der Arbeit von Poggi et al. (2010) ein Modell für unterschiedliche Bedeutungskategorien von Kopfnicken vorgeschlagen (Abbildung 3.4). Darin wurde auch angekündigt, es würden auf Basis dieses Modells weitere Arbeiten zur automatischen Erkennung folgen. Entsprechende Veröffentlichungen blieben bisher jedoch aus.

Insbesondere bezüglich des Backchannel-Verhaltens einer zuhörenden Person wurde eine Kategorisierung für Bedeutungsformen des Kopfnickens vorgenommen. So bilden die Kategorien 1.B1 bis 1.B3 einen steigenden Grad an 'Commitment'-Stärken wie folgt:

1.B1 Bedeutet etwa 'Ich bestätige, dass ich dir zuhöre.'

1.B2 Bedeutet etwa 'Ich halte das Gesagte für relevant.'

1.B3 Bedeutet etwa 'Ich teile deine Meinung, ich stimme dir zu.'

Um zu untersuchen, ob eine derartige Einteilung auch auf einer Datenbasis von Merkmalsvektoren wiederzufinden ist, wird zunächst Videomaterial mit entsprechenden Annotationen benötigt. Eine ähnliche Kategorisierung von Zuhörer-Feedback findet sich in der Arbeit von Malisz et al. (2016), in der auch der sogenannte ALICO-Korpus als Videomaterial vorgestellt wurde. Dieser beinhaltet 20 Interaktionen zwischen einer erzählenden und einer zuhörenden Person. Dabei wurde insbesondere das Zuhörer-Verhalten detailliert analysiert und annotiert. Unter anderem kam ein Kodierschema zum Einsatz, das drei ansteigende Intensitätsstufen von Grounding-Tiefe oder Bewertungs-Stärke beschreibt. Es zeigt sich eine starke Ähnlichkeit der Bedeutungs-Kategorien. Definiert wurden diese grundlegenden funktionalen Feedback-Kategorien: (Buschmeier et al. (2011)

P1 Bedeutet etwa 'Perzeption erfolgt, ich höre, bitte fahre fort'.

P2 Bedeutet etwa 'Verstehen was gesagt wurde'.

P3 Bedeutet etwa 'Akzeptanz/Zustimmung mit Sprecher-Aussage'.

Diese drei Kategorien P1-P3 können somit ähnlich zu 1.B1-1.B3 als hierarchische Skala mit steigender Stärke von Commitment angesehen werden.

Diese drei Kategorien wurden von drei unabhängigen Annotatoren annotiert, wobei Unstimmigkeiten gemeinsam disktutiert wurden und durch Mehrheitsentscheidungen die letztendlichen Annotationen gewählt wurden. Auffällig war hierbei, dass die Kategorie P2 sowohl am häufigsten verwendet wurde, als auch die größte Übereinstimmung zwischen verschiedenen Annotatoren aufwies.

In der Kategorie P3 war man sich jedoch besonders oft uneinig.

Erwähnenswert ist dabei, dass laut Definition die Kategorien sich nicht gegenseitig ausschließen, sondern teilweise sogar bedingen. So impliziert die Kategorie P3, dass auch P1 und P2 erfüllt sind. Schließlich muss man eine Aussage gehört und verstanden haben, um dieser auch zustimmen zu können. Dabei wurde auch das meiste Nicken einer der ersten beiden Kategorien zugeordnet.

Die annotierten Kategorien P1-P3 des ALICO-Korpus sollen nun im folgenden Abschnitt analysiert werden.

7.2.1 Datensatz und Analyse

Um die annotierten Nickgesten für eine Datenanalyse zugänglich zu machen, werden Merkmalsvektoren mitilfe des Kopfwinkel-Schätzers aus Kapitel 5.5 extrahiert.

Anhand einer Untersuchung der annotierten Zeitintervalle wie sie in der Tabelle 7.2 aufgelistet sind, wurde für die gemeinsame Merkmalsgrundlage ein 600ms Zeitfenster festgelegt. Durch ein einheitliches Zeitfenster für alle Annotationen lassen sich alle Merkmalsvektoren in denselben Merkmalsraum einbetten und somit gut analysieren.

Erste Auffälligkeiten zeigen sich in der Klasse P3. So sind die annotierten Zeitintervalle in dieser Klasse im Mittel etwa 100 Millisekunden kürzer als in den anderen Klassen. Auch zeigen die Daten eine wesentlich stärkere Ausprägung der nach unten gerichteten Nickbewegung als in den weiteren Klassen.

Aus den annotierten Videos wurden Frame-weise Kopfwinkel mit einer fixen Fensterbreite von 600ms extrahiert. Mithilfe der Annotationen wurden aus den Merkmalen diejenigen mit dem entsprechenden Zeitstempel aussortiert und mit einem der drei Kategorien versehen.

Tabelle 7.2: Annotierte Zeitintervalle in Millisekunden der Kategorien P1, P2, P3.

Kategorie	Mittelwert	Median	Max.Länge von 90%
'P1'	620	600	800
'P2'	658	640	880
'P3'	552	520	720
'P1+P2+P3'	623	840	800

Die Abbildung 7.3 zeigt oben für jede Kategorie ein Koordinatensystem, in dem alle annotierten Merkmalsvektoren der jeweiligen Kategorie übereinander gezeichnet wurden. Darunter zeigt sich für jede Kategorie der über alle Instanzen gemittelte Merkmalsvektor. Während sich dieser für P2 und P3 nicht wesentlich unterscheidet, so grenzt sich P1 doch deutlich durch eine insgesamt schwächere Nick-Amplitude von den anderen Kategorien ab.

Für jede der drei Kategorien wurde jeweils ein spezialisierter SVM-Klassifikator trainiert, indem die vorgegebene Klasse als Positiv markiert und die jeweils übrigen beiden Klassen als Negativ markiert wurden. Die folgende Tabelle 7.3 zeigt nun die Ergebnisse der Kreuzvalidierung für die Kategorien P1, P2 und P3.

Tabelle 7.3: Ergebnisse der Kreuzvalidierung eines SVM-Klassifikators für die Kategorien P1, P2 und P3. Als Negative wurden jeweils die anderen beiden Kategorien genommen.

Kategorie	Anz.Positive	Anz.Negative	TPR	TNR
'P1'	60	113	0.70	0.50
'P2'	69	104	0.61	0.57
'P3'	44	129	0.57	0.57

Die Erkennungsergebnisse sind allerdings nur etwas besser als der Zufall.

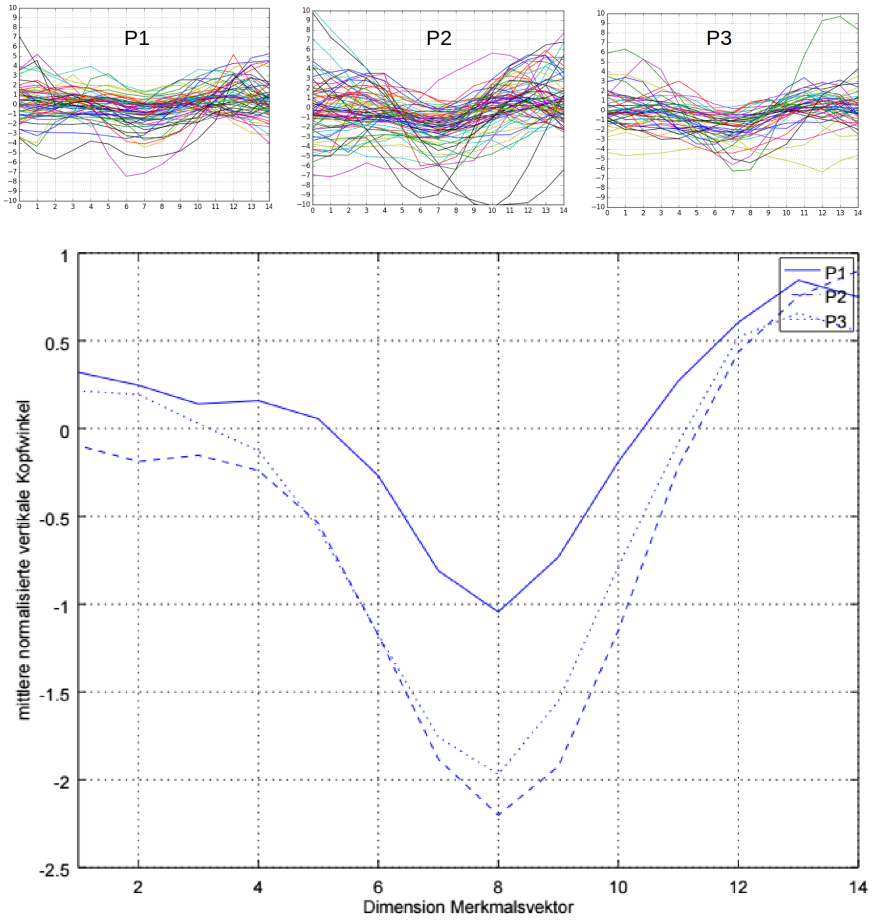


Abbildung 7.3: Oben sind die annotierten Instanzen der drei Kategorien P1,P2 und P3 dargestellt. Mittelt man die Mengen der Instanzen in den einzelnen Dimensionen, so ergeben sich die unten gezeichneten Merkmalsvektoren (durchgehende Linie für P1, gestrichelt für P2 und gepunktet für P3).

Wie es anhand der Abbildung 7.3 bereits zu erwarten war, lässt sich P1 noch am besten von den übrigen Kategorien abgrenzen.

7.2.2 Ergebnisse und Diskussion

In diesem Unterkapitel wurde einer Kategorisierung von Poggi et al. (2010) bezüglich dreier Commitment-Stufen nachgegangen. Es wurde mit dem ALICO-Korpus ein Datensatz mit ähnlicher Annotations-Kategorien ausfindig gemacht und für eine entsprechende Datenanalyse verwendet. Aus den annotierten Zeitintervallen wurden diejenigen mit erkennbarem Kopfnicken herausgegriffen und als Merkmalsvektoren extrahiert. Während die Unterschiede zwischen den Kategorien P2 und P3 auf Basis der Merkmale nicht signifikant ausfielen, zeigte sich jedoch eine Tendenz in der ersten Kategorie mit einer erhöhten Vorhersage-Wahrscheinlichkeit. Somit wäre zumindest eine Tendenz zu einer zwei-stufigen Commitment-Hierarchie erkennbar. Mit einer derartigen Einstufung eines Kopfnickens in eine vergleichbare Hierarchie könnte von einem System genutzt werden, um zum Beispiel nicht nur binär ein Kopfnicken zu erkennen, sondern zwischen starkem und schwachem Feedback zu unterscheiden. Bei erkanntem Kopfnicken, das jedoch in die erste Kategorie eingeordnet wird, könnte somit nochmal explizit nachgefasst werden, ob die entsprechende Information tatsächlich verstanden wurde.

7.3 Ausblick für weitere Anwendungen

Eine ausführliche Untersuchung der restlichen im Abschnitt 3.2 vorgestellten Interpretationsansätze auf der Grundlage realer Daten würde den Rahmen dieser Dissertation sprengen. Daher seien im Folgenden weitere besonders interessante mögliche Anwendungsbereiche erwähnt.

7.3.1 Backchannel oder Turn-Taking

In den Arbeiten beispielsweise von (Yngve (1970) wird besagt, dass alle Signale, die nicht eine Turn-Taking Funktion innehaben, als Backchannel zu verstehen sind. Somit sind nach dieser Kategorisierung alle Kopfnicken entweder Turn-Taking Signale oder Backchannel Signale.

Ist es möglich, bereits an den Eigenschaften eines Kopfnickens mit einer über dem Zufall liegenden Wahrscheinlichkeit zu erkennen, ob es sich um ein Turn-Taking Signal handelt, so könnte dadurch die Regulation innerhalb eines Gesprächs verbessert werden. Eine Anwendungsmöglichkeit für ein Assistenzsystem wäre, bei einem wahrscheinlichen Turn-Taking Signal

des Zuhörers bewusst eine kurze Pause für Feedback einzulegen, um diesem die Möglichkeit zu geben, seinen Anspruch auf Wortergreifung nachzukommen.

7.3.2 Anpassung der Agenten-Verbosität anhand des Nutzer-Nickverhaltens

Verbosität wird auch als Wortreichtum bezeichnet. Ein hoher Grad an Verbosität bedeutet also, dass viele Wörter verwendet werden und besonders ausführlich formuliert wird.

Da bei Assistenzsystemen ein Gesprächsverhalten implementiert wird, wird dabei auch ein bestimmmter Grad an Verbosität festgelegt. Fasst der Agent sich in seinen Ausführungen sehr kurz, so kann es mancher Nutzer als ungenau oder unhöflich empfinden. Drückt dieser sich jedoch sehr ausführlich aus, zeigt ein Nutzer gerade bei häufiger Nutzung oft Ungeduld. In der Regel wird ein Systemverhalten programmiert, welches möglichst einen guten Kompromiss für den konkreten Anwendungsfall darstellen soll.

Besser als eine Kompromisslösung wäre, Evidenzen für die Nutzererwartung an Verbosität zu registrieren und sich daran anzupassen.

Laut der Veröffentlichung von Matarazzo et al. (1964) gilt, dass häufiges Nicken eines Zuhörers dazu führt, dass der Sprecher eher bereit dazu ist, seine Mitteilungen umfangreicher und mit mehr Worten auszuführen.

Überträgt man diese beobachtete menschliche Eigenschaft auf das Systemverhalten, so könnte das System bei überdurchschnittlich häufigem Erkennen von Kopfnicken in einen 'ausführlicheren Sprechmodus' und bei fast ausbleibenden Nick-Signalen in einen sehr kurzgefassten Sprechmodus schalten, um so dynamisch auf verschiedene Nutzerprofile zu reagieren.

Um diesen Effekt zu untersuchen, wäre etwa folgende Studie sinnvoll:

Eine Menge an Versuchspersonen werden durch zwei aufeinanderfolgende ähnlich gestaltete Interaktionen geführt. In der ersten Interaktion wird ein Nutzerprofil bezüglich des Nickverhaltens extrahiert. Die jeweils zweiten Interaktionen werden in zwei Konditionen aufgeteilt; In der einen Kondition äußert sich der Gesprächspartner oder virtuelle Agent sehr verbos, während dieser sich in der zweiten Kondition sehr kurz fasst. Anschließend wird die Nutzerzufriedenheit bezüglich der erfahrenen Kommunikation abgefragt. Die Hypothesen wären hierbei folgende:

Nutzer mit häufigem Kopfnicken in der ersten Runde sind mit stark verbosem Systemverhalten zufriedener als Nutzer mit wenig Kopfnicken.

Eine einfache technische Umsetzung zur Extraktion eines Nutzerprofils für Kopfnicken in einem Live-System könnte folgendermaßen aussehen: Es wird eine Aktivierungsfunktion definiert, die durch Nicken gezeigtes Engagement repräsentiert. Die Aktivierung sinkt stetig mit der laufenden Interaktions-Zeit. Durch Detektion von Kopfnicken wird die Aktivierung mit einem Wert aufaddiert und somit gestärkt. Ab einem Aktivierungs-Schwellwert schaltet der Agent in einen verboseren Zustand. Wird dieser Schwellwert eine gewisse Zeit unterschritten, verlässt dieser ihns wieder.

So könnte eine Art einfühlsames Verhalten des Agenten verwirklicht werden, das sich ein Stück weit dem kommunikativen Nutzerprofil eines Nutzers anpasst.

8 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde ein Kopfnick-Erkennungssystem entwickelt und Kopfnicken auf Interpretationsansätze hin untersucht.

Zuerst wurden in dem Kapitel 2 grundlegende relevante Methoden der Bildverarbeitung und Verfahren zur automatischen Detektion erklärt. Neben den technischen Aspekten der Bildverarbeitung und Gestenerkennung wurde anschließend im Kapitel 3 das menschliche Kopfnicken hinsichtlich der sozialen Bedeutsamkeit und der Rolle in der zwischenmenschlichen Kommunikation erörtert. Verschiedene Erkenntnisse über die Interpretation von Kopfnicken und dessen funktionelle Eigenschaften wurden aus aktuellen wissenschaftlichen Untersuchungen herausgearbeitet. Die gewonnenen Erkenntnisse wurden dahingehend untersucht, ob sie derart automatisiert erkannt und verarbeitet werden können, so dass ein virtueller Agent daraus einen für den Kontext relevanten Informationsgewinn erzielen könnte.

Die Entwicklung eines Systems zum Erkennen von Kopfnicken wurde in Kapitel 5 in mehreren Schritten aufgezeigt. So wurde eine Support Vector Regression (SVR) zur Kopfwinkelschätzung trainiert und dessen Ergebnisse als Merkmale zur Detektion von Kopfnicken genutzt. Als Detektions-System wurde ein auf Dynamic Time Warping basierender Algorithmus entwickelt. Alternativ wurde auch ein Ansatz mittels Support Vector Machine verfolgt und umgesetzt.

Das darauffolgende Kapitel 6 befasste sich anschließend mit der Auswertung und Beurteilung der Leistungsfähigkeit der im vorigen Kapitel ausgearbeiteten Ansätze zur Detektion von Kopfnicken. In dem Rahmen wurden die Besonderheiten und Eigenheiten ausgearbeitet, die bei der Evaluation von kommunikativen Gesten wie Kopfnicken relevant sind. Zur Evaluation wurden angepasste Verfahren entwickelt und angewendet. Es zeigte sich insgesamt eine Leistungsfähigkeit, die zwar leicht unter vergleichbaren aktuellen State-Of-The-Art Ansätzen liegt, sich jedoch dennoch aufgrund der besonders geringen Anforderungen an Hardware hervorhebt.

Unter den beiden Detektionssystemen lieferte der auf SVM basierte An-

satz geringfügig bessere Ergebnisse als DTW. Ein Vorteil von SVM zeigte sich in der besseren Verwertbarkeit der Merkmalsdaten in Form von Normalisierung der Rohdaten, während bei DTW durch Verwendung von Ableitungen Informationen verloren gehen. Zudem wird bei SVM durch eine beliebige Menge von Stützvektoren der Klassifikationsraum umfassender definiert. Demgegenüber wird bei DTW eine Lösung lediglich aufgrund der Differenz zu einem einzelnen Referenz-Prototyp ausgemacht. Dennoch bietet DTW auch besondere Vorteile und Potentiale. So ist man nicht wie bei SVM an eine fixe Fensterbreite gebunden, was gerade bei starken Variationen in den Annotations-Zeitintervallen vorteilhaft ist. Zudem ist statt der Suche nach einem einzelnen Referenz-Prototypen auch eine Suche nach einer Gruppe von Referenz-Prototypen zum Beispiel durch eine Cluster-Analyse denkbar, um den Klassifikations-Raum wesentlich genauer abzudecken. Entsprechende Ansätze mit mehreren Prototypen erfordern im Trainings-Schritt allein schon aufgrund verschiedener Schwellwerte pro Prototyp mathematische Optimierungs-Heuristiken, um eine Kreuzvalidierung zu realisieren.

Die geringen Leistungsunterschiede zwischen DTW und SVM können aber auch ein Hinweis darauf sein, dass beide Verfahren die verfügbaren Informationen aus den Merkmalen gut nutzen und die möglichen Ergebnisse eher aufgrund unzulänglicher Merkmale oder Annotation beschränkt werden. Wesentliche Verbesserungen wären also erst durch eine effektivere Merkmalsgenerierung durch den Kopfwinkelschätzer denkbar. Der Bereich der Kopfwinkelschätzung ist in der aktuellen Forschung immer noch ein großes Thema und es werden viele vielversprechende neue Verfahren vorgestellt, die in das vorgestellte System modular integriert werden könnten.

Anschließend wurden im Kapitel 7 Ansätze zur automatischen Interpretation und Kategorisierung von Kopfnicken erörtert und an Datensätzen getestet. Das Modell von Hadar mit den drei Kategorien für Zuhörer-Nicken 'Bestätigung', 'Antizipation' und 'Synchronisation' ließ sich anhand von Testdaten nachvollziehen. So wurde gezeigt, dass auf dieser Basis Anwendungen zur Erkennung von nonverbaler Ankündigung von Äußerungen, sowie zur Verringerung von Unterbrechungen denkbar sind.

Auch das 3-stufige Modell von 'Commitment'-Stärken von Poggi lieferte Hinweise darauf, dass sich die physikalische Ausführung von Kopfnicken zumindest zwischen schwachem und stärkerem Commitment mit einer über dem Zufall liegenden Wahrscheinlichkeit unterscheidet.

Des Weiteren wurde ein Ausblick über weitere Anwendungsmöglichkeiten

von Verarbeitung von Kopfnicken aufgezeigt. So wurde ein Modell zur dynamischen Anpassung der System-Verbosität auf Basis einer Erfassung des Nutzer-Nick-Profils vorgeschlagen.

Als weiteren Ausblick sei besonders die Erweiterung des DTW-Verfahrens mit mehreren Prototypen erwähnt, welche neue Forschungsfragen eröffnet. Auch wäre es sicher sinnvoll, zur Erkennung von Kopfnicken nicht nur die physikalische Nick-Bewegung zu betrachten, sondern mit multimodalen Ansätzen den Dialog-Kontext wie Sprach-Rythmus und Emotionserkennung mit zu berücksichtigen.

Wesentliche Fortschritte in der Leistungsfähigkeit und Praxistauglichkeit derartiger Detektionssysteme werden möglicherweise dann zu verzeichnen sein, wenn in den kommenden Jahren heute noch spezielle Hardware wie hochauflösende Stereo-Kameras oder Grafikkarten-beschleunigte Deep-Learning-Ansätze auch in kleine Alltags-Geräte immer mehr Einzug finden.

Literaturverzeichnis

Adolphs, S. and Carter, R. (2007). Beyond the word: New challenges in analysing corpora of spoken English. *European journal of English studies*, 11(2):133–146. 3.2

Asor, E. (2014). *The timing of head nods is constrained by prosodic structure*. PhD thesis. 3.2

ASUSTeK (2019). Bewegungssensor Xtion Pro 'www.asus.com/de/3D-Sensor' (zuletzt aufgerufen am 01.11.2019). 2.2

Bamoallem, B. S., Andrew, J., and Gordon, M. (2016). The impact of head movements on user involvement in mediated interaction. *Computers in Human Behaviour*, 55:424–431. 3.2, 3.5, 3.6

Bavelas, J., Gerwing, J., Sutton, C., and Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language - J MEM LANG*, 58:495–520. 1.1

Bellman, R. and Kalaba, R. (1957). Dynamic Programming and Statistical Communication Theory. In *Proc. of the National Academy of Sciences of the United States of America*, volume 43, pages 749–751. 2.7.1

Bernieri, F. and Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. *RS Feldman & B. Rimé (Eds.), Studies in emotion & social interaction. Fundamentals of nonverbal behavior*, pages 401–432. 3.3

Birdwhistell, R. L. (1970). *Kinesics and context*. 3.2, 3.6

Boholm, M. and Allwood, J. (2010). Repeated head movements, their function and relation to speech. In *Proc. of the 2010 LREC*. 3.2

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. 5.1

Brinol, P. and Petty, R. E. (2003). Overt Head Movements and Persuasion: A Self-Validation Analysis. *Journal of Personality and Social Psychology*, 84(6):1123–1139. 3.3

Brown, R. (1986). *Social Psychology*. 3.2

Buschmeier, H. and Kopp, S. (2011). Towards Conversational Agents That Attend to and Adapt to Communicative User Feedback. In *Proceedings ofthe 11th International Conference on Intelligent Virtual Agents*, pages 169–182. 4.2

Buschmeier, H., Malisz, Z., Włodarczak, M., Kopp, S., and Wagner, P. (2011). 'Are you sure you're paying attention?' – 'Uh-huh' Communicating understanding as a marker of attentiveness. In *Proc. of the 2011 Interspeech*, number August, pages 2057–2060. 7.2

Canton-Ferrer, C. (2006). Head Pose Detection Based on Fusion of Multiple Viewpoint Information. In *Multimodal Technologies for Perception of Humans*. 2.5

Capper, S. (2000). Nonverbal Communication and the Second Language Learner: Some Pedagogical Considerations. *The Language Teacher 24*, 5:19–23. 3.2

Carmona, J. M. and Climent, J. (2012). A Performance Evaluation of HMM and DTW for Gesture Recognition. In Alvarez, L., Mejail, M., Gomez, L., and Jacobo, J., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 7441 of *Lecture Notes in Computer Science*, pages 236–243. Springer Berlin Heidelberg, Berlin, Heidelberg. 2.6

Chen, Y., Yu, Y., and Odobez, J.-M. (2015). Head Nod Detection from a Full 3D Model. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 528–536. IEEE. 2.6, 6.2, 6.6, 6.7

Chu, S. and Tanaka, J. (2012). Head Nod and Shake Gesture Interface for a Self-portrait Camera. *ACHI 2012 - 5th International Conference on Advances in Computer-Human Interactions*. 2.6, 6.2

Clancy, P. (1982). Written and spoken style in Japanese narratives. In *D. Tannen - Spoken and Written Language: Exploring Orality and Literacy*, pages 55–76. 3.2

Darwin, C. (1872). The expression of the emotions in man and animals. *London: John Murray, Albemarle Street.* 3.2

Davis, J. W. and Vaks, S. (2001). A Perceptual User Interface for Recognizing Head Gesture Acknowledgements. In *Proc. of the 2001 workshop on Perceptive user interfaces*, pages 1–7. 2.6, 6.2

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conservations. *Journal of Personality and Social Psychology*, 23(2):283–292. 3.2

Ekman, P. and Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press, Palo Alto, CA*. 2.4

Fanelli, G., Weise, T., Gall, J., and Gool, L. V. (2011). Real time head pose estimation from consumer depth cameras. In *DAGM'11 Proceedings of the 33rd international conference on Pattern recognition*, pages 101–110. Springer-Verlag. (document), 5.5, 5.4

Förster, J. and Strack, F. (1996). Influence of Overt Head Movements on Memory for Valenced Words : A Case of Conceptual-Motor Compatibility. *Journal of Personality and Social Psychology*, (October 1996). 3.3

Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., and Kobayashi, T. (2004). A conversation robot using head gesture recognition as para-linguistic information. *IEEE International Conference on Robot and Human Interactive Communication.* 2.6

Fusaro, M., Vallotton, C. D., and Harris, P. L. (2014). Beside the point: Mothers' head nodding and shaking gestures during parent–child play. *Infant Behavior and Development*, 37(2):235–247. 3.2

Giges, B. (1975). ÜSING YOUR HEAD": Notes on Nodding. *Transactional analysis journal : formerly: The Transactional analysis bulletin*, 5(3):264–266. 3.2

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11):419–429. 1.1

Gopakumar, K. and Suni, S. (2016). A real time decision support system using head nod and shake. *2016 International Conference on Circuit, Power and Computer Technologies*. 2.6, 6.3

Graumann, W. and Sasse, D. (2003). *CompactLehrbuch Anatomie*. 3.1

Guo, G., Dyer, C. R., and Huang, T. S. (2008a). Head Pose Estimation : Classification or Regression ? In *Proc. of the 19th International Conference on Pattern Recognition 2018*. 2.5

Guo, G., Huang, T. S., and Dyer, C. R. (2008b). Locally Adjusted Robust Regression for Human Age Estimation. In *Proc. of the 2008 IEEE Workshop on Applications of Computer Vision*. 2.5

Hadar, U., Steiner, T., and Clifford Rose, F. (1985). Head Movement during listening turns in conservation. *Journal of Nonverbal Behavior*, 9(4):214–228. (document), 3.2, 3.2, 3.4, 3.6, 7.1

Hale, J., Ward, J. A., Buccheri, F., Oliver, D., and Hamilton, A. F. D. C. (2018). Are you on my wavelength? Interpersonal coordination in naturalistic conversations. (November). 3.3

Hall, G., Knapp, D. E., and Huemmrich, K. F. (1997). Physically based classification and satellite mapping of biophysical characteristics in the southern boreal forest. *Journal of Geophysical Research*, 102(97):29,567–29,580. 3.2

Harrigan, J. A., Rosenthal, R., and Scherer, K. R. (2005). *The new Handbook of Methods in Nonverbal Behaviour Research*. 1

Heitmann, A., Guttkuhn, R., Aguirre, A., Trutschel, U., and Numerics, A. (2001). Technologies for the Monitoring and Prevention of Driver Fatigue. In *Proc. of the First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*. 2.2

Helweg-Larsen, M., Cunningham, S. J., Carrico, A., and Pergram, A. M. (2004). To Nod or Not to Nod: An Observational Study of Nonverbal

Communication and Status in Female and Male College Students. *Psychology of Women Quarterly*, 28:358–361. 3.3

Horn, B. K. and Schnuck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203. 2.5

Iddan, G. and Yahav, G. (2000). 3D Imaging in the studio. *SPIE*, 4298:pp. 48. 2.2

Jacquin, A., Eleftheriadis, A., Laboratories, T. B., and Hill, M. (1995). Automatic location tracking of faces and facial features in video sequences. In *Proc. of the International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland*, number June. 2.6

Johnson, D. O. and Cuijpers, R. H. (2013). Predicting gaze direction from head pose yaw and pitch. In H.R. Arabnia, L. Deligiannidis, J. Lu, F.G. Tinetti, J. Y., editor, *Proceedings of the IPCV'13 - The 2013 International Conference on Image Processing, Computer Vision, & Pattern Recognition*, pages 662–668, Las Vegas, Nevada, USA. World Academy Of Science. 2.5

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems 1. *Journal of Basic Engineering*, 82(1):35–45. 2.6

Kapoor, A. and Picard, R. W. (2001). A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces - PUI '01*, New York, USA. ACM Press. 2.6, 6.2

Kawato, S. and Ohya, J. (2000). Real-Time Detection of Nodding and Head-Shaking by Directly Detecting and Tracking the 'Between-Eyes'. In *Proceedings ofthe Fourth IEEE International Conference on Automat-ic Face and Gesture Recognition 2000*, page 40. 2.6, 6.2

Kazemi, V. and Sullivan, J. (2014). One Millisecond Face Alignment with an Ensemble of Regression Trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. (document), 2.4, 2.1, 5.4

Kendon, A. (1972). Some relationships between body motion and speech. *A. Seigman and B.Pope, editors, Studies in Dyadic Communication*, pages 177–216. 3.2

Kihara, H., Fukushima, S., and Naemura, T. (2016). Analysis of Human Nodding Behavior during Group Work for Designing Nodding Robots. *Proceedings of the 19th International Conference on Supporting Group Work - GROUP '16*, pages 433–436. 3.5

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *The Journal of Machine Learning Research*, 10:1755–1758. 5.1

Kopp, S., Brandt, M., Buschmeier, H., Cyra, K., Krämer, N., Kummert, F., Opfermann, C., Pitsch, K., Schillingmann, L., Straßmann, C., and Wall, E. (2018). Conversational Assistants for Elderly Users – The Importance of Socially Cooperative Dialogue. In *Proceedings of the AAMAS 2018 Workshop on Intelligent Conversational Agents in Home and Geriatric Care Applications*, pages 10–17. 4

Kreienbrock, L., Pigeot, I., and Ahrens, W. (2012). *Epidemiologische Methooden*. 6

Kurtenbach, G. and Hulteen, E. (1990). Gestures in Human-Computer Communication. *The Art and Science of Interface Design*, pages 309–317. 2.1

Langholz, E. H. and Brasher, R. (2018). Real-time on-device nod and shake recognition. pages 1–6. 2.6, 6.2

Lee, C., Lesh, N., Sidner, C. L., Morency, L.-p., and Kapoor, A. (2004). Nodding in Conversations with a Robot. In *Extended abstracts of the 2004 Conference on Human Factors in Computing Systems*. 2.6, 6.2

Lee, J., Neviarouskaya, A., and Marsella, S. (2009a). Learning Models of Speaker Head Nods with Affective Information. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops 2009. (ACII 2009)*. 3.5

Lee, J., Rey, M., and Rey, M. (2009b). Learning a Model of Speaker Head Nods using Gesture Corpora. In *Proc. of the 8th International Conference on AAMAS*, pages 289–296. 3.5

Li, R. and Danielsen, M. (2007). Head Pose Tracking and Gesture Detection Using Block Keywords :. In *Proc. of the 2007 Mobility*, number 2, pages 572–575. 2.6

Liu, C., Ishi, C. T., Ishiguro, H., and Hagnita, N. (2012). Generation of Nodding, Head Tilting and Eye Gazing for Human-Robot Dialogue Interaction. In *HRI '12 Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 285–292. 3.5

Lucas, B. D. and Kanade, T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proc. of Imaging Understanding Workshop*, pages 121–130. 2.6

Mahalanobis, P. C. (1936). On the generalized distance in statistics. In *Proc. of the National Institute of Science of India. Vol. 2, Nr. 1*, pages 49–55. 2.7.1

Malisz, Z., Włodarczak, M., Buschmeier, H., Skubisz, J., Kopp, S., and Wagner, P. (2016). The ALICO corpus: analysing the active listener. *Language Resources and Evaluation*, 50(2):411–442. 7.2

Matarazzo, J., Saslow, G., Wiens, A., M, W., and Allen, B. (1964). No Title. *Psychotherapy: Theory, Research & Practice*, 1(2):54–63. 7.3.2

Maynard, S. K. (1987). Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. Journal of Pragmatics. *Journal of Pragmatics*, 11:589–606. 3.2

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7):855–878. 3.2

McGrew, W. (1972). An ethological study of children's behavior. 3

Mehrabian, A. (1972). Nonverbal communication. *Aldine-Atherton, Illinois: Chicago*. 3.3

Mehrabian, A. and Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, pages 109–114. 1

Meyer, C. and Stiller, A. (2011). Was hinter dem Sprachassistenten Siri steckt. *Mac & i Heft 4*, 4:146–149. 1

Microsoft (2019). Microsoft Kinect 'https://developer.microsoft.com/en-us/windows/kinect' (zuletzt aufgerufen am 01.11.2019). 2.2

Morency, L.-P. and Darell, T. (2002). Stereo Tracking using ICP and Normal Flow Constraint. *Object recognition supported by user interaction for service robots*, pages 367–372. 5.5

Morency, L.-p. and Darrell, T. (2003). Adaptive View-Based Appearance Models. 2.6

Moretti, S. and Greco, A. (2018). Truth is in the head. A nod and shake compatibility effect. *Acta Psychologica*. 3.3

Morimoto, C. H., Koons, D., Amir, A., and Flickner, M. (2000). Pupil detection and tracking using multiple light sources. *Image and Vision Computing*, 18(4):331–335. 2.6

Moubayed, S. A., Beskow, J., and Centre, K. T. H. (2009). Effects of Visual Prominence Cues on Speech Intelligibility. In *AVSP-2009*, pages 43–46. 3.2

Müller, M. (2007). *Information retrieval for music and motion*. 5.6.1

Murphy-Chutorian, E. and Trivedi, M. M. (2007). Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation. In *Proc. of the 2007 IEEE, Intelligent Transportation Systems Conference*, pages 709–714. 2.5

Murphy-Chutorian, E. and Trivedi, M. M. (2009). Head pose estimation in computer vision: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626. 2.5, 5.5

Navarretta, C., Ahlsén, E., Allwood, J., and Jokinen, K. (2012). Feedback in Nordic First-Encounters: a Comparative Study. In *Proceedings of the LREC 2012*. 3.2, 3.4

Nguyen, L., Odobez, J.-M., and Gatica-Perez, D. (2012). Using Self-Context for Multimodal Detection of Head Nods in Face-to-Face Interactions. *Icmi'12*, pages 1–4. 2.6, 6.2

Nori, F., Lipi, A. A., and Nakano, Y. (2011). Cultural Difference in Nonverbal Behaviors in Negotiation Conversations: Towards a Model for Culture-adapted Conversational Agents. *Stephanidis C. (eds) Universal Access in Human-Computer Interaction. UAHCI 2001*, 6765. 3.2, 3.4

Nunn, R. and Maya, T. (2003). Head Nodding in Intercultural Conversation. *Japan Journal of Multilingualism and Multiculturism*, 9:69–86. (document), 3.2, 3.3, 3.4

Oertel, C., Mora, K. A. F., and Odobez, J.-m. (2014). Who Will Get the Grant? A Multimodal Corpus for the Analysis of Conversational Behaviours in Group Interviews. In *Proc. of the UM3I*, pages 27–32. 6.2

Oka, K. (2005). Head Pose Estimation System Based on Particle Filtering with Adaptive Diffusion Control. In *Proc. of the IAPR Conference on Machine Vision Applications*, pages 586–589. 5.5

Ota, S., Jindai, M., Yasuda, T., and Sejima, Y. (2017). Development of Nodding Detection using Neural Network Based on Communication Characteristics. In *Proc. of the 56th Annual Conference of the Society of Instrument and Control Engineers of Japan*. 2.6

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine and Journal of Science*, 5982. 7.1.3

Petukhova, V. and Bunt, H. (2009). Grounding by nodding. In *GESPIN proceedings, vol. I*, number June, pages 4629–4629. 3.2, 3.6

Poggi, I., D'Errico, F., and Vincze, L. (2010). Types of Nods. The Polysemy of a Social Signal. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2570–2576. (document), 3.2, 3.4, 7.2, 7.2.2

Quinlan, J. R. (1986). Induction of Decision Trees. pages 81–106. 2.7

R. Rabiner, L. (1989). A Tutorial on Hidden MArkov Models and Selected Applications in Speech Recognition. In *Proc. of the IEEE*, pages 257–286. 2.7

Raghavendra, S. and Deka, P. C. (2014). Support Vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, 19:372–386. 2.7.2

Rekha, N. and Kurian, M. Z. (2014). Face Detection in Real Time Based on HOG. *International Journal of Advanced Research in Computer Engineering & Technology*, 3(4):1345–1352. 2.3

Rizvi, D. Q. M. (2011). A Review on Face Detection Methods. *Journal of Management Development and Information Technology*, 11. 2.3

Sacks, H. (1992). Lectures on Conversation. Technical report. 3.2

Sakoe, H. and S. Chiba (1971). A dynamic programming approach to continuous speech recognition. In *Proc. Int. Cong. Acoust.*, pages 20C–13. 2.7.1

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639. 5.6.2

Senechal, T., Mcduff, D., and Kaliouby, R. (2015). Facial Action Unit Detection using Active Learning and an Efficient Non-Linear Kernel Approximation. *IEEE International Conference on Computer Vision Workshop*. 2.4

Sidner, C. L., Lee, C., Morency, L.-P., and Forlines, C. (2006). The effect of head-nod recognition in human-robot conversation. *Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction - HRI '06*, page 290. 1.1

Simion, F. and Di Giorgio, E. (2015). Face perception and processing in early infancy : inborn predispositions and developmental changes. *Frontiers in psychology*. 2.3

Stigler, S. M. (1981). Gauss and the Invention of Least Squares. *The Annals of Statistics*, 9(3):465–474. 2.7.2

Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on Language and Social Interaction*, 41(1):31–57. 3.2

Svinhufvud, K. (2016). Nodding and note-taking: Multimodal analysis of writing and nodding in student counseling interaction. *Language and Dialogue 6:1*, 1:81–109. 3.2

T. Dittmann, A. and G. Llewellyn, L. (1968). Relationships Between Vocalizations and Head Nods as Listener Responses. *Journal of Personality and Social Psychology*, 9:79–84. 3.2

Tan, W. and Rong, G. (2003). A real-time head nod and shake detector using HMMs. *Expert Systems with Applications*, 25(3):461–466. 2.6

Terven, J., Raducanu, B., and Salas, J. (2014). Robust Head Gestures Recognition for Assistive Technology. In *Proc. of the 6th Mexican Conference, MCPR 2014 Cancun*. 2.6, 6.2

Thelen, E., Schöner, G., Scheier, C., and Smith, L. B. (2001). The dynamics of embodiment: a field theory of infant perseverative reaching. *The Behavioral and brain sciences*, 24:1–86. 1

Toivio, E. and Jokinen, K. (2012). Multimodal Feedback Signaling in Finnish. *Human Language Technologies - The Baltics Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012*, 9:247–255. 3.2

Valstar, M. and Pantic, M. (2006). Fully Automatic Facial Action Unit Detection and Temporal Analysis. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*. 2.4

Vapnik, V. and Chervonenkis, A. (1974). *Theory of pattern recognition*. 2.7.2

Viola, Paul and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. *International Journal of Computer Vision*, 57:137–154. 2.3

Wall, E., Schillingmann, L., and Kummert, F. (2017). Online Nod Detection in Human-Robot Interaction. In *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 5.6, 6.3, 6.6

Walsh, E. and Daems, W. (2015). An optical head-pose tracking sensor for pointing devices using IR-LED based markers and a low-cost camera. (November). 2.2

Wang, J.-G. and Sung, E. (2007). EM enhancement of 3D head pose estimated by point at infinity. *Image and Vision Computing*, 25. 5.5

Wei, H., Scanlon, P., LI, Y., Monaghan, D., and O'Connor, N. E. (2013). Real-time head nod and shake detection for continuous human affect recognition. In *Image Analysis for Multimedia Interactive Services (WIAMIS)*. 2.6, 6.2

Wells, G. L., Petty, R. E., and Wells, G. L. (1980). The Effects of Over Head Movements on Persuasion: Compatibility and Incompatibility of Responses The Effects of Overt Head Movements on Persuasion : Compatibility and Incompatibility of Responses. *Basic and Applied Social Psychology*, pages 219–230. 3.3

Wilbur, R. B. (2000). Phonological and Prosodic Layering of Nonmanuals in American Sign Language. In *The Signs of Language revisited*, pages 215–244. 3.2

Xiong, X. and Torre, F. D. (2013). Supervised Descent Method and its Applications to Face Alignment. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539. 2.6

Xu, J., Gannon, P. J., Emmorey, K., Smith, J. F., and Braun, A. R. (2009). Symbolic gestures and spoken language are processed by a common neural system. *Proc. of the National Academy of Sciences of the United States of America*, 106(49):20664–20669. 3.3

Xu, X. and Kakadiaris, I. A. (2017). Joint Head Pose Estimation and Face Alignment Framework Using Global and Local CNN Features. In *Proceedings of IEEE 12th International Conference on Automatic Face & Gesture Recognition*. 2.6

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting, Chicago Linguistics Society*, pages 567–578. 3.2, 7.3.1

Yu, W. and Gang, L. (2011). Head Pose Estimation Based on Head Tracking and the Kalman Filter. *Physics Procedia 22*, pages 420–427. 2.5

Zabulis, X., Sarmis, T., and Argyros, A. A. (2009). 3D head pose estimation from multiple distant views. In *British Machine Vision Conference, London, UK*. 2.5

Zelinsky, A. and Heinzmann, J. (1996). Real-Time Visual Recognition of Facial Gestures for Human-Computer Interaction. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition.* 2.6, 6.2

Zhu, Y. and Fujimura, K. (2003). 3D Head Pose Estimation with Optical Flow and Depth Constraints Kikuo Fujimura. In *Proc. of the Fourth International Conference on 3-D Digital Imaging and Modeling.* 2.5