Multimodale Bestimmung des visuellen Aufmerksamkeitsfokus von Personen am Beispiel aufmerksamer Umgebungen

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der Fakultät für Informatik des

genehmigte

Dissertation

von

Michael Voit

aus Pforzheim

Tag der mündlichen Prüfung:	**14.07.2011**
Erster Gutachter:	**Prof. Dr.-Ing. Rainer Stiefelhagen**
Zweiter Gutachter:	**Prof. Dr.-Ing. Jürgen Beyerer**

Kurzfassung

Mit dem Wissen darüber wohin eine Person ihren Blick richtet, lassen sich Rückschlüsse auf den Kontext ihres Handelns, und der Situation in der sie sich befindet, ziehen. Ein Beobachten ihrer Augenpaare erfordert allerdings eine restriktive und für die Person unnatürliche Sensoranbringung. Stetes Zugewandtsein zu einer frontal auf das Gesicht ausgerichteten Kamera oder ihrer dedizierten Anbringung an den Körper der Person, um auch bei Bewegung das Gesicht zuverlässig einfangen zu können, schränken Handlungs- und Bewegungsradius ein. In diesem Zusammenhang kann die Annäherung des Blicks anhand der Kopfdrehung als Möglichkeit angesehen werden, Personen aus gegebener Distanz zu beobachten und bei Fehlen einer hochauflösenden, frontalen Aufnahme, dennoch eine Aussage herzuleiten, wohin deren Aufmerksamkeit gerichtet wird.
In der vorliegenden Arbeit wird ein System entworfen, das visuell die Kopfdrehung von Personen schätzt und hiervon auf deren jeweilige visuelle Aufmerksamkeitszuwendungen schließt. Erkannt werden soll dabei auf welches Objekt oder welche weitere Person im Raum sie in ihren Handlungen Bezug nehmen.
Für eine verdeckungsfreie und vollständige Erfassung der Szene, werden hierzu mehrere Kameras für eine gleichzeitige Aufnahme aus unterschiedlichen Blickwinkeln eingesetzt. Damit wird den beteiligten Personen vollständige Bewegungs- und Handlungsfreiheit zugesichert, was aber in Folge bei großen Aufnahmedistanzen zu niedrig aufgelösten und so herausfordernden Kopfbeobachtungen führen kann. Es soll daher untersucht werden, in wie weit das Benutzen gleichzeitiger, komplementärer Ansichten diese Reduktion der Motivqualität ausgleichen und so zu einer hinreichend feingranularen Unterscheidung der Aufmerksamkeitszuwendungen beitragen kann.
Der erste Teil der Arbeit stellt hierzu die Implementierung eines Partikelfilters vor, der im Beobachtungsmodell Gradientenhistogramme und Künstliche Neuronale Netze einsetzt und damit eine Zustandshypothese des Kopfs hinsichtlich dessen dreidimensionaler Position, Größe und Orientierung im Raum schätzt. Damit das System invariant gegenüber der Anzahl vorhandener Kameras reagiert, werden die Zustandshypothesen in jede Bildebene projiziert, bewertet und erst auf Entscheidungsebene fusioniert. So können Ansichten zur Laufzeit entfernt, ignoriert oder weitere Kameras hinzugefügt werden ohne ein erneutes Einlernen einzelner Systemkomponenten zu erfordern. Als wissenschaftlicher Beitrag setzt sich dieser Abschnitt erstmalig mit dem Nutzen mehrerer Kameraansichten desselben Kopfs auseinander, um unter herausfordernden Aufnahmebedinungen wie der erwähnten, etwaigen niedrigen Auflösung der Kopfmotive

aber auch unterschiedlicher Lichtverhältnisse innerhalb des Raums und Verdeckungen umgehen zu können. Mit der zusätzlichen Information aus komplementären Beobachtungswinkeln wird untersucht, ob ein Mehrwert in der Hypothesenschätzung erreicht werden kann und welche Hypothesengenauigkeit bei unterschiedlichen Fusionspermutationen der Ansichten zu erwarten ist. Der hierzu aufgenommene Datensatz für die Evaluationen wurde im Laufe dieser Arbeit als Grundlage weiterer Veröffentlichungen und Arbeiten dritter eingesetzt und stellt damit eine erste Referenz für solche Aufnahmebedinungen dar.
Im zweiten Teil der Arbeit wird die Kopfdrehung dazu eingesetzt, auf die individuelle Aufmerksamkeitszuwendung aller Personen in einem multimodal erfassten Raum schließen zu können. Vom erfassten Blickfeld wird hierzu auf mögliche Zielpersonen oder -objekte geschlossen, auf die eine Person ihre Aufmerksamkeit am wahrscheinlichsten richtet. Um den Kontext der Situation dabei nicht unberücksichtigt zu lassen, werden Bewegungsgeschwindigkeit und Sprachaktivität aller Personen als beobachtbare Merkmale des Geschehens in einen probabilistischen Rahmen einbezogen. Mit der Berücksichtung bewegter Aufmerksamkeitsziele, wird hierbei erstmalig im Rahmen dieser Arbeit auf dynamische Szenen eingegangen, in denen weder die vorhandene Anzahl möglicher Aufmerksamkeitsziele noch deren Positionen oder Trajektorien vorab gegeben und bekannt sind. Anhand eines dediziert aufgezeichneten Datensatzes, in dem alle vorhandenen statischen Objekte vermessen und alle beinhalteten Personen durch die automatische Kopfdrehungsverfolgung unterschieden und erfasst wurden, wird die Tauglichkeit des Systementwurfs an realen Besprechungen und Präsentationen in einem Arbeitsumfeld getestet und bestätigt.

Danksagung

Mit dem Abschluss dieser Arbeit blicke ich auf mehrere Jahre spannender Forschung und Mitarbeit an Themenfeldern zurück, die mich ursprünglich überhaupt erst dazu veranlassten Informatik zu studieren. Ich durfte während dieser Zeit Teil einer Wissenschaftsgemeinde sein, deren Arbeiten mit Sicherheit die Art und Weise beeinflussen werden, mit der wir in Zukunft mit Maschinen - und Computern im Speziellen - interagieren und diese in unseren Alltag einbetten werden. An dieser Stelle möchte ich deswegen all denjenigen meinen Dank aussprechen, die mir den Weg hierher geebnet haben und ohne deren Hilfe diese Arbeit in ihrer jetzigen Form nie vorliegen würde.

Mein erster Dank geht dabei an Prof. Dr. Rainer Stiefelhagen, der mit seiner Arbeit nicht nur zündender Funke für meine Dissertation war sondern über seine Rolle als Betreuer und Doktorvater hinaus auch freundschaftlich immer mit Rat und Tat zur Seite stand. Ich kann mich glücklich schätzen noch die Anfänge seiner Bildverarbeitungsgruppe am Lehrstuhl von Prof. Dr. Alex Waibel am damaligen Insitut für Logik, Komplexität und Deduktionssysteme der Universität Karlsruhe (TH) miterlebt zu haben. Mein Dank gilt in diesem Zusammenhang deswegen auch Prof. Waibel für dessen Unterstützung zu Beginn meiner beruflichen Laufbahn und die aufregende, außergewöhnliche und erlebnisreiche Zeit, die er mich am damaligen CHIL-Projekt teilhaben ließ.

Erkenntlich zeigen möchte ich mich daneben auch Herrn Prof. Dr. Jürgen Beyerer dafür, dass er in seiner Rolle als Korreferent sowohl meinen Sinn für das Wesentliche schärfte als auch in unseren konstruktiven Gesprächen meinen eigenen Horizont stets erweiterte und darüber hinaus als Institutsleiter am Fraunhofer IOSB meinen beruflichen Werdegang gerade in der Endphase dieser Arbeit unterstützte.

Mein ganz besonderer Dank gilt aber all jenen Kollegen, mit denen ich während dieser Zeit zusammenarbeiten durfte. In der Hoffnung niemanden zu vergessen, richte mich ich dabei in erster Linie an Kai Nickel, Hazım Kemal Ekenel, Matthias Wölfel, Maria Danninger und Keni Bernardin, mit denen ich die erlebnisreichen Erinnerungen an das CHIL-Projekt teile und die nicht nur mit einer stets offenen Bürotür sondern auch einem offenen Ohr Ansprechpartner für viele richtungsweisende Gespräche und mitunter auch notwendige Perspektivenwechsel waren. Florian van de Camp, Joris Ijsselmuiden und Alexander Schick bin ich für die außergewöhnliche Zeit am Fraunhofer IOSB dankbar, während der sie mich gerade in der Endphase dieser Arbeit begleitet und vor allen Dingen auch unterstützt haben; der gemeinsame Aufbau unseres Labors, vor allem aber die freundschaftliche Zusammenarbeit innerhalb unserer Gruppe und die vielen

spannenden Demonstrationen und Pressearbeiten machen diese Zeit zu etwas wirklich ganz besonderem.
Darüber hinaus danke ich Dr. Gordon Freeman für seine schweigsame Geduld während unserer Zusammenarbeit. Seine Unermüdlichkeit wies mir während dieser Zeit oft die einzuschlagende Richtung.
Ebenso dürfen die studentischen Hilfskräfte nicht unerwähnt bleiben, die mir insbesondere durch ihre Ausdauer beim Annotieren der Daten hilfreich unter die Arme gegriffen und damit das Fundament der Experimente in dieser Arbeit beigesteuert haben - auch ihnen danke ich hierfür ganz besonders.
Zuletzt gebührt aber meiner Familie und meinen Freunden mein Respekt, denn ohne sie wäre ich niemals dahin gekommen wo ich heute bin. Ihre Unterstützung und Begeisterung für meine Arbeit spendeten mir oft Motivation und Energie. Ich bewundere sie für die Ausdauer, die sie dafür zeitweise an den Tag legen mussten.

Karlsruhe, im September 2011 *Michael Voit*

Für meine Mutter

Inhaltsverzeichnis

Notation

Vektoren werden in Form klein geschriebener, fetter lateinischer Buchstaben angegeben. Matrizen in Form groß geschriebener, fetter lateinischer Buchstaben. Skalare werden anhand kleiner und kursiv geschriebener lateinischer oder griechischer Buchstaben bezeichnet. Hochgestellte Indizes beschreiben Elemente einer Menge.

$\mathbb{N}, \mathbb{R}$	Körper der natürlichen bzw. reellen Zahlen
$\mathbb{S}$	Allgemeiner Zustandsraum
$\mathbf{s}$	Zustandsvektor des Kopfs
s_x,s_1	x-Komponente bzw. erste Komponente des Vektors $\mathbf{s}$
$\theta_{pan}, \theta_{tilt}$	Horizontaler bzw. vertikaler Drehwinkel des Kopfs
π^i	Gewichtung des Partikels i
$\mathbf{s}^i$	Durch Partikel i repräsentierter Zustandsvektor
$\mathbf{o}_t^c$	Bild der Kamera c zum Zeitpunkt t
$\mathbf{K}^c$	Kalibrierungsmatrix für Kamera c
$g^c(\cdot)$	Operator der eine Zustandshypothese des Kopfs auf eine Rechteckregion in Kamera c projiziert
$w^{GH}(\cdot,\cdot)$	Bewertungsoperator für Gradientenhistogramme
$m^{GH}(\cdot)$	Merkmalsbeobachtung für Gradientenhistogramme
F^j	Aufmerksamkeitsziel (Person oder Objekt) mit Index j
$\mathscr{F}$	Menge aller Aufmerksamkeitsziele
$\tilde{\theta}$	Mittelwert der zu beobachtenden Kopfdrehungen wenn auf ein Fokusziel oder Voxel geblickt wird
$\hat{\theta}$	Tatsächlicher Blickwinkel zu einem Fokusziel oder Voxel
$V^{j,l}$	Voxel mit Index l des Fokusziels j
$\mathscr{V}$	Voxelmenge
$N_{\mathscr{V}}, \lvert\mathscr{V}\rvert$	Anzahl der Elemente in Menge $\mathscr{V}$

1. Einleitung

Die Blickrichtung eines Menschen deutet uns, wohin er sieht. Indem wir erkennen wohin eine Person ihre Aufmerksamkeit richtet, vervollständigt sich unser Empathievermögen. Wir werden damit in die Lage versetzt begreifen zu können, worauf sich unser Gegenüber bezieht, über was er spricht und auf wen er zeigt. Tätigkeiten werden damit in einen Kontext gesetzt, was dazu beiträgt Absichten zu verdeutlichen. Damit spielt der Blick im gesellschaftlichen Alltag eine herausragende Rolle. In jedem Moment menschlicher Interaktion, gibt die Aufmerksamkeit einen wichtigen Aufschluss über die Umstände der stattfindenden Handlungen. Das gilt sowohl für beteiligte Personen als auch für außen stehende Beobachter.
Das automatische Erkennen der Blickrichtung eines Menschen baut auf dieser Motivation auf. Ein Computersystem das nachvollziehen kann wohin eine Person blickt, kann Aufschluss darüber geben wohin sie ihre Aufmerksamkeit richtet. Das wiederum gibt Kenntnis über den Rahmen, in der sich die beobachtete Person aufhält. Sinnvoll kann das immer dann sein, wenn Menschen mit Maschinen arbeiten oder eingebettete Systeme den menschlichen Alltag beobachten. So müssen humanoide Roboter zum Beispiel erst erkennen lernen, wann sie angesprochen werden oder auf wen sonst sich gegebenenfalls jeweilige Interaktionspartner beziehen, um einem stattfindenden Dialog folgen zu können. Aufmerksame Umgebungen hingegen protokollieren das Geschehen in ihnen und reagieren entsprechend auf Aktionen der beobachteten Benutzer. Das Ziel der Aufmerksamkeit einer Person, stellt damit einen wesentlichen Faktor im Verstehen ihrer Absichten dar.

1.1. Der Visuelle Aufmerksamkeitsfokus

Die Blickrichtung lässt sich erkennen, wenn die Augen der jeweiligen Person beobachtet werden. Das ist allerdings nur unter kontrollierten Bedingungen möglich, weil sichergestellt werden muss, dass die Augen dabei auch sichtbar sind und bleiben. Um mit Kameras die Blickrichtung aufzuzeichnen, müssen diese damit immer in unmittelbarer Nähe zu der zu beobachtenden Person angebracht sein. Unmittelbar bedeutet in diesem Fall, dass sie entweder direkt vor dem Gesicht platziert sind und die Person frontal betrachten oder dass sie in einer Art und Weise an der Person befestigt werden, so dass Kopfdrehungen oder andersartige Bewegungen keinen Einfluss auf die Sicht der Kameras nehmen. Für eine zuverlässige Bestimmung der Blickrichtung, wird deshalb häufig eine Helmkamera eingesetzt, wie sie zum Beispiel in Abbildung 1.1 dargestellt wird. Damit können Probanden normalen Arbeits- oder Alltagshandlungen nachgehen,

Abb. 1.1.: Das EyeLink II-System der Firma SR Research [Res]. Tragbare Kamerasysteme wie das abgebildete eignen sich für die detaillierte Beobachtung der Pupillen in unrestriktiven Handlungen und Umgebungen.

ohne dass die Kamera ihre Sicht auf die Pupillen verliert. Möchte man den Umstand vermeiden beeinträchtigende Sensorik am Körper tragen zu müssen, besteht die Herausforderung darin, dass die Augen gegebenenfalls nur zu seltenen Zeitpunkten sichtbar sind. Statt einer frontalen Aufnahme, ist so nur eine Profil- oder sogar nur Hinterkopfansicht der jeweiligen Person zu beobachten.

1.1.1. Sekundärindikatoren bei offenen Aufmerksamkeitsverschiebungen

Neben *verdeckten* Aufmerksamkeitsverschiebungen, bei denen keine physiologischen Änderungen von außen sichtbar sind, geschehen *offene* Zuwendungen mit einer wahrnehmbaren Veränderung der Blickrichtung, Kopforientierung oder Körperhaltung. Die Kopfdrehung, gibt die Orientierung des Blickfelds vor - jenem Bereich, der bei stillstehendem Kopf von den Augen überblickt werden kann [Sch73]. In Situationen, in denen dieses Primärmerkmal, das Augenpaar, nicht einsehbar ist, erscheint somit der Versuch intuitiv, für hinreichend disjunkt positionierte Gegenstände oder Interaktionspartner im Blickfeld einer Person anhand Sekundärindikatoren wie eben der Blickfeldorientierung selbst, auf das Ziel der momentanen Aufmerksamkeitsverschiebung zu schließen. Dass dies trotz mangelnden Wissens über die tatsächlichen Fixationspunkte der Augen möglich ist, konnte bereits in verwandten Arbeiten, wie zum Beispiel [SFYW99, BO04, FS08], nachgewiesen werden: Die *visuelle Aufmerksamkeitszuwendung* einer Person - das Ziel in ihrem Blickfeld, auf das sie zum Zeitpunkt der Beobachtung blickt - konnte darin erfolgreich allein durch sekundäre Indikatoren erkannt werden.

1.1.2. Von der Kopfdrehung zur visuellen Aufmerksamkeit

Dass die Kopfdrehung ein wichtiges Indiz darstellt der Aufmerksamkeit einer Person zu folgen, lässt sich bereits in der Kunst nachvollziehen. Gee und Cipolla verdeutlichten das zum Beispiel an Gemälden der Renaissance [GC94], in denen die Kopfdrehung der Bildprotagonisten den Blick des Betrachters gezielt auf vereinzelte Elemente im Bild lenken sollte.
Ein analytisches und empirisches Untersuchen, ob auch im gegenwärtigen Arbeitsalltag von der Kopfdrehung auf Interaktionsziele geschlossen werden kann, fand sich aber erst mit den Arbeiten von Stiefelhagen et al. [SFYW99]. Darin wurden in Besprechungen die Kopfdrehungen aller Teilnehmer beobachtet und unter Bezug ihrer jeweiligen Aufmerksamkeitszuwendungen untersucht. Stiefelhagen stellte dabei einen deutlichen Zusammenhang fest, wie Menschen ihre Kopfdrehung dazu einsetzen Gesprächsteilnehmer um sich herum anzuvisieren - eine Beobachtung die auch in Untersuchungen aus der kognitiven Psychologie bestätigt werden konnte [FS08]: Wenn Menschen wiederholt dasselbe Ziel an derselben Position betrachten, dann lassen sich Ballungen feststellen, um die sich die jeweilige Kopfdrehung ausrichtet. Die Beobachtung der Kopfdrehung unterliegt dabei natürlich zusätzlichem Rauschen und das generelle Ausmaß der Ballung hängt darüber hinaus von individuellen Faktoren der entsprechenden Person ab. Aber für ausreichend disjunkt positionierte Gesprächspartner konnte Stiefelhagen damit ein systematisches Abbilden beschreiben, das zu jedem Zeitpunkt allein aufgrund der Kopfdrehung auf den wahrscheinlichsten Interaktionspartner schließen lässt.

1.2. Problemstellung: Dynamische Szenarien

Die Verteilung möglicher Aufmerksamkeitsziele im Blickfeld einer Person und eine entsprechende Modellierung der zu erwartenden Kopfdrehungen, wenn diese betrachtet werden, setzen voraus, dass sich die relative Lage der Ziele zum Beobachter nicht sonderlich ändert.
In der Literatur finden sich zwei grundlegende Ansätze um von der Kopfdrehung auf die Aufmerksamkeitszuwendung zu schließen. Dabei wird entweder eine direkte Abbildung der Kopfdrehwinkel auf ihnen entsprechende Zielzuwendungen eingelernt oder ein probabilistisches Modell der zu erwartenden Kopfdrehungen eingesetzt, wenn ein bestimmtes Ziel anvisiert wird. Bisher wurde dabei jedoch immer außer Acht gelassen, dass die Anordnung der Ziele in der Praxis nur unter kontrollierten Bedingungen als vorgegeben betrachtet werden kann. Sich bewegende Objekte verändern ihre relative Winkelposition zu einem gegebenen Betrachter und eingelernte Modelle werden damit ungültig. Aufgrund der Vielzahl möglicher Zielanordnungen, Trajektorien und Interaktionssituationen können eventuelle Kombinationen im Vorfeld nur spärlich abgedeckt und damit im Nachhinein berücksichtigt werden. Ein Anpassen der Modelle zur Laufzeit wird damit unumgänglich, blieb in der Literatur bisher allerdings unbeachtet. Stattdessen wurden Anwendungsfälle darauf reduziert, dass nur von Situationen ausgegangen

werden musste, die zuvor gemachten Annahmen über die Szene im weiteren Verlauf entsprachen [SBGPO07]. Modelle konnten so optimiert werden, die Disjunktivität zwischen den repräsentierten Aufmerksamkeitszielen zu maximieren und damit über die gesamte, vordefinierte Szene hinweg optimal klassifizieren zu können.

1.2.1. Dynamische versus statische Szenarien

In Situationen mit hoher Dynamik ist weder die relative Lage der Aufmerksamkeitsziele zu einem Betrachter noch deren Anzahl bekannt. Wie erwähnt, kann die Menge möglicher Kombinationen der Objekt- und Personenanordnungen nur spärlich im Vorfeld berücksichtigt werden. Darüber hinaus stellt sich die Frage, ob unter dem Begriff der Szenendynamik nur die Positionierung aller sich darin befindlichen Aufmerksamkeitsziele verstanden werden kann oder ob die darin stattfindenden Aktivitäten ebenfalls Einfluss auf die Aufmerksamkeitszuwendungen der Personen nehmen.
Die wenigen Arbeiten, die sich mit dieser Problemstellung gezielt auseinander setzen, berücksichtigen allein die Bewegung derjenigen Person, deren Aufmerksamkeitszuwendung erkannt werden soll [SBGPO07]. Der Anwendungsfall ist darauf ausgerichtet, dass der Blick zu einem an einer Wand aufgehängten Poster detektiert werden soll, wenn eine Person vor diesem passiert. Mit der Einschränkung sich nicht bewegender Ziele, kann die Menge der zu berücksichtigenden Modelle niedrig gehalten und weiteren Problemen, die in unrestriktiven, dynamischen Szenen aufgeworfen würden, ausgewichen werden: Gäbe es nämlich neben dem fest angeordneten auch weitere, sich bewegende Aufmerksamkeitsziele, so wäre eine Verdeckung des Posters durch sie ebenfalls zu berücksichtigen. Dabei stellt sich dann insbesondere die Frage, ob bisherige Repräsentationen der Objekte und Personen in einer Szene darauf übertragbar sind und falls ja, welche Auswirkungen eventuelle Verdeckungen auf die Entscheidung einer angenäherten Blickrichtung haben können.
Das Beobachten echter Alltagssituationen verlangt nach einer Berücksichtigung nicht-statischer Geschehnisse. Nur wenn allen sich darin befindlichen Personen vollständige Bewegungs- und Handlungsfreiheit zugesprochen wird, können Systeme Alltagsaktivitäten in ihrem Kontext erfassen und darauf reagieren. Infolgedessen stellt das Problem die konsequente Weiterführung des Grundgedankens dar, der im Vorfeld zu der Frage führte, ob aufgrund der Kopfdrehung auf die visuelle Aufmerksamkeitzuwendung einer Person geschlossen werden kann.

1.3. Die visuelle Aufmerksamkeit in aufmerksamen Umgebungen

Um in der Lage zu sein Alltagssituationen beobachten zu können, kommt man nicht umher die sensorgestützte Erfassung einer Szene so in den Hintergrund zu stellen, dass den Protagonisten das Gefühl genommenen werden kann unter ständiger Beobachtung zu stehen. Der Begriff der

aufmerksamen Umgebung bezeichnet in diesem Rahmen Räumlichkeiten, in denen Sensoren und Aktoren als geschlossenes System auftreten und Personen die sich darin befinden den Eindruck vermitteln, sich in einem aufmerksamen und reaktiven Raum aufzuhalten. Das System beobachtet hierzu aus der Ferne die Tätigkeiten jeweiliger Personen und setzt sukzessive mit Hilfe perzeptiver Verarbeitungskomponenten die sensorgestützten Beobachtungen in Wissen über die Situation und Rahmenhandlung zusammen. Mit dem Ziel Absichten und Aktivitäten zu verstehen und darauf zu reagieren, soll eine echte Interaktion mit den agierenden Personen umgesetzt werden können. Wie diese schließlich gestaltet ist, hängt dabei vom jeweiligen Anwendungsfall ab. Die erkannten Tätigkeiten können zum einen protokolliert werden, um im Nachhinein nachvollzogen werden zu können - unter sicherheitsrelevanten Aspekten könnte so eine Überwachung erlaubter Tätigkeiten stattfinden. Zum anderen kann die Perzeption als Ersatz zu bisherigen Eingabemodalitäten dienen, so dass Spracherkennung und Gestik die Basis für einen intuitiven Dialog mit dem System und seinen Eigenschaften bereitstellen. Schlussendlich kann die Umgebung proaktiv unterstützend in eine Handlung eingreifen und dem Benutzer sowohl bei der Steuerung als auch bei seiner Absicht entgegenkommen. Die Möglichkeiten einer solchen, den Mensch verstehenden, Systemsteuerung, erlaubt viele Anwendungsgebiete, die den Rahmen dieser Arbeit sprengen würden. Als solche interessiert jedoch besonders die im Hintergrund angebrachte Sensorik, die die wahrnehmende Ausstattung einer solchen Umgebung darstellt.

1.3.1. Niedrig-aufgelöste Beobachtungen

Die Genauigkeit einer bildverarbeitenden Komponente hängt wesentlich von der Darstellungsqualität des zu verarbeitenden Motivs ab. So ist auch im Fall des automatischen Erkennens der Aufmerksamkeitszuwendung einer Person, die Qualität des erfassten Kopfausschnitts ein ausschlaggebender Faktor für eine hinreichend genaue Erkennung der Blickfeldorientierung. Optimale Aufnahmebedingungen erscheinen deswegen sinnvoll, um die größtmögliche Genauigkeit bei der Winkelschätzung sicherstellen zu können.
Die Herangehensweise, die dabei in der Literatur dominant ist, reduziert zum einen die zu erkennenden Motive auf Vorderkopf- bis maximal Profilansichten. Zum anderen wird in der Regel eine nahestehende Platzierung der jeweiligen Kamera vorausgesetzt, um Gesichtsmerkmale detailreich aufnehmen und die Erkennung der Kopfdrehung damit unterstützen zu können.
Eine derart indiskrete Anbringung widerspricht dem genannten Argument, Menschen unrestriktiv bei natürlichen Interaktionen und Handlungen beobachten zu können. Darüber hinaus kann eine solche Kameraplatzierung zwar unter labortechnischen Bedingungen, nicht aber in Alltagssituationen, sichergestellt werden. Eine unaufdringliche Anbringung der Kameras unterstützt hingegen das Vorhaben, bedeutet aber unter praktischen Gesichtspunkten ein Verlust in der Aufnahmequalität, weil gegebenenfalls größere Abstände zu der zu beobachtenden Person

auftreten. Einerseits ist das der Fall, weil die Anbringung und Ausrichtung der Kamera damit suboptimalen Bedingungen Folge leisten muss, aber auch weil uneingeschränkte Aktivitäten, die Person nicht an einen Handlungsradius vor der Kamera bindet, der unverdeckte und ideale Aufzeichnungen garantieren würde. Objektive mit verstellbarer Brennweite erlauben zwar ein verlustfreies Vergrößern einzelner Bildmotive, in denen der Kopf zu beobachten ist, weisen aber den Nachteil auf, dass ein Ausrichten und Fokussieren auf die entsprechende Bildregion Zeit in Anspruch nimmt und währenddessen ein Fortbewegen der Person nicht ausgeschlossen werden kann. Ferner setzt das die Annahme voraus, dass die fokussierende Kamera eine optimale Sicht auf den zu beobachtenden Kopf bietet und nicht stattdessen eine zweite Kamera an anderer Stelle vorteilhafter wäre, um gegebenenfalls eine Vorderkopfansicht aufnehmen zu können. Mit einer hohen Anzahl vorhandener Kameras ist es zwar möglich diesem Problem entgegenzuwirken, indem mehrere gleichzeitig auf die entsprechende Stelle im Raum fokussieren. Jedoch stellt sich dann die Frage nach einer notwendigen Ressourcenverwaltung, wenn mehrere Personen im Raum gleichzeitig beobachtet und unterstützt werden sollen. So kommt man nicht umhin, auf Ansichten mit fester Brennweite zurückzugreifen. Mit weiten Öffnungswinkeln bieten diese eine großflächige Aufnahme der gesamten Szene und erlauben das gleichzeitige Beobachten verschiedener Regionen. Mit ihrer Sicht müssen jedoch die beschriebenen Eigenschaften in Kauf genommen werden, die Einfluss auf die Darstellungsqualität der zu beobachtenden Motive nehmen.

1.3.2. Mehrkameraumgebungen

Mit dem Einsatz von Kameras mit fester Brennweite und großem Öffnungswinkel können Alltagsszenen gesamtheitlich optimal erfasst werden. Die damit verbundene Herausforderung stellt sich in der zur Verfügung stehenden Auflösung der Kopfausschnitte und dem Umgang mit Kopfdrehungen im gesamten Winkelwertebereich. Die Herausforderung auf Motiven mit niedriger Auflösung Kopfdrehungen zu erkennen, ist in der Literatur bisher noch größtenteils unberücksichtigt. Insbesondere im Zusammenhang mit dem Schließen auf den visuellen Aufmerksamkeitsfokus einer Person existieren bis dato nahezu keinerlei Arbeiten. Dies hängt unter anderem damit zusammen, dass notwendige Datensätze zu Entwicklungs- und Evaluationszwecken nicht vorhanden sind und Forschungsschwerpunkte ferner entweder auf das Erkennen der Kopfdrehung in Überwachungs- und Identifikationanwendungen setzen oder dem Problem der Aufmerksamkeitserkennung nachgehen - nicht aber beidem gemeinsam.
Prinzipiell stellt sich die Frage, welche Kopfdrehwinkelgranularität unter solchen Bedingungen erwartet werden kann. Für das Bestimmen der visuellen Aufmerksamkeit sind feine Winkelschätzungen dabei deutlich von Vorteil. Im Anwendungsfall einer aufmerksamen Umgebung, hängt die Erfassung aber nicht nur von den jeweiligen Kameraeigenschaften ab sondern wird daneben auch vom Verhalten der Personen beeinträchtigt. Wird dabei nur eine Kamera verwen-

det, ist fraglich ob Hinterkopfaufnahmen oder sich fortbewegende Personen nicht den Rahmen des Möglichen sprengen. Der Einsatz mehrerer, unterschiedlich angebrachter Kameraansichten bietet im Vergleich dazu den Vorteil, eine Vorderkopfansicht - oder zumindest Profilansicht - zu jedem Zeitpunkt sicherstellen zu können. Mit der prinzipiellen Möglichkeit in einer aufmerksamen Umgebung Multisensorik einsetzen zu können, würde eine solche Anbringung mehrerer Kameras darüber hinaus den Vorteil bieten, das Innenvolumen im Gesamten erfassen zu können und Verdeckungen daneben gezielt entgegenwirken zu können.

1.4. Ziel der Arbeit

Im Rahmen dieser Arbeit wird ein System entworfen, das die visuelle Aufmerksamkeit von Personen in einer aufmerksamen Umgebung nachvollziehen kann. Als Anwendungsbeispiel dienen dabei Besprechungsszenarien im Arbeitsalltag.
Die beteiligten Personen sollen in ihren Handlungen weder durch an ihnen angebrachte Sensorik noch durch nah platzierte Kameras eingeschränkt werden. Diese werden in hinreichender Entfernung unaufdringlich unterhalb der Raumdecke in den Ecken des Raums installiert. Weil damit keine Sicht auf die Pupillen der beobachteten Personen gewährleistet werden kann, wird als Annäherung der Blickrichtung die Orientierung des jeweiligen Blickfelds genutzt. Hierzu sollen die Kopfdrehungen aller Person visuell erfasst und vollautomatisch erkannt werden. Dabei wird insbesondere der Frage nachgegangen, inwieweit das System invariant gegenüber Auflösungsunterschieden der erfassten Kopfausschnitte ist und ob der Einsatz mehrerer Kameras aus unterschiedlichen Blickwinkeln einen tatsächlichen Mehrwert bietet.
Durch einen probabilistischen Verfahrensansatz wird das System für eine zu beobachtende Person in Echtzeit Schätzungen ausgeben, welches Objekt oder welche andere Person im Blickfeld am wahrscheinlichsten im Fokus ihrer visuellen Aufmerksamkeit ist. Die Beantwortung der Fragen wer mit wem interagiert und auf wen oder was sich jeweilige Besprechungsteilnehmer beziehen ist essentiell für ein System, um den Kontext menschlicher Handlungen und Absichten nachvollziehen zu können. Das Verständnis darüber erscheint somit wichtig, um Mensch-Maschine-Kommunikation intuitiver und zugänglicher gestalten zu können.
Speziell der Anwendungsfall unrestriktiver Dynamik während der Besprechungen in der Literatur, blieb bis zu dieser Arbeit bisher unberücksichtigt. Mit den im Laufe der Arbeit publizierten Auszügen konnte das insofern bestätigt werden, als dass dieser Frage in der Literatur zunehmend nachgegangen wird und Bestandteil weiterer Systemansätze wurde [BHO09]. Die hierfür notwendige Herausforderung Kopfdrehungen vollautomatisch in einer Multisensor-Umgebung wie der beschriebenen erkennen zu können, trug dazu bei mit Hilfe der CLEAR-Evaluationen 2006 und 2007 den im Rahmen dieser Arbeit aufgenommenen und -bereiteten Datensatz als die bis dato alleinige Referenz in dieser Domäne publik zu machen. Mit den genannten Evaluationen fanden erstmalig öffentliche Evaluationen statt, die Kopfdrehung mit Hilfe mehrerer,

kombinierter Kameraansichten niedriger Auflösung zu schätzen, was in Folge zu einem verstärkten Interesse dieses Themas führte.

1.4.1. Beiträge der Arbeit

Neben dem Befassen mit zwei Problemstellungen, die in der Forschung bisher jeweils unberücksichtigt blieben, lassen sich die Beiträge dieser Arbeit auch in der Verfahrensweise und implementierten Algorithmik finden. Im Folgenden sollen diese im Einzelnen hervorgehoben werden.

Echtzeitbasiertes Schätzen der Kopfdrehung in Multisensor-Umgebungen

Im Kontrast zu bisherigen, in der Literatur veröffentlichten Verfahren, werden dabei alle Kameraansichten kombinatorisch auf Entscheidungsebene vereint. Der Drehwinkelschätzung wird die Herausforderung gegenüber gestellt, erstmalig unter solchen Bedingungen Hypothesen in Echtzeit mit einer Winkelgranularität von 1° auszugeben. Wegen der dabei notwendigen Verarbeitung mehrerer Kamerabilder in Echtzeit wird der entworfene Ansatz parallelisiert und für die Unterstützung inzwischen im Desktop-Markt verfügbarer Hardware implementiert.

Unterstützter Winkelwertebereich und Bildauflösung der Kopfregionen

Der beschriebene Anwendungsfall einer aufmerksamen, unaufdringlichen Umgebung setzt voraus, dass keine Beschränkungen in der Bewegung und Kopfdrehung der beobachteten Personen gestellt werden dürfen. Damit dürfen Kopfausschnitte sowohl in weiter Distanz als auch im gesamten Winkelwertebereich ($[0^\circ, 360^\circ)$ horizontal, $[-90^\circ, 90^\circ]$ vertikal) zu den jeweiligen Kameras erscheinen. Aufgrund der Größe der genutzten Umgebung, weisen diese dabei je nach ihrer Distanz eine minimale Auflösung von bis zu 20×30 Pixel auf. Die hierbei gestellten Aufgaben setzen dabei voraus, dass die zur automatischen Schätzung notwendige Kopflokalisierung sowohl auflösungs- als auch rotationsinvariant ist und die Drehwinkelschätzung infolgedessen unterstützt. Der dabei zum Einsatz kommende Partikelfilter kombiniert dabei Gradientenhistogramme und Neuronale Netze. Dabei genügt die Kopfform als Merkmal, insbesondere den genannten Bedingungen, womit deren Gradienten eine stabile Lokalisierung bei hinreichend gegebenem Kontrast ermöglichen. Zum anderen bewiesen Neuronale Netze mit ihrer guten Generalisierungsfähigkeit ihre Stärke bei der Drehwinkelschätzung über den gesamten Winkelwertebereich, während deren Sensibilität gegenüber Lokalisierungsfehlern aber durch den probabilistischen Hintergrund des Partikelfilters und der Robustheit der Kopflokalisierung durch Gradientenhistogramme entkräftet werden konnte.

Datensätze als Referenz öffentlicher Evaluationen

Das Erkennen der Kopfdrehung als auch der Aufmerksamkeitszuwendung sind unter den genannten Bedingungen in der Literatur bisher unberücksichtigt geblieben. Infolgedessen waren bislang weder öffentliche Datensätze verfügbar noch vergleichbare Systemevaluationen, die sich dieser Fragestellung angenommen hatten.
Die Datensätze, die im Rahmen dieser Arbeit aufgenommen und annotiert worden sind, wurden beide veröffentlicht. Jener zum Evaluieren der Kopfdrehung, wurde dabei Bestandteil der internationalen CLEAR-Evaluationen 2006 und 2007 [SBB+07]. Damit konnte nicht nur öffentliches Interesse für diese neuartige Problemstellung geweckt, sondern erstmalig eine vergleichbare Evaluation unter den beschriebenen Herausforderungen durchgeführt werden. Bis dato stellt der Datensatz damit eine Referenz dar.

Bestimmung der visuellen Aufmerksamkeit in dynamischen Szenen

Bisherige Arbeiten, die sich mit dem Nachvollziehen der visuellen Aufmerksamkeit durch die Kopfdrehung auseinandersetzten, beschränkten den Anwendungsfall auf rein statische Positionierungen der Aufmerksamkeitsziele: Objekte und Personen wurden dabei an fest vorgegebenen Stellen platziert, während ferner die Menge der vorhandenen Aufmerksamkeitsziele vorgegeben blieb.
Diese Arbeit setzt sich als erste mit der in der Praxis unvermeidbaren Dynamik in einer Szene auseinander. Hierzu wurden die in der Literatur gängigen, dem Arbeitsalltag entsprechenden Besprechungsszenarien als Anwendungsdomäne zugrundegelegt. Die Menge und Anordnung aller Aufmerksamkeitsziele zu denen eine Person blicken kann, wurde dabei allerdings bewusst inkonstant gehalten. Als Folge stellten sich erstmalig Fragen nach der Sichtbarkeit von Objekten und Personen, die im Kontrast zu bisherigen Ansätzen eine weiterentwickelte Repräsentation und Modellierung der Ziele erforderte. Ferner stellten die Variabilitäten der Zielanordnungen und unberechenbaren Situationskompositionen das System vor die Aufgabe, zur Laufzeit eine Adaption seiner Modelle einzubeziehen, um der Dynamik innerhalb einer beobachteten Szene folgen zu können.

Offene Menge möglicher Aufmerksamkeitsziele

Die in Besprechungsszenarien dominanten Aufmerksamkeitsziele lassen sich auf die Besprechungsteilnehmer, den Besprechungstisch und gegebenenfalls eine Projektionsleinwand für Foliendarstellung reduzieren. Mit der Herausforderung einer robusten Kopfdrehungserkennung und den in den berücksichtigten Situationen vorhandenen Mehrdeutigkeiten, fanden Evaluationen vergleichbarer Ansätze ausschließlich auf einer solch geschlossenen Menge möglicher Zielobjekte statt.

Im Rahmen der Arbeit wurde deshalb ein dedizierter Datensatz aufgezeichnet, dessen Schwerpunkt gezielt auf einen uneingeschränkten Handlungsrahmen aller Besprechungsteilnehmer gelegt wurde. Die aufgezeichneten Geschehnisse beziehen sich deshalb auch auf weitere, der genannten Menge hinausgehender Objekte. Inwieweit dabei der Mensch in der Lage ist, die Aufmerksamkeitszuwendung einer Person unter solchen Bedingungen erkennen zu können, soll deswegen ebenfalls Bestandteil dieser Arbeit sein.

2. Verwandte Arbeiten

Aktuelle Forschungs- und Entwicklungsarbeiten bestätigt den in der Konsumelektronik erkennbaren Trend, dass die Bedienung und Steuerung von Alltagsgeräten für den Menschen intuitiver und natürlicher werden soll. An klassischen Desktop-Arbeitsplätzen dominieren zwar noch Tastatur und Maus als die wesentlichen Eingabemodalitäten, aber Smartphones weisen bereits viele Vorteile auf, die sich mit Fingergesten erreichen lassen um, eine intuitive Bedienung der Benutzeroberflächen anzubieten. Darüber hinaus darf nicht vergessen werden, wie selbstverständlich kontextbezogene Dienste bereits Werkzeuge unseres Alltags sind, um zum Beispiel in Suchanfragen auf mögliche Bedürfnisse des Benutzers eingehen zu können.
Das Nachvollziehen der Absicht einer Person ist ein wesentlicher Bestandteil zur Feststellung, in welchem Kontext sie agiert. Das gilt gleichermaßen für menschliche Beobachter wie auch für Systeme, die ihr Wissensmodell über die jeweilige Person vervollständigen und dementsprechend auf sie eingehen sollen. Deswegen scheint es essentiell, sich sowohl mit den soziologischen Gesichtspunkten menschlicher Interaktion, als auch den naturwissenschaftlichen Modellen und Annäherungen des damit verbundenen Kontexts auseinanderzusetzen.
In diesem Kapitel wird hierzu der Frage nachgegangen, welche in der Literatur bestehenden Arbeiten, den im hiesigen Problemstellungen nahekommen. Dabei wird zunächst die Aufgabe eruiert, Kopfdrehungen in Multisensor-Umgebungen zu erkennen. Im zweiten Abschnitt wird anschließend eine Übersicht gegeben, mit welchen Ansätzen bisher mit Hilfe erkannter Kopfdrehungen auf Aufmerksamkeitszuwendungen geschlossen wird und welcher Nutzen in der derzeitigen Forschung daraus gewonnen werden kann.

2.1. Kopfdrehungserkennung in niedrig aufgelösten Multisensor-Daten

Kopfdrehungen geben nicht nur Aufschluss über die Orientierung des Blickfelds, sie sind auch mit der Fragestellung nach der Identität einer Person verknüpft. Neben Beleuchtungsunterschieden stellen Kopfrotationen eine der bedeutsamsten Ursachen für Varianzen in der Erscheinung eines Gesichts dar. Besonders bei der Problemstellung menschliche Gesichter in Bildern oder Videoströmen zu detektieren, stellt die Kopfdrehung damit insbesondere bei jenen Ansätzen eine wesentliche Herausforderung dar, die einer robusten Detektion gewisser Gesichts- oder Kopfmerkmale ausgesetzt sind.
Einer der ersten veröffentlichten Berichte in dem die Kopfdrehung als zu berücksichtigende Varianz beschrieben wird, findet sich in W. Bledsoes Arbeit über semiautomatische Identifikation

von Gesichtern aus dem Jahr 1964 [Ble64]. Bledsoes Vorgehen bestand daraus, anhand der relativen Positionierung verschiedener Gesichtsmerkmale einen Bezug zur Identität der jeweiligen Person herzustellen. Durch verschiedene Kopfdrehungen in seinen Aufnahmen beobachtete er allerdings nicht-lineare Verschiebungen der von ihm ausgewählten Merkmale, die er mit Hilfe einer eigens dafür entworfenen Normalisierung eindämmen wollte.

Die ersten vollautomatischen Klassifikationen von Kopfdrehungen beginnen im Kontrast dazu erst verhältnismäßig spät mit den Arbeiten von Beymer et al. ab 1993 [Bey94]. Auch bei Beymer liegt die Motivation in der visuellen Identifikation von Gesichtern. Das Erkennen der Kopforientierung sollte helfen, Repräsentanten aus derselben Drehwinkelklasse bei der Detektion auswählen zu können. Die singulären Drehwinkel hierfür werden dabei geometrisch bestimmt, indem Merkmale wie die Position der Augen, Nase und Mundwinkel vollautomatisch detektiert werden und deren Lage jeweils zueinander zur geometrischen Berechnung der Kopfdrehwinkel führt.

Dass Kopfdrehungen auch bereits Rückschlüsse auf die Blickrichtung eines Menschen erlauben, beschreiben als erste Gee und Cipolla 1994 in [GC94]. Darin nehmen sie Bezug auf Beispiele von Künstlern wie Michelangelo und Botticelli und hoben gestalterisch eingesetzte Kopfdrehungen als wichtiges Element hervor, den Blick eines Betrachters auf ausgesuchte Bereiche im Bild zu lenken. Für einen Versuch dies automatisch nachvollziehen zu können, schließen die Autoren dabei von einem sehr schmalen Blickfeld mit einer Winkelbreite von $\pm 10^{\circ}$ auf jene Bereiche im Gemälde, die innerhalb dessen liegen und so der Aufmerksamkeitszuwendung der gemalten Person entsprechen.

Inzwischen umfasst das Erkennen menschlicher Kopfdrehungen Anwendungsbereiche, die sich von rein frontal aufgenommenen Ansichten, wie sie in der Regel zur Identifikation oder Verifikation von Personen herangezogen werden, maßgeblich unterscheiden. Neben der nun bestehenden Vielzahl verschiedener Ansätze, kommt damit auch eine große Varianz bezüglich der zugrunde gelegten Aufnahmebedingungen und der dabei zu erwartenden Winkelgranularität solcher Systeme hinzu, die es erschwert, die unterschiedlichen Verfahren in einheitliche Kategorien einteilen und vergleichen zu können. Insbesondere Arbeiten, die sich mit Multisensor-Umgebungen auseinander setzen, verwenden in der Regel einzelkamerabasierte Methodiken, die auf den speziellen Anwendungsfall übertragen und zum Beispiel eine vorangehende Kameraselektion erwarten, auf der schließlich die Drehwinkel hypothetisiert werden. Im Folgenden soll der Überblick verwandter Arbeiten daher zunächst eine Zusammenfassung prinzipieller Methodiken einzelkamerabasierter Verfahren darstellen. Im Anschluss werden diese beschriebenen Ansätze dann in einen Zusammenhang mit den in dieser Arbeit zu berücksichtigenden Problemstellungen gebracht und die wenigen veröffentlichten Arbeiten zu diesem Themengebiet näher erläutert.

2.1.1. Kategorisierung prinzipieller Methodiken zur automatischen Schätzung der Kopfdrehung

Ein bereits sehr umfassender Vergleich bisheriger Verfahren findet sich in einer Veröffentlichung von Murphy-Chutorian et al. aus dem Jahr 2009 [MCT09]. Die Autoren schlagen darin eine Kategorisierung der prinzipiellen Methodiken vor, womit fast ausnahmslos ein Überblick über einzelkamerabasierte Methoden, sowie mögliche Übertragungen auf weitere Anwendungsbereiche gegeben wird. Murphy-Chutorian et al. weisen die Ansätze dabei den Kategorien Template-Matching, Detektoren, Regressoren, Unterraumabbildungen, Graphstrukturen, geometrische und schlußendlich trackingbasierte Verfahren zu. Besonders im Hinblick auf verfügbare Datensätze und damit verbundenen vergleichenden Evaluationen geben die Autoren einen umfangreichen Überblick über den aktuellen Stand der Forschung und Entwicklung, der an dieser Stelle sowohl den Rahmen der Übersicht sprengen würde als auch nicht substantiell mit den Herausforderungen der hiesigen Arbeit verwandt wäre. Die im Folgenden aufgeführten Techniken wurden deshalb ausschließlich wegen ihres Übertragens auf dieser Arbeit verwandte Anwendungsbereiche ausgesucht. Für einen darüber hinausgehenden Vergleich weiterer Ansätze, sei der interessierte Leser allerdings auf weiterführende Literatur verwiesen.

Geometrische Herleitung der Drehwinkel

Geometrische Ansätze basieren auf der Motivation, dass in der Wahrnehmungstheorie durch Experimente nachgewiesen werden konnte, dass das menschliche Erkennen von Kopfdrehungen maßgeblich auf der Ausrichtung der Nasenspitze und Symmetrieeigenschaften innerhalb des Gesichts aufbaut [WWLC00]. Detektiert und beobachtet man verschiedene Merkmale in einem menschlichen Gesicht und deren relative Lage zueinander, kann man auf die Drehwinkel des Kopfs Rückschlüsse ziehen. Bei einer Rotation findet eine nicht-lineare Verschiebung dieser Merkmale statt. Geometrische Ansätze verfolgen diese Tatsache und versuchen trigonometrisch die Drehwinkel zu berechnen [HZ98, HYD96, WS05, NP00, GHC04]. Prinzipiell erlauben geometrische Verfahren die feingranularste Auflösung der Winkelerkennung, setzt man voraus, dass die Detektion der Merkmale hierfür ausreichend robust geschieht. Eine hochauflösende Aufnahme des beobachteten Kopfes ist dabei Voraussetzung. Des Weiteren beeinflussen Verdeckungen die Merkmalsfindung deutlich. Dies geschieht dabei nicht nur bei tatsächlichen Verdeckungen durch Objekte im Vordergrund, sondern auch bereits bei ausgeprägten Kopfdrehungen, die durch das Relief des Gesichts Merkmale verdecken lassen [Bro01].
Viele Verfahren bauen auf dem Drei-Punkt-Modell von Huttenlocher und Ullman [HU90] auf, das Lage und Orientierung dreidimensionaler, rigider Objekte in zweidimensionalen Darstellungen durch nur drei unterschiedliche Punktkorrespondenzen von Bild- und Modellpunkten ermittelt. Im Beispiel von [HZ98] und [NP00] werden hierzu Augen und Mund bzw. linkes oder rechtes Ohr, Braue und Mundwinkel als Merkmalspunkte herangezogen. Horprasert et

al. [HYD96] nutzen im Gegensatz dazu Augen und Nasenspitze. Und mit der relativen Lage des detektierten Mundes zum Gesichtsmittelpunkt bestimmen Wenzel und Schiffmann [WS05] die Kopfdrehung auf Kamerabildern, bei denen Augenregionen durch Brillen verdeckt werden, um die geometrische Granularität auch dann zu erhalten, wenn die Personen Kamerahelme, -brillen oder Displayhelme für *Virtual Reality*-Anwendungen tragen.
Die Robustheit geometrischer Ansätze hängt in hohem Maße von der Lokalisierungsgenauigkeit der zugrundegelegten Merkmalspunkte im Kamerabild und Gesicht ab. Bei nicht-hybriden Ansätzen geht insbesondere kein explizites Erkennen einer zwar groben, aber zumindest hilfreichen Kopfdrehung voraus, die bei der eigentlichen Detektion der Merkmale hilfreich sein könnte. Stattdessen werden Kandidaten für Gesichtsregionen in der Regel mit Hilfe von rotationsinvarianten Merkmalen reduziert und in den reduzierten Bildbereichen gezielt nach Merkmalspunkten gesucht. Ein populäres Vorgehen ist dabei durch Hautfarbsegmentierung Gesichtshypothesen vom Hintergrund zu trennen [WS05, NP00, GHC04].

Templatebasierte Klassifikation und Interpolation

Im Gegensatz zu geometrischen Ansätzen, nutzen template-basierte Verfahren wie zum Beispiel [NF96, Bey94, BO04, BO05, PA00, PB98] die gesamte Darstellung des Gesichts beziehungsweise des Kopfs, indem sie das Bild mit hinterlegten Beispielen vergleichen und schließlich diejenigen Drehwinkel zuweisen, die am besten den jeweiligen Winkelrepräsentanten entsprechen.
Im einfachsten Fall wird die gefundene Kopfregion dabei pixelweise mit den hinterlegten Beispielbildern verglichen. Niyogi und Freeman verfolgen diesen Ansatz in [NF96], worin sie die Lokalisierung des Kopfes durch Minimieren des Mittleren Quadratischen Fehlers über vorliegende Templatebeispiele umsetzen. Pappu et al. gehen ähnlich vor, erstellen die Beispiele in [PB98] aber für jede Person im Einsatz initial neu, indem sie einmalig eine frontale Aufnahme des Gesichts voraussetzen, um durch Warping angenäherte Darstellungen unterschiedlicher Drehungen zu berechnen. Nachfolgende Beobachtungen werden dann mit den gerenderten Beispielen verglichen und das nächstliegendste mit seinen zugehörigen Drehwinkeln als Ergebnis ausgegeben.
Hinter Templatevergleichen steckt die vereinfachende Annahme, dass paarweise Ähnlichkeiten im Bildraum maßgeblich durch ähnliche Kopfdrehungen hervorgerufen werden. Weitere Varianzen, die zum Beispiel durch unterschiedliche Identitäten hervorgerufen werden, bleiben dabei unberücksichtigt [MCT09]. Um Templates invariant gegenüber Beleuchtungs- und Identitätsschwankungen ausfallen zu lassen, transformieren manche Verfahren das Eingabebild zunächst. So wird versucht, bedeutsame Merkmale herauszuarbeiten, von denen erwartet wird, dass sie in ihrer Erscheinung zu einer wesentlichen Unterscheidung der Kopfdrehungen beitragen. Ba et al. falten beispielsweise in [BO04, BO05] die extrahierte Kopfregion zunächst mit

Gauss- und Gaborfiltern und ballen die durch Konkatenation erhaltenen Merkmalsvektoren mit Hilfe des k-Mean-Verfahrens in Klassenregionen. Park et al. dagegen unterteilen in [PA00] das Kopfbild in 3×4 lokale Zellen und konkatenieren die für jede Zelle entsprechenden Momente zu einem zwölfdimensionalen Vektor der schließlich für den Templatevergleich benutzt wird.
Verfahren, die aufbauend auf der Darstellung des Kopfes oder lokaler Teilbereiche aufsetzen, bieten den Vorteil, dass sie prinzipiell auch auf Kamerabilder übertragen werden können, in denen der Kopfausschnitt mit nur niedriger Auflösung beziehungsweise kleiner Größe dargestellt ist. Das resultiert aus der einfachen Tatsache, dass Gesichter in Aufnahmen mit niedriger Auflösung nur unscharf abgebildet werden und nur selten die notwendigen Details beinhalten, die ein robustes Lokalisieren von bestimmten Merkmalspunkte möglich machen. Ansichtsbasierte Methoden nutzen dagegen die Erscheinung des modellierten Bildbereichs als Ganzes, statt nur das zu erkennende Merkmal. Doch auch sie setzen eine hinreichend robuste und konsistente Lokalisierung der (Teil-)Regionen voraus, damit Hintergrundrauschen minimiert und der zu vergleichende Merkmalsvektor größtenteils frei von weiteren Einflüssen außer der zu erkennenden Drehung bleibt. In probabilistischen Verfahren wie [BO04, BO05] und auch [NF96] wird dem zu einem gewissem Grad entgegengewirkt, indem Rotation und Lokalisierung in Form einer gemeinsamen Zustandshypothese bewertet werden und ein Minimieren des Fehlers bzw. der Distanz zu den jeweiligen Klassenrepräsentanten, zum Beispiel durch einen Partikelfilter, umgesetzt wird.

Detektorreihen

Detektoren entscheiden für ein gegebenes Motiv mit einer binären Ja-/Nein-Entscheidung, ob das Dargestellte dem gelernten Modell entspricht. Das Motiv wird hierzu aus dem Bild extrahiert und dem Detektor zur Entscheidung vorgelegt. Durch deren vereinfachte Ausgabe liegt ein wesentlicher Vorteil von Detektoren in ihrer Geschwindigkeit und prinzipiellen Robustheit durch die starke Einschränkung des auszugebenden Wertebereichs. Man setzt sie deswegen häufig auf dem gesamten Kamerabild an, um möglichst schnell nach Motivkandidaten des gesuchten Zielobjekts zu suchen. Da mit unterschiedlicher Beobachtungsdistanz des Objekts allerdings dessen Größe stark variieren kann, muss das in der Suche berücksichtigt werden. Häufig wird dazu das Eingabebild in seiner Größe auf- und abskaliert, so dass eine Bildpyramide entsteht, deren einzelnen Ebenen dem Detektor jeweils als Kamerabild vorgelegt werden. Ohne beim Einlernen Beispiele mit unterschiedlicher Größe einbinden zu müssen, kann der Detektor so durch Suchen auf unterschiedlichen Skalierungsebenen mit verschiedenen Größenverhältnissen umgehen.
Durch ihren Wertebereich eignen sich Detektoren ausschließlich für Entscheidungsfindungen bezüglich der Zugehörigkeit zu einer einzelnen Klasse an Motiven. Für Kopfdrehungen muss deswegen zunächst der Winkelwertebereich diskretisiert und in jeweilig zu detektierende Klas-

sen eingeteilt werden. Für jede Winkelklasse wird dann schließlich ein eigener Detektor angewandt. Ein Zusammenführen der jeweiligen Detektorergebnisse in einem einzelnen Bild, sogar gegebenenfalls mehrerer positiver Detektionen unterschiedlicher Klassen auf demselben Motiv, ist darüber hinausgehend Aufgabe des Gesamtsystems.

Als erster Detektor für Kopfdrehungen kann die Arbeit von Kanade und Schneidermann aus dem Jahr 2000 angesehen werden [SK00]. Mit zwei Detektoren wird darin zwischen Frontal- und Profilgesichtern klassifiziert, wobei Profilansichten sowohl in ihrer eigentlichen Erscheinung als auch gespiegelt klassifiziert werden, um mit demselben Detektor zwischen rechts- und linksseitigen Profilansichten unterscheiden zu können. Die Detektoren nutzen dazu Histogramme, in die die Verteilung von Wavelet-Koeffizienten, sowie deren Position im Motiv, gelernt wurden. Mit einer großen Menge solcher Histogramme wird versucht die Varianzen der verschiedenen aber zugehörigen Motive pro Klasse zu erfassen. Zur Zeit der Veröffentlichung brauchte der Algorithmus allerdings auf einem 320×240 Pixel großem Kamerabild und lediglich vier Skalierungsstufen knapp eine Minute Berechnungszeit, was für mehrere Drehwinkelklassen auch heute noch ein ungünstiges Ausmaß an Rechenzeit bedeutet.

Anhand dieses Beispiels wird deutlich, dass die genannten Vorteile der Detektoren durch zwei wesentliche Eigenschaften relativiert werden: zum einen bestehen Einschränkungen in der notwendigen Diskretisierung der Drehwinkel, die sowohl groß genug sein muss um Varianzen in der eigenen Klasse genügend eingrenzen zu können aber klassenübergreifende Unterschiede deutlich trennen können muss. Zum anderen können hohe Laufzeiten pro Detektorsuche den praktischen Einsatzzweck benachteiligen, da sich dieser natürlich linear mit der Anzahl der Detektoren erhöht, was so eher für eine nur grobe Diskretisierung des Winkelwertebereichs spricht.

Im Jahr 1998 beschrieben Viola und Jones einen Gesichtsdetektor für frontale Ansichten, der durch seine schnell zu berechnenden Merkmale Detektionen in Echtzeit ermöglicht [VJ01]. Hierzu nutzen das System Sammlungen von Differenzwerten definierter Bildregionen im jeweiligen Motiv. Die Regionen sind dabei meist rechteckig und für jeden rechteckigen Bereich werden die umfassten Pixelintensitäten akkumuliert und von einem zweiten entsprechend umschriebenen Bildbereich subtrahiert. Die Idee ist es, für generelle Objektdetektion lediglich die jeweils hinreichende Menge dem Motiv entsprechender Differenzpaare zu finden und dem Detektor einzulernen. Die Geschwindigkeit beim Berechnen der Differenzwerte zieht ihren Vorteil dabei insbesondere durch Nutzung eines Integralbildes, bei dem jedem Pixel die Summe seiner Vorgänger aufaddiert wird. Damit können Pixelsummen über rechteckige Bereiche durch nur wenige Zugriffe auf das Integralbild bestimmt werden, was zu einer deutlichen Beschleunigung der Merkmalsberechnung führt. Des weiteren besteht der Detektor aus kaskadierten Klassifikatoren, deren Komplexität mit jeder Stufe erhöht wird: Lehnt der Klassifikator in einer Stufe ein vorgelegtes Motiv ab, wird die Kaskade sofort unterbrochen und das Motiv verworfen. Ansons-

ten wird es einem nachfolgenden, komplexeren Klassifikator für eine rechenintensivere Entscheidung durchgereicht. Hintergrundmotive, die den eingelernten Beispielen unähnlich sind, können damit bereits in den ersten Stufen abgelehnt werden und die Iteration über ein gesamtes Kamerabild so deutlich beschleunigen. Nur entsprechende Beispiele werden den komplexeren Kaskadenstufen unterzogen. Der dabei entstehende Geschwindigkeitsvorteil erlaubt es statt nur eines einzelnen Detektors gleich mehrere einsetzen zu können und eine eventuell doch feingranularere Diskretisierung durch die damit eingesparte Rechenzeit durchführen zu können. Zhang et al. nutzten diesen Ansatz in [ZZLZ02] und erklärten ihr System als das erste damals echtzeitfähige, detektorbasierte zur Lokalisierung und Winkelklassifikation gedrehter Köpfe. Dazu kaskadieren sie drei Detektorschichten, indem sie das Bild mit einem ersten Detektor absuchen, der allgemein auf Köpfe innerhalb des Winkelwertebereichs $[-90°, +90°]$ reagiert. Nicht-verworfene Kandidaten werden schließlich erst drei weiteren Detektoren unterzogen, die ihrerseits den anfänglichen Winkelwertebereich äquidistant aufteilen. Die hier positiv klassifizierten Hypothesen werden zuletzt einer dritten Ebene an Detektoren durchgereicht, in der neun weitere Detektoren wiederum feiner unterteilen und die endgültige Klassifikation darstellen.
Detektoren haben inzwischen den Vorteil, dass sie bei Wahl geeigneter Merkmale nicht nur schnell sind, sondern auch die Lokalisierung des Kopfausschnitts implizit mitliefern. Im Gegensatz zu den bisherigen Ansätzen, muss keine konsistente Kopfdetektion vorab durchgeführt werden, die anschließend zur Drehwinkelschätzung genutzt wird. Detektoren suchen gezielt nach der eingelernten Klasse entsprechenden Motive und geben neben der Klassifikation auch den jeweiligen Bildbereich zurück.
Im Fall einer doch zu groben Winkeldiskretisierung, kann die detektierte Kopfregion immer noch einer feineren Klassifikation oder sogar Regression unterzogen werden. Die vorab lokalisierte Bildregion des Kopfes reduziert dabei den Lokalisierungsaufwand eines nachgestellten, die Drehwinkel verfeinernden Systems. So veröffentlichten Murphy-Chutorian et al. 2007 einen Ansatz, der die von Viola und Jones beschriebenen Detektorkaskaden zur Lokalisierung des Fahrerkopfes in Automobilen nutzt [MCDT07]. Die Autoren nehmen hierzu jeweils einen Detektor für das Lokalisieren linksgedrehter Profilmotive, frontaler Aufnahmen und rechtsgedrehter Köpfe. Die dabei grobe Winkelschätzung verfeinern sie anschließend, indem mit Hilfe einer Support Vektor Maschine die beobachteten Drehwinkel des Kopfs regressiert werden.

Neuronale Netze und Support Vektor Maschinen als Klassifikatoren oder Nichtlineare Regressoren

Mit dem Begriff *Neuronale Netze* werden Familien unterschiedlicher Topologien zusammengefasst, die alle grundlegend auf der Vernetzung sogenannter *Neuronen* aufbauen. Neuronen sind atomare Einheiten, die miteinander verbunden einen Netzverbund darstellen. Dabei bekommen dedizierte Eingangsneuronen einen Merkmalsvektor des zu bearbeitenden Motivs ange-

legt, reagieren darauf entsprechend einer ihnen zugewiesenen Funktion und leiten ihre Ausgabe an nachgeschaltete Neuronen weiter. Diese reagieren schließlich auf dieselbe Art und Weise, womit die Verarbeitung sukzessive durch das Netz gereicht wird. Zwei populäre Methoden um Kopfdrehungen mit Hilfe Neuronaler Netze zu schätzen, sind dabei sogenannte *Mehrschichtige Perzeptronen* und *Selbstorganisierende Karten*. Erstgenannte Topologie beschreibt die Aneinanderreihung vereinzelter Neuronenschichten: Eine Eingabeschicht empfängt den Merkmalsvektor, verarbeitet diesen und leitet durch Verbindungen in die nachfolgende Schicht die Ausgaben aller Neuronen weiter. Die der zweiten Schicht empfangen durch die Verbindungen die vorherigen Ausgaben, fassen diese zusammen, reagieren entsprechend und geben wiederum ihre Ausgaben an eventuell wiederum nachgeschaltete Schichten fort. Dieser Prozess wird fortgesetzt, bis die Ausgabeschicht erreicht ist, deren Neuronenreaktionen die jeweilige Assoziationen des Netzes auf den angelegten Merkmalsvektor darstellen.
Selbstorganisierende Karten empfangen hingegen den Merkmalsvektor in der Eingabeschicht, welche ihre Reaktionen anschließend an eine zweite, in der Regel zugleich letzten Schicht weitergibt. Diese *Karte* soll dabei eine Ähnlichkeitsabbildung beschreiben und den höherdimensionalen Vektorraum des Eingabemerkmals in eine planare Darstellung zueinander gehöriger Muster transformieren. Wird ein Signal an das Netz angelegt, werden nur diejenigen Gebiete der Karte angeregt, für die gilt, dass Merkmale aus diesen Bereichen hinreichend Ähnlichkeiten zum Eingangssignal aufweisen. Die hauptsächlichen Unterschiede zwischen den beiden Topologien, bestehen in der verschiedenartigen Verarbeitung der Merkmale und in der Art und Weise diese einzulernen. Mehrschichtige Perzeptronen müssen im Gegensatz zu Kartentopologien überwacht eingelernt werden. Dazu werden für gegebene Merkmalsvektoren, die ihnen zugehörigen Zielausgaben des Netzes sowie notwendige Verbindungseigenschaften parametrisiert. Selbstorganisierende Karten lernen hingegen unüberwacht und ballen eine gegebene Menge von Trainingsbeispielen automatisch in entsprechende Kartenassoziationen.

Die Einsatzbereiche von Neuronalen Netzen sind sehr vielfältig. Durch geeignete Wahl hinreichend vieler Neuronen und Netztopologien, können beliebige Funktionen approximiert werden. Neuronale Netze können damit sowohl zur Klassifikation von Kopfausschnitten in vorgegebene Winkelklassen als auch zur Regression einer Drehfunktion über die angelegten Motive eingesetzt werden. Maßgebliche Arbeiten in diesem Bereich finden sich in [BAMPS98, SYW98, KBS00, Sti04, SNS04], die im Folgenden kurz beschrieben werden sollen.

Bruske et al. [BAMPS98] verarbeiten zum Beispiel durch Farbsegmentierung gefundene Gesichtsregionen mit Gabor-Wavelets. Die erhaltenen Koeffizienten der Gabor-Darstellung werden dabei als Merkmalsvektor interpretiert und dieser schließlich durch eine Selbstorganisierende Karte auf Kopfdrehungen abgebildet. Durch die unterschiedliche Anregung jeweiliger Kartengebiete interpolieren die Autoren eine regressierende Schätzung der Kopfdrehung. Die Experimente der Arbeit beschränken sich allerdings nur auf Aufnahmen maschinell gedrehter

Puppenköpfe. Das gibt wenig Aufschluss hinsichtlich der Übertragbarkeit auf reale Motive. Auch werden nur horizontale Drehungen im Bereich $\pm 75^\circ$ und vertikale im Bereich $\pm 30^\circ$ erkannt. Selbst frontale Aufnahmen sind damit schon soweit eingeschränkt, dass Profilansichten die Möglichkeiten des Systems überanspruchen.

Stiefelhagens Arbeiten, wie zum Beispiel [SYW98], verwenden ebenfalls Gesichtsmotive, die im Kamerabild durch Hautfarbsegmentierung lokalisiert wurden. Das dabei zum Einsatz kommende mehrschichtige Perzeptron besteht aus einer Eingabeschicht für das Motivbild, einer weiteren Schicht mit 10 Neuronen und einer Ausgabeschicht, die das Grauwertbild des Gesichts in eine von vier möglichen Drehklassen (links, rechts, frontal, nach unten geneigt) zuordnet. Nachfolgende Arbeiten evaluieren ferner weitere Netztopologien mit unterschiedlicher Anzahl zwischengeschalteter Neuronen in der zweiten Schicht. Ebenso berücksichtigen sie sowohl klassifizierende und regressive Ausgaben. Zuletzt untersuchte Stiefelhagen des weiteren eine Verfeinerung der Eingaben durch Konkatenierung des Grauwertbilds an horizontale und vertikale Gradienten einer Sobelfaltung [Sti04].

Seemann et al. veröffentlichten im Jahr 2004 einen ähnlichen Ansatz, der allerdings die von einer Stereokamera stammenden Tiefen- bzw. Disparitäteninformation als zusätzliches Merkmal zum Grauwertbild des Gesichtsmotivs in das Neuronale Netz einspeist [SNS04]. Für die horizontale, vertikale und neigende Kopfdrehung wird hierzu jeweils ein Neuronales Netz mit einer regressiven Ausgabe des Drehwinkels eingelernt.

Support Vektor Maschinen eignen sich wie auch Neuronale Netze sowohl zum Regressieren einer Funktion als auch zum Klassifizieren. Support Vektor Maschinen (engl. Support Vector Machines, im Folgenden SVM genannt) bilden implizit Merkmalsvektoren in einen höherdimensionalen Vektorraum ab, in welchem sie die Vektoren linear separieren können. Das Parametrisieren der Separierungsebene erfolgt im Trainingsschritt, in welchem die Stützvektoren - sogenannter *Support Vektoren* - aus dem Datensatz an Trainingsbeispielen so ausgewählt werden, dass diese durch eine dazwischen eingebrachte Trennebene optimal voneinander separiert werden. Prinzipielles Vorgehen ähnelt dabei dem Neuronaler Netze: eine Kopf- oder Gesichtsregion wird lokalisiert und in einen Merkmalsvektor transformiert. Dieser wird der Support Vektor Maschine eingespeist, welche schließlich die eigentliche Klassifikation unternimmt. Beispiele darauf aufbauender Systeme finden sich insbesondere in [NG99, LGSL00] und [MCT08b].

Die Arbeit von Li et al. aus dem Jahr 2000 [LGSL00] setzt ihren Schwerpunkt auf der Detektion von Gesichtern mit unterschiedlichen Drehungen, ähnlich zu detektorbasierten Ansätzen die in Abschnitt 2.1.1 vorgestellt wurde. Das System reduziert seine Suche zunächst auf interessante Bildregionen, die durch Hautfarb- und Vordergrundsegmentierung extrahiert wurden. Die lokalisierten Gesichtshypothesen werden daraufhin einem SVM-Regressor eingespeist, um eine grobe aber hinreichend genaue Schätzung der Kopfdrehung in dem extrahierten Motiv zu

erhalten. Dabei setzen die Autoren jeweils einen Regressor für das Erkennen der horizontalen Drehung und einen für die vertikale ein. Um die Dimensionalität des Motivs zu reduzieren, wird es zuvor mit Hilfe einer Hauptkomponentenanalyse transformiert und auf seine dominanten Merkmale reduziert bevor es zur Regression vorgelegt wird. Die anschließende Schätzung nutzen die Autoren schließlich, um mit einem speziell auf die entsprechende Drehwinkelklasse eingelernten Gesichtsdetektor das extrahierte Gesichtsmotiv nochmalig als Gesicht zu bekräftigen und bestätigen.
Ng und Gong beschreiben in [NG99] ebenfalls einen Detektor auf Basis von Support Vektor Maschinen. In ihrer Arbeit implementierten sie einen Ansatz, der insgesamt fünf Detektoren für den horizontalen Drehbereich $[0°, 180°]$ vorsieht. Jeder Detektor besteht aus einer Support Vektor Maschine, die auf hauptkomponententransformierten Motivbildern hinsichtlich seiner entsprechenden Drehklasse klassifiziert. Die Detektoren werden ähnlich zu den vorgestellten Detektorverfahren in einer Bildpyramide über das gesamte Kamerabild iteriert. In Kombination mit einem Trackingverfahren wird die Suche in darauf folgenden Kamerabildern zwar auf einen stark reduzierten Bereich eingeschränkt, das exzessive Suchverfahren in Kombination mit den aufwendig zu berechnenden SVM-Klassifikationen ermöglicht keine Echtzeitfähigkeit per se.
Die Arbeiten von Murphy-Chutorian aus den Jahren 2007 und 2008 setzen ebenfalls auf Support Vektor Maschinen [MCDT07, MCT08b]. Die Autoren konzentrieren sich darin auf das Erkennen der Kopfdrehung von Autofahrern und erhalten die Kameraaufnahmen von vor dem Fahrer im Cockpit angebrachten Kameras. Wie bereits im Abschnitt über Detektorreihen, setzen die Autoren in [MCDT07] dazu klassische Gesichtsdetektoren nach Viola und Jones ein, um in Echtzeit die jeweilige Gesichtsregion des Fahrers zu lokalisieren. Nachdem diese in ein Gradientenhistogramm umgerechnet wurde, wird es als Merkmal in eine Support Vektor Maschine zur Regression der Drehwinkel eingespeist. Darauf aufsetzend kommt in der nachfolgenden, hybriden Arbeit [MCT08b] ein paralleles Trackingverfahren zum Einsatz, das neben dem reinen Schätzen noch hypothetisierte Kopfdrehungen dreidimensional rendert und texturell mit dem im Kamerabild lokalisierten Gesichtsmotiv vergleicht. Sinkt die Konfidenz der getrackten Hypothese unter einen vordefinierten Schwellwert, wird der Tracker mit der parallel stattgefundenen Schätzung des SVM-Regressors reinitialisiert und neu gestartet.

Modellieren der Verteilung in Vektorraumabbildungen

Indem man ein extrahiertes Motiv aus einem Kamerabild pixelweise ausliest, kann es als Vektor zur weiteren Verarbeitung interpretiert werden. Die zugrunde liegende Annahme bei Template-Vergleichen ist die, dass solche Motivvektoren auf Ähnlichkeit untersucht werden können. Bezogen auf Kopfdrehungen bedeutet das im einfachsten Fall, dass Kopfmotive derselben Orientierung einander ähnlich genug erscheinen, um im direkten Vergleich mit hinterlegten Beispielen zu bestehen. Diese Ähnlichkeit kann dabei im zugehörigen Vektorraum, in dem die

Motivvektoren aufgespannt werden, zum Beispiel durch geometrische Distanzmaße ermittelt werden.
Wie in Abschnitt 2.1.1 bereits beschrieben wurde, unterscheiden sich Gesichts- und Kopfmotive nicht nur anhand ihrer jeweiligen Kopfdrehungen. Vektoren gleicher oder ähnlicher Kopforientierungen unterliegen weiteren Varianzen, die auch durch Wahl unterschiedlicher Distanzmaße nur marginal eingeschränkt werden kann. Grundidee ist jedoch eine geometrische Verteilung der verschiedenen Kopfposen in einem Vektorraum, die parametrisch oder gegebenenfalls nichtparametrisch, beschrieben und modelliert werden kann. Ein neues Motiv wäre so durch seine Lage in der beschriebenen Verteilung singulär. Im Fall von Posenballungen im Vektorraum durch seine Zugehörigkeit zu einer entsprechenden Klasse, durch Interpolation oder im Fall eines kontinuierlich beschreibenden Modells sogar durch eine Approximation seiner Regression.

Vektorraumabbildungen haben die Grundlage die Merkmalsvektoren so abzubilden, dass die zur Unterscheidung notwendigen Varianzen berücksichtigt bleiben, unwesentliche Merkmale darüber hinaus aber auf ein Minimum reduziert werden. Ein Beispiel dafür ist die *Hauptkomponentenanalyse* (engl. Principal Component Analysis, im folgenden PCA genannt), die in einem gegebenen Datensatz die Dimensionsachsen der dominanten Varianzen erkennt und eine Koordinatentransformation derart vornimmt, dass die transformierten Hauptachsen entlang der Varianzen ausgerichtet sind. Damit kann eine wesentliche Reduktion der Dimensionalität in der Motivbeschreibung erreicht werden, indem weniger dominante Varianzdimensionen ignoriert werden können.
Die Arbeiten von [SB02, DMP96] verfolgen diesen Ansatz. In beiden Fällen werden für alle Kopforientierungsklassen jeweils Eigenraumabbildungen berechnet und neue Motive durch Abbilden in die entsprechenden, hypothetisierten Räume über eine einfache Distanzminimierung einer Klasse zugeordnet. Während Srinivasan in [SB02] allerdings nur die reine Klassifikation mittels PCA untersucht und dazu einen vorgegebenen Datensatz extrahierter Kopfausschnitte heranzieht, beschreibt Darrell in [DMP96] eine Einbettung seiner Drehschätzung darüber hinaus in ein Trackingsystem, das den Kopf und dessen Orientierung kontinuierlich verfolgt und Hypothesen glättet. Dabei wird mit einer Kamera mit Weitwinkelobjektiv die Position der Person im Raum ermittelt, indem mit einer Differenzsegmentierung des Vordergrunds vom Hintergrund die Körpersilhouette der Person extrahiert wird. Darauf aufbauend ermittelt Darrell anschließend die Kopfposition heuristisch. Eine zweite Kamera mit verstellbarer Brennweite fokussiert schließlich den ermittelten Kopfkandidaten und extrahiert das beobachtete Kopfmotiv zur weiteren Klassifikation.

Eine weitere Abbildung ist die *Lineare Diskriminanz-Analyse* (engl. Linear Discriminant Analysis, im Folgenden LDA genannt) die den Vektorraum so abbildet, dass durch eine Trennebene zwei Klassen linear separiert werden können. Diesem Ansatz folgen insbesondere Chen et al. in ihrer 2003 veröffentlichten Arbeit [CZH+03]. In ihrem Verfahren beschreiben die Autoren

eine Regression der Kopfdrehung auf Basis einer LDA zweier Drehklassen. An Support Vektor Maschinen erläutern die Autoren den Vorteil von Transformationen auf Basis einer LDA, da mit dieser Motive derselben Klasse auch nach der Abbildung geometrisch noch immer dicht beieinander angeordnet bleiben. SVM-Ansätze optimieren im Vergleich dazu lediglich die Separierung der beiden Klassen, lassen die klasseninterne Struktur der Vektoren hingegen unberücksichtigt. Durch die geometrisch dicht beieinander angesiedelten Motivvektoren, implementieren Chen et al. eine Interpolation über Kopfdrehungen von $-10°$ hin zu Kopfdrehungen von $+10°$: Dafür separieren die Autoren mit Hilfe der LDA die beiden Klassen voneinander und wandern entlang der Gradienten in der Abbildung von der einen Klasse hin zur anderen. Trotz der ursprünglichen Diskretisierung in nur zwei Winkelklassen, erreichen sie damit eine angenäherte Regression des Drehwinkels, die sich an der geometrischen Verteilung im abgebildeten Vektorraum ausrichtet und so nur wenige Beispiele zum Einlernen der Struktur braucht.

2.1.2. Der Sonderfall niedrig aufgelöster Kamerabilder und Drehungen im Bereich $\pm 180°$

Sind Kameras in großer Entfernung von der zu beobachtenden Person angebracht, so sind die Aufnahmen des Kopfes klein und unterliegen einer niedrigen Auflösung. Augen oder die Nasenspitze sind darin, falls überhaupt als solche erkennbar, nur noch unscharf abgebildet und eine präzise Lokalisierung solcher Merkmale in der Praxis nur unter Umständen umzusetzen. Solche Aufnahmen stellen bisherige Systeme vor neue Herausforderungen. Diese werden aber auch gerade durch die gesteigerte Nachfrage von Überwachungssystemen zunehmend gefragter und interessanter.
So ist bereits die Detektion von Gesichtern und Köpfen im Allgemeinen schwierig: In Umgebungen, in denen die Kameras unaufdringlich angebracht wurden, sind Kopfdrehung nicht mehr im Bereich $\pm 90°$ zu einer Kamerasicht zu erwarten, sondern erscheinen auch vollständig von der Kamera weg gedreht. Solche Beobachtungen bringen mit Drehungen bis $\pm 180°$ zusätzliche Varianzen durch Frisuren, Kopfbedeckungen und bei schlechten Darstellungen auch durch den fehlenden Kontrast zum Hintergrund. Zum anderen basieren alle bisherigen Vorgehensweisen prinzipiell auf den Darstellungsunterschieden gedrehter Gesichter, nicht aber Köpfe: Neben den Gesichtsmerkmalen bieten Köpfe nur wenige weitere, die ihre jeweilige Drehung bestimmbar machen. Ist das Gesicht von der Kamera weg gedreht und Gesichtsmerkmale dadurch nicht vorhanden, deuten Ohren und das Fehlen entsprechender Hautfarbbereiche im Motiv auf den groben Drehwinkel hin. Je nach Frisur fällt die Hautfarbverteilung dabei jedoch unterschiedlich aus. Entsprechend verhält es sich mit durch die Haare verdeckter Ohren.

Das Erkennen der Kopfdrehung unter Aufnahmebedingungen mit niedriger Auflösung ist ein junges Anwendungsgebiet. So umgehen viele Arbeiten die Kopfdrehung gänzlich und appro-

ximieren die Orientierung des Blickfelds einer Personen allein durch deren Trajektorie oder anderen Beobachtungen ihrer Aktivität [Bux03, GSRL98, DH04, EBMM03, JH95, ME02].

Eine tatsächliche und erste Analyse bezüglich der Übertragbarkeit regulärer Verfahren zur Kopfdrehungserkennung, veröffentlichte Lisa Brown im Jahr 2002 [BT02]: Darin fasst sie die bis dato wenigen Verfahren zusammen, die auf nur groben Bildauflösungen und gegebenenfalls auch größeren Winkelwertebereichen Kopforientierungen schätzten [KBS00, NF96, RR98, WT00, ZPC02]. Alle darin aufgeführten Arbeiten sind ansichtsbasiert und unterscheiden sich allein in der Wahl der jeweiligen Motivabbildung oder der Auswahl zu detektierender Merkmale, sowie der eigentlichen Klassifikation. Mit einer Winkelgranularität von nur 1° stellt die Arbeit von Krüger et al. dabei die feingranularsten Hypothesen, beschränkt sich aber im Gesamten auf Kopfrotationen im Bereich $\pm 20^\circ$ [KBS00]. Krügers Verfahren baut auf der Arbeit von Bruske et al. auf [BAMPS98], die ein Selbstorganisierendes Netz für die Drehwinkelschätzung zugrunde gelegt. Ebenso wie Bruske benutzt Krüger die Koeffizienten einer Gabor-Repräsentation der lokalisierten Kopfregion. Der hier neue Ansatz, der ein ganzes Netzwerk an Gabor-Wavelets einsetzt, um das Gesicht feiner und genauer zu lokalisieren, erlaubt nur deutlich eingeschränkte Kopfdrehungen bis zu $\pm 20^\circ$ in horizontaler Richtung - damit stets frontale Aufnahmen des Gesichts vorliegen und die Lokalisierung nicht überbeansprucht wird. Ebenso wie Bruske evaluierte Krüger sein System nur auf Kopfaufnahmen einer Puppe und stellte bislang keinen Bezug zu realistischeren Datensätzen her.
Mit Niyogi und Freemans Arbeit aus dem Jahr 1996 [NF96] wurde dasselbe System zum Vergleich herangezogen, das bereits in Abschnitt 2.1.1 vorgestellt wurde. Ihre Winkelgranularität ist mit 20° für horizontale, 30° für vertikale Drehungen und angenommenen Kopforientierungen in einem Winkelwertebereich von nur $\pm 50^\circ$ aber vergleichsweise grob aufgelöst.
Mit der Arbeit von Wu und Toyama [WT00] werden Gabor- und Gaussfilter auf dem lokalisierten Kopfausschnitt angewandt und deren Ausrichtung an vordefinierten Stellen innerhalb des Kopfmotivs probabilistisch modelliert. Durch ein Maximum A-Posteriori-Verfahren wird im Anschluss die Wahrscheinlichkeit einer neuen Beobachtung berechnet, eingelernten Beispielmotiven einer angenommenen Winkelklasse zu entsprechen und die Kopfdrehung so erstmalig über einen horizontalen Wertebereich von $\pm 180^\circ$ erkannt - allerdings ebenfalls mit einer groben Winkeldiskretisierung von 20°-umfassenden Klassen. Jedoch sind die dabei zum Einsatz kommenden Motivbeispiele mit einer Auflösung von 32×32 sehr klein und verdeutlichen neben dem unterstützten Winkelwertebereich auch die überdurchschnittlich hohe Klassifikationsgenauigkeit trotz der herausfordernden Motivdarstellungen.
Mit der Arbeit von Zhao et al. [ZPC02] stellt Brown darüber hinaus in ihrem Vergleich ein System gegenüber, dass mit einem Neuronalen Netz direkt auf dem lokalisierten Kopfbild die Kopfdrehung in horizontaler bzw. vertikaler Drehrichtung schätzt und Winkelbereiche von jeweils $\pm 90^\circ$ und einer Auflösung von 10° erlaubt. Die Netze schätzen dabei von kontrastverstärkten,

auf 48×48 Pixel skalierten Grauwertbildern die Wahrscheinlichkeitsfunktion des Winkelwertebereichs der jeweiligen Rotation. Die vorangehende Histogrammnormalisierung des Grauwertbilds wird dabei darin begründet, dass durch Anheben des Kontrasts Beleuchtungsvarianzen im Bild entgegengewirkt werden kann.
In einer eigenen Evaluation implementierte Brown den Ansatz von [WT00] und ein Schätzen der Kopfdrehung mit Neuronalen Netzen ähnlich zu [ZPC02] nach und untersucht darin gezielt die Sensibilität der Ansätze bezüglich Motivgröße - und damit auch Schärfe - sowie deren Robustheit gegen Lokalisierungsvarianzen in Form von verschobenen und suboptimal ausgerichteten Kopfmotivregionen. Brown unterzieht beide Ansätze damit auch erstmalig demselben Datensatz und verdeutlicht die nahezu äquivalenten Präzisionen beider Ansätze bis zu Motivauflösungen von nur 8×8 Pixel. Mit so einer niedrigen Auflösung der Kopfausschnitte reduziert sich die nutzbare Information im Motivbild erheblich; infolgedessen zeigen sich beide Ansätze als sehr anspruchsvoll gegenüber einem konsistenen Lokalisieren des Kopfes. Besonders anfällig sind beide dabei insbesondere für Motive, in denen der Kopf nicht komplett umfasst und eingegrenzt wird, sondern nur ein Teil von ihm, wie zum Beispiel nur bis zum Mund statt zum Kinn. Hintergrundrauschen, das daneben durch zu tolerante Lokalisierungen mit einfließt, findet hingegen jedoch kaum Auswirkungen auf die Genauigkeit. Lediglich bei Inkonsistenzen in der Motivbreite und -höhe zeigen sich die beiden Verfahren gegenüber sensibel. Darüber hinaus kann Brown eine starke Beeinträchtigung erkennen, wenn die Kamera nicht frontal vor dem Benutzer angebracht ist, sondern diesen von oben herab beobachtet. Köpfe erscheinen so in ihrer Ruhelage bereits nach unten geneigt - eine Tatsache, die bereits jetzt auf mögliche Schwierigkeiten hindeutet, Kopfdrehungen aus weiter Distanz mit Hilfe von Kameraaufnahmen unterhalb der Raumdecke schätzen zu wollen.

Weitere Arbeiten, die sich mit der Problematik niedrig-aufgelöster Kopfdarstellungen auseinandersetzen, gibt es nur wenige. 2004 veröffentlichten Wang und Ji einen Ansatz, der Detektoren einsetzt, um gedrehte Gesichter in realistischen Überwachungsaufnahmen finden zu können. Dabei werden allerdings ausschließlich Gesichter mit horizontalen Drehungen im Bereich $\pm 90°$ berücksichtigt, die mit Kaskaden von hierarchischen Support Vektor Maschinen in eine von sieben möglichen Drehwinkelklassen klassifiziert werden. Mit Motivgrößen von 20×20 bis 35×35 Pixeln ist die Auflösung zwar auch gering, die grobe Winkelgranularität jedoch ist typisch für detektorbasierte Ansätze.

Robertson und Reid beschrieben 2006 ein Verfahren, das Kopfdrehungen bei Motivgrößen von 20×20 erkennt [RR06]. In ihrem System finden sie durch Vordergrund- und Hautfarbsegmentierung mögliche Gesichts- bzw. Kopfregionen im Bild, auf denen anschließend die Drehung erkannt wird. Die Kopfdrehung wird dabei in Winkelklassen von je $45°$-Breite klassifiziert und über den gesamten Wertebereich von $360°$ erkannt. Für ein Motiv berechnen die Autoren hierfür pro Pixel die Zugehörigkeit einem initialen Hautfarbmodell zu entsprechen. Die Rückprojektion

dieser Konfidenz in das Kamerabild, dient schließlich als Merkmal, indem die Pixelintensitäten innerhalb der Motivregion konkateniert werden. Die Klassifikation selbst geschieht darauf über Templatevergleiche. Die jeweiligen Repräsentanten der Winkelklassen sind dabei in einer Baumstruktur hinterlegt, um den Such- und Vergleichsprozess zu beschleunigen.
Auf [RR06] setzt hingegen das System von Benfold und Reid aus dem Jahr 2008 auf [BR08]. Als Weiterentwicklung beschreiben die Autoren dabei die geringe Größe der Kopfmotive von nun möglichen 10×10 Pixeln, sowie eine detailliertere Farbsegmentierung durch jeweilige Ballungen für Haare, Haut- und Hinterdarstellungen. Die Klassifikation erfolgt auch hier in acht Klassen über einen horizontalen Drehbereich von 360°, indem ebenfalls über den Abstieg in einer Baumstruktur Templates mit dem Motiv verglichen werden. Wohingegen jedoch in der ursprünglichen Arbeit ein manueller Initialisierungsschritt per Hand notwendig war, um ein initiales Modell der Farbverteilung zu erstellen, benutzen Benfold und Reid hier ein über 1000 Beispielköpfen automatisch eingelerntes Farbmodell und umgehen so die zuvor nur mögliche Semiautomatik.
Im Jahr 2009 beschrieben Orozco et al. ein weiteres System für den praktischen Einsatz in Überwachungsaufnahmen, um Kopfregionen der Größe 10×20 bis 40×60 Pixel verarbeiten zu können [OGX09]. Auch hier unterstützen die Autoren wie auch [RR06, BR08] zwar horizontale Drehwinkel bis zu 360°, unterteilen aber auch wieder in insgesamt nur acht 45° große Drehklassen. Dagegen konzentrieren sich Orozco et al. auf reale Aufnahmen, in denen viele Personen vorhanden sind und damit viele Überlappungen und dicht beieinander auftretende Personensilhouetten die Kopflokalisierung und -drehwinkelschätzung herausfordern. Indem angenommen wird, dass die maßgebliche Bewegung im Bild von Personen stammt, wird Vordergrundsegmentierung eingesetzt um die eigentlichen Personen vom Hintergrund zu trennen. Detektoren für Torso und Unterkörper separieren schließlich die Körperteile der segmentierten Silhouetten. Zur eigentlichen Drehschätzung bilden die Autoren Differenzbilder der Kopfmotive zu den acht jeweils hinterlegten Drehrepräsentanten und klassifizieren durch ein einfaches Mehrheitsvotum über die Ausgaben einer Support Vektor Maschine für mehrere Klassen.

2.1.3. Komplementarität und Redundanz durch Einsatz von Multisensorik

Wie im vorigen Abschnitt dargestellt wurde, bieten bisherige Ansätze, die sich mit Kopfmotiven niedriger Auflösung auseinandersetzen, lediglich Klassifikationen mit einer groben Winkelgranularität von in der Regel 45°. Zum Einsatz kommen dabei immer nur einzelne Ansichten. Wie im Rahmen dieser Arbeit, ist es in intelligenten Räumen oder größeren Umgebungen dagegen möglich mehrere Kameras so anzubringen, dass sie zwar denselben Bereich, diesen aber aus unterschiedlichen Blickwinkeln, diesen so aber komplementär zueinander, beobachten können. Ist in einer Ansicht ein Hinterkopf zu sehen, kann derselbe Kopf in einer anderen Ansicht so gegebenenfalls frontal aufgezeichnet werden. Und eine dritte wäre in der Lage eine zusätzliche

Profilansicht des Kopfs beizusteuern. Die Motivation dabei ist es, durch Ausnutzen verschiedener Blickwinkel auf dasselbe Motiv, zum einen die Robustheit bei Kopfdrehungen über den gesamten Winkelwertebereich zu erhöhen als auch eine feingranularere Auflösung des Winkelwertebereichs umsetzen zu können.

In [RHE03] beschreiben Ruddarraju et al. 2003 ein erstes System, das mittels eines Kameraarrays, das im Kreis um eine Person aufgereiht ist, versucht, Gesichtsmerkmale in zwei benachbarten Ansichten zu detektieren, um geometrisch Kopfdrehungen über 360° erkennen zu können. Durch den künstlichen Aufbau der Kameras stellt sich allerdings die Frage, inwiefern sich ein solches Vorgehen auf reale Aufnahmebedingungen übertragen lässt - selbst bei der heutzutage hohen Verfügbarkeit hochauflösender Sensoren.
Tatsächliche Aufnahmebedingungen in Umgebungen, in denen Kameras in weiter Entfernung von den zu beobachtenden Personen angebracht sind, stellen durch die große Entfernung mit der die Person aufgenommen wird eine wesentliche Herausforderung dar. Je nachdem wie nah an die Kamera herangegangen werden kann, variiert die beobachtbare Kopfgröße dabei sehr. Systeme können dabei nicht annehmen konstanten Aufnahmebedingungen ausgesetzt zu sein. So kommt es vor, dass Frontalgesichter nah an einer Kamera beobachtet werden können, während eine entgegengesetzte Sicht den Hinterkopf in weiter Entfernung aufnimmt oder umgekehrt die Frontalansicht in weiter Entfernung zu einer Kamera steht, während die nahe Ansicht den Hinterkopf detailliert beobachten kann.

Mit [TBC+03, KSS+06, ZSA09] lassen sich drei Systeme finden, die sich unabhängig voneinander mit Kopfdrehungen auf realistischeren Aufnahmebedingungen in Mehrkameraumgebungen auseinandersetzen, in denen Kopfmotive lediglich eine Darstellungsgröße von 8×8 beziehungsweise 20×20 Pixel vorweisen.
Tian et al. setzen in [TBC+03] dazu auf ihren vorherigen Ansatz [BT02], in dem Neuronale Netze gegen Gaborrepräsentationen auf niedrig-aufgelösten Kopfregionen experimentell evaluiert wurden. Köpfe werden darin lokalisiert, indem Personensilhouetten in jeder Kameraansicht separat durch eine Vordergrundsegmentierung extrahiert und heuristisch auf Gliedmaßen, Beine und Kopf unterteilt werden. Die so detektierten Kopfregionen werden anschließend unabhängig voneinander mit einem Mehrschichtigen Perzeptron in ihrer jeweiligen Drehung geschätzt: Die Neuronalen Netze hypothetisieren dabei die Wahrscheinlichkeitsfunktion über diskrete aber nur grobe Drehwinkelklassen; die Schätzungen aller Ansichten werden anschließend mit einem Maximum-A-Posteriori-Verfahren zu einer Gesamthypothese vereint.
Kobayashi et al. stellen im Kontrast dazu ein Verfahren vor, das den Kopf mit seiner Position, Maße und Orientierung dreidimensional hypothetisiert und den Kandidaten durch Projektion in die jeweiligen Kameraansichten bewertet und verfolgt [KSS+06]. Die Bewertung in den Ansichten geschieht dabei durch Detektoren auf Basis von Viola und Jones [VJ01], indem die geschätzte und erwartete Kopfdrehung dazu dient, die entsprechenden Detektoren

auszuwählen und auf der in das Kamerabild projizierten Kopfregion anzuwenden. Die Menge final durchlaufener Kaskadenstufen interpretieren die Autoren probabilistisch. Eine endgültige und gemeinschaftliche Bewertung ergibt sich schließlich aus dem Produkt der ansichtsbasierten Wahrscheinlichkeiten.
Zabulis et al. setzen daneben ebenfalls Detektoren ein, fusionieren die Ansichten allerdings zuerst auf Signalebene: Der hypothetisierte Kopfkandidat wird dabei in die Kameraansichten rückprojiziert und die entsprechende Beobachtung der Rückprojektion auf ein dreidimensionales Kopfmodell übertragen [ZSA09]. Auf der damit gewonnenen, planaren Kopftextur wird im Anschluss die Orientierung mit Hilfe einer Gesichtsdetektion durch entsprechende Detektoren ermittelt. Bezogen auf das Weltkoordinatensystem, lässt sich hieraus schließlich die endgültige Drehwinkelschätzung herleiten.

Die CLEAR Evaluationen

In der mangelnden Verfügbarkeit entsprechender Datensätze liegt der Hauptgrund für fehlende, weitere Ansätze und Systeme die sich mit dieser Thematik befassen. Die notwendige Sensorausstattung und Datenaufbereitung ist bisher mit vielen Anwendungsgebieten in der Literatur und Forschung nicht konform gewesen. Mit Aufkommen zunehmender Überwachungssystematik und aufmerksamen Umgebungen wie der hier beschriebenen, wird die Problemstellung aber zunehmend populärer.
Als erste öffentliche Evaluation bezüglich solcher und darüber hinaus auch verwandten Problemstellungen in aufmerksamen Umgebungen stellen die CLEAR-Evaluationen, in den Jahren 2006 und 2007 die bisherige Referenz aktueller Forschung dar. Wie bereits im Eingangskapitel dargelegt, wurde hierfür der in dieser Arbeit aufgenommene und -bereitete Datensatz zur Kopfdrehungserkennung publik gemacht und dient bis dato als Basis vergleichbarer Lösungsansätze, um die Kopfdrehung in Multisensor-Umgebungen auf Kopfmotiven zu erkennen, die nur in niedriger Auflösung vorliegen. Im Zuge der Evaluationen, wurden so die einzig wenigen, vergleichbaren Ansätze hinsichtlich derselben Problemstellung veröffentlicht. Im Folgenden soll deswegen hervorgehoben auf diese im Einzelnen eingegangen werden. Alle bisher auf CLEAR-Daten evaluierten Systeme bauen grundlegend auf ansichtsbasierten Verfahren auf [ZHLH06, CFCP06, PZ07, LB07, BO07b, CFCP07, YZF+08].

Detektor-basiert gehen die Arbeiten von Potamianos et al. [PZ07] und Zhang et al. [ZHLH06] vor, indem sie mit Detektoren kandidierende Drehwinkelklassen auf allen jeweiligen Kameraansichten separat erkennen und die Hypothesen anschließend probabilistisch, zum Beispiel über ein Trackingverfahren, vereinen.

Auf Signalebene, fusionieren Yan et al. in [YZF+08] die Merkmale aller Ansichten, indem sie die lokalisierten Kopfausschnitte aus allen Kamerabildern konkatenieren und anschließend eine

Vektorraumabbildung vornehmen, mit der sie die Verteilung der Kopfdrehungen modellieren und nachbilden.

Canton-Ferrer et al. fusionieren in [CFCP06] die Ansichten ebenfalls bereits auf Signalebene: der Kopf wird darin als dreidimensionaler Ellipsoid modelliert und das Ergebnis einer Hautfarbsegmentierung aus allen Kameraansichten auf diesen projiziert. Die so erhaltene Textur des Ellipsoiden, die die Verteilung von Hautfarbe auf dem approximierten Kopf darstellt, gibt Aufschluss über die beobachtete Kopfdrehung. Dazu nutzen Canton-Ferrer et al. die Lage des Erwartungswerts der Hautfarbverteilung als eigentliche Angabe der Kopforientierung. In einer darauf folgenden Arbeit erweitern die Autoren ihr System und setzen zusätzlich einen Partikelfilter ein, dessen Hypothesen sowohl Orientierung als auch Position und Größe des Kopf beschreiben [CFCP07]. Die Textur, die auch hier durch Farbprojektion auf den Kopfkandidaten erhalten wird, wird darüber hinaus allerdings mit gelernten Mustern verglichen, die die optimale Textur der Hautfarbverteilung bei einer gegebenen Kopfdrehung vorgibt.

Ähnlich gehen auch Lanz et al. mit ihrem System in [LB07] vor: Statt einer reinen Farbsegmentierung, bauen sie auf einer initialen Aufnahme der Person aus unterschiedlichen Ansichten auf, mit der sie Histogramme der Farbverteilungen von Kopf, Torso und Beine pro Person erstellen. Ebenfalls in einem Partikelfilteransatz werden die Hypothesen für Lage und Orientierung des jeweiligen Körperbereichs in das Bild rückprojiziert und das dort vorzufindende Histogramm gegen die initial erstellten Modelle verglichen. Die geschätzte Orientierung dient dabei insbesondere dazu, die Referenzhistogramme gewichtet zu interpolieren, um so eine kontinuierliche Schätzung der Drehung erhalten zu können.

Statt einer Merkmalsfusion entscheiden sich Ba et al. in [BO07b] für eine Kameraselektion, so dass die Kopfdrehung nur auf Basis von zwei der vier vorhandenen Ansichten geschätzt wird. Diese werden dabei so ausgesucht, dass sie hierzu den größten Hautfarbanteil im Kopfkandidaten vorweisen. Der Kopfkandidat wird über ein Trackingverfahren nachverfolgt. Auf

Abb. 2.1.: Beispielmotive aus dem CLEAR'07-Datensatz. Dargestellt sind Kopfdrehungen für unterschiedliche Personen aus jeweils allen Kameraansichten. Gesichtsmerkmale sind aufgrund der geringen Größe der Motive nur hinreichend erkennbar. Wegen der verschiedenen Blickwinkel der Kameras sind darüber hinaus Hinterkopfansichten zu berücksichtigen.

den beiden Ansichten wenden die Autoren schließlich ihr in früheren Arbeiten veröffentlichtes, einzelkamerabasiertes Verfahren an, das auf Gauss- und Gaborfiltern aufsetzt und Repräsentationen der Drehwinkelklassen durch Ballungen der Merkmalsvektoren zusammensetzt [BO05]. Die Hypothesen der beiden Ansichtsschätzungen werden schließlich gemittelt um die endgültige Winkelschätzung auszugeben.

2.2. Deduktion des visuellen Aufmerksamkeitsfokus

Mit erkannten Kopfdrehungen steht noch immer die Problematik aus, tatsächliche Aufmerksamkeitszuwendungen anhand der Orientierung des Blickfelds abzuleiten. Im hiesigen Abschnitt soll deshalb auf diese nachhaltige Fragestellung eingegangen und ein umfassender Überblick über verwandte Arbeiten in der Literatur gegeben werden, die sich ebenfalls damit auseinandersetzen zunächst von der Kopfdrehung auf den Aufmerksamkeitsfokus einer Person schließen zu können.
Geprägt wurde der Begriff der *visuellen Aufmerksamkeit* im Bereich der Mensch-Maschine-Interaktion von Langton et al., um das Nachvollziehen der Blickrichtung in einen Zusammenhang mit der Kopfdrehung zu bringen. Die Grundidee war, damit ein umfassenderes Verständnis der visuellen Aufmerksamkeitszuwendung einer Person zu erhalten [LWB00]. Nach Langtons Auffassung ist ein reines Beobachten der Ausrichtung der Augen nicht unbedingt ausreichend um eine zuverlässige Aussage über die Aufmerksamkeit einer Person tätigen zu können. Stattdessen wird dieser Bezug erst durch weitere, sekundäre Indikatoren hergestellt, so dass außen stehende Beobachter erst anhand der Kopfdrehung und gegebenenfalls einer zusätzlichen Gestenerkennung die soziale Ausrichtung der Aufmerksamkeit auf ein Gegenüber feststellen.
Aufgrund des damit in Verbindung gebrachten Umfangs notwendiger Sensoraufzeichnungen, sind bis heute dafür ausgelegte Datensätze und Studien hauptsächlich auf Besprechungs- und Diksussionssituationen beschränkt, da dort schon recht früh eine Grundlage für einen hinreichend kontrollierten Sensoraufbau inklusive der darüber hinaus notwendigen Beobachtungsqualität durch die vorherrschende Arbeitsumgebung gewährleistet werden konnte.
Die im Folgenden beschriebenen Ansätze lassen sich in zwei unterschiedliche Vorgehensweisen aufteilen: Zum einen gibt es geometrische Verfahren, die die Kopfdrehung nicht nur als annähernden Indikator der Blickrichtung verstehen, sondern sie der eigentlichen Blickrichtung sogar gleichsetzen. Mit den erkannten Drehwinkeln des Kopfes kann ein Vektor dabei entsprechend der Orientierung des Gesichts ausgerichtet werden. Dieser Vektor wird dann als eigentliche Blickrichtung interpretiert, so dass ein Schließen auf ein mögliches Interaktionsziel damit auf einen einfachen, geometrischen Vergleich reduziert werden kann, welches Objekt oder welche Person diesem Vektor am nächstliegendsten erscheint.
Der eigentlich nur groben Annäherung der Blickrichtung werden im Gegensatz dazu andere Verfahren gerecht, die die Kopfdrehung lediglich als Orientierung des Blickfelds auffassen.

Aufmerksamkeitsziele werden darin über ihr Schnittvolumen bestimmt oder deren Anordnung im Blickfeld bei jeweiliger Aufmerksamkeitszuwendung probabilistisch modelliert. Gegebenenfalls werden darüber hinausgehende Kontextbeobachtungen, die die Aufmerksamkeitszuwendung einer Person in der vorherrschenden Situation beeinflussen können, als A-Priori-Wahrscheinlichkeiten einbezogen. Dadurch ergibt sich ein umfassenderes Verständnis der stattfindenden Szene, was dem Versuch entspricht, die faktische Motivation hinter einer Zuwendung modellieren zu können.
Ein Vorteil rein geometrischer Herangehensweisen ist der, dass ein Bestimmen der Aufmerksamkeitszuwendung beziehungsweise dem Erkennen der Interaktion zweier Personen miteinander, keine weiteren Situationsannahmen vorausgesetzt werden müssen. Eine disjunkte Platzierung der Personen und Objekte in der Umgebung erscheint damit ausreichend um die Zuwendung allein aufgrund der erkannten Kopfdrehung unterscheiden zu können. Dementsprechend sind solche Verfahren schnell berechen- und in Echtzeit ausführbar. Die Disjunktivität der Ziele beschränkt jedoch die Menge möglicher Interaktionssituationen von Anfang an. Darüber hinaus trifft man mit der direkten Abbildung der Kopfdrehung auf die eigentliche Blickrichtung die grundlegend falsche Annahme, dass der Blick zum Beispiel nicht von der Situation und der Interaktionsdynamik innerhalb einer Diskussionsgruppe beeinflusst wird. Das zum Richtungsvektor einer Kopfdrehung nächstgelegene Ziel, erscheint zwar zunächst intuitiv korrekt und naheliegend, entspricht aber häufig nicht der tatsächlichen Blickrichtung! Diese hängt tatsächlich vom Kontext einer Situation ab, so dass Kopfdrehung und Augenbewegung auch getrennt voneinander eingesetzt werden - je nachdem wie beispielsweise die Interaktion zwischen Personen stattfindet und ein Beobachter dem Geschehen zu folgen versucht.

2.2.1. Kopfdrehung als geometrischer Ausdruck der Blickrichtung

Geht man von der Annahme aus, dass alle möglichen Aufmerksamkeitsziele für einen Beobachter hinreichend disjunkt voneinander platziert sind und die Situation vorgibt, dass dieser Beobachter seinen Kopf bei einer Aufmerksamkeitszuwendung stets dem jeweiligen Ziel zudreht, dann findet sich ein intuitiver Ansatz darin, das entsprechende Ziel geometrisch zu bestimmen. Ein Verfahren, dass direkt von der Kopfdrehung auf Aufmerksamkeitsziele schließt, beschreiben zum Beispiel Murphy-Chutorian und Trivedi in [MCT08a]. Sie bilden darin ein klassisches Besprechungsszenario in einer Arbeitsumgebung nach, in dem Personen um einen Tisch verteilt Platz nehmen, miteinander diskutieren und jeweils zu vereinzelten Präsentationen vor einer Leinwand aufstehen. Durch automatisches Erkennen der Kopfdrehung, ist dem System die dreidimensionale Mittelpunktposition aller Köpfe der jeweiligen Teilnehmer bekannt, was gleichzeitig auch allen möglichen Interaktionszielen entspricht. Als Aufmerksamkeitsziel der Gruppe wird anschließend jenes Ziel klassifiziert, das euklidisch insgesamt am dichtesten zu den Vektoren positioniert ist, die den Orientierungen der Kopfposen der Teilnehmer entsprechen.

Ähnlich gehen Farenzena et al. in [FTB+09] vor, berücksichtigen dabei jedoch gezielt Umgebungen, in denen die Beobachtungsqualität der Kopfmotive nur grobe Schätzungen in vier Klassen zulässt: Nord, Ost, Süd und West. Sie beobachten damit eine eingerichtete Teeküche, in der Mitarbeiter regelmäßig zusammenkommen und miteinander interagieren. Basierend auf den erkannten Kopfdrehungen, spannen die Autoren Sichtpyramiden auf, die bei zugewandten Kollegen zu Schnittvolumen führen - je nachdem wie dicht beieinander die Personen sich gegenüber stehen und zugewandt sind. Heuristisch legen die Autoren für Zuwendungen dabei eine maximale Distanz von 2 m fest. Eine Interaktion wird folglich detektiert, wenn sich zwei Personen so mindestens 10 Sekunden lang zugewandt erscheinen.
Direkt auf das Schnittvolumen aufgespannter Sichtpyramiden, setzen dagegen Canton-Ferrer et al. in [CFSP+08]. Anstatt jedoch für die jeweiligen Person auf ein Aktionsgegenüber zu schließen, wird allein der Schnitt aller orientierter Blickfelder erkannt und davon auf von mehreren Personen anvisierte Bereiche im Raum geschlossen. Die Schnittvolumina werden aber lediglich als Bereiche gesteigerten Interesses hervorgehoben.

2.2.2. Probabilistisches Beschreiben wahrscheinlicher Aufmerksamkeitszuwendungen

Um den Kontext einer Szene erfassen zu können, reicht ein geometrisches Distanzmaß allein nicht aus. Dem gegenüber stehen deshalb probabilistische Ansätze. In ihnen wird die Kopfdrehung als Beobachtung aufgefasst und gegeben den jeweiligen Personen und Objekten die prinzipiell angeschaut werden können, bedingt modelliert. Für darüber hinausgehende Kontextbeobachtungen besteht dabei die Möglichkeit, diese als zusätzliche Zufallsvariablen mitzuberücksichtigen. In den nachfolgend vorgestellten Arbeiten reichen diese dabei von Bewegungsmustern hin zu Situationsheuristiken, die die A-Priori-Wahrscheinlichkeit der Ziele angesehen zu werden, beeinflussen sollen.
Erste empirische Versuche gehen dabei zurück auf Stiefelhagen et al. [SFYW99]. Aufmerksamkeitszuwendung werden darin als stochastischer Markov-Prozess in Form eines Hidden Markov Modells umgesetzt. Die entsprechenden Zustände umfassen dabei die übrigen Teilnehmer der Besprechung, die Übergänge den Wechsel der Zuwendung zwischen ihnen. Die Übergangswahrscheinlichkeiten legte Stiefelhagen in der Arbeit hierzu empirisch festgelegt, die Emissionswahrscheinlichkeiten stammen jedoch aus gelernten Normalverteilungen, deren Mittelwert und Standarabweichung unüberwacht auf Trainingsdaten eingelernt wurden. Stiefelhagen zufolge entsprechen die von ihm beobachteten Kopfdrehungen in ihrer Form Normalverteilungen, wenn wiederholt dasselbe Ziel angeblickt wurde. Dabei beschreibt er, dass die Mittelwerte keineswegs den eigentlichen Blickwinkeln zu den jeweiligen Aufmerksamkeitszielen entsprechen, sondern verschoben zu diesen liegen. Initial positioniert Stiefelhagen die Gausskomponenten daher mit den eigentlichen Blickwinkeln zu den jeweiligen Interaktionspartnern, adaptiert die-

se aber anschließend mit einem Maximum-A-Posteriori-Verfahren auf die eigentlichen Ballungszentren in den Trainingsdaten. Stiefelhagen konnte so als erster einen systematischen Versuch vorweisen, Fixationen auf Gesprächspartner, allein aufgrund der Kopfdrehung, erfolgreich nachvollziehen zu können.

Die Beobachtungsmodellierung der Kopfdrehungsballungen in Form von Normalverteilungen, wird bis heute in allen stochastischen Ansätzen als Grundlage eingesetzt. So auch in Folgearbeiten des Autors, in denen er anstatt der bisher eingesetzten Markovmodelle, Mischverteilungen für ein unüberwachtes Lernen der Ballungen und anschließendem Klassifizieren verwendet [SYW01a]. Sein System erweitert er damit dahingehend, dass nur noch die Anzahl der Ziele, bzw. Personen in der Besprechung, vorab bekannt sein muss, um zwischen ihnen bei der Zuwendung unterscheiden zu können. Mit Hilfe des Expectation Maximization Algorithmus, adaptiert der Autor dabei seine Modelle auf die beobachteten Kopfdrehungsballungen während der Besprechungen. Den dabei eingesetzten Mischverteilungen legt Stiefelhagen die gleiche Anzahl Komponenten zugrunde, wie Aufmerksamkeitsziele im Umfeld der jeweiligen Person vorhanden sind. Die Dichtefunktionen der Komponenten benutzt Stiefelhagen anschließend in einem naiven Bayes-Klassifikator als bedingte Modelle für die Beobachtungswahrscheinlichkeit der Kopfdrehung. Die A-Priori-Gewichte der Komponenten in der Mischverteilung setzt der Autor dabei als Verteilung über die Ziele ein.
Stiefelhagen zeigte den Mehrwert seines Ansatzes zunächst nur auf Besprechungsszenarien. In [SYW01b] übertrug er die Methodik aber ferner auch auf weitere Situationen in der Mensch-Maschine Interaktion. Konkret unterscheidet er darin, ob Sprachkommandos an einen Haushaltsroboter oder einen Videorekorder gerichtet werden, je nachdem wohin der erkannte Aufmerksamkeitsfokus deutet. Im Kontrast zum vorherigen Anwendungsfall ist hier nicht mehr das Beobachten der Szene durch Dritte (dem System das den Aufmerksamkeitsfokus nachvollziehen soll) der Fall, sondern eine direkte Interaktion mit verschiedenen Systemen, die anhand der Zuwendung erkennen sollen ob sie angesprochen werden und entsprechend reagieren sollen.

Zwar in Folgearbeiten, dafür zu ähnlichen Ergebnissen kommen Ba et al. mit ihren in [BO07a, BO06, OB07] beschriebenen Systemen. Darin gehen sie insbesondere näher auf die von Stiefelhagen et al. vorgestoßene Richtung ein, eine Beobachtungsmodellierung anhand normalverteilter Kopfdrehungsballungen zugrunde zu legen. Dem gegenüber stellen die Autoren Markovmodelle, um vergleichsweise auf deren Vorteil der Glättung in Hypothesensequenzen einzugehen. Im dargestellten Vergleich fällt die Entscheidung hierbei zu Gunsten der Glättung, die das durch die Kopfdrehungsschätzung verursachte Rauschen deutlich reduzieren kann und so kurzzeitige Ausreißer bei der Drehschätzung umgeht.

Mit Otsukas Arbeiten in [OTYH05, OYTM06] zeichnet sich derweil die erste Erkenntnis ab, dass die Beobachtungsmodellierung der Kopfdrehung allein nicht ausreicht und die Anordnung der Mittelwerte variabel auf unterschiedliche Situationen angepasst werden sollte. Otsuka fasst

darin die Mittelwertverteilung als eine weitere Zufallsvariable auf, die im Zuge einer Maximum-A-Posteriori-Evaluation über eine Besprechung gleichzeitig mit optimiert wird. Eine Einschränkung während der Adaption geben die Autoren dabei in Form einer A-Priori-Verteilung vor, die insbesondere eine mögliche Divergenz der Mittelwerte verhindern soll und jene insbesondere auf Winkelwerte nah der eigentlichen Ziele einschränkt. Ba und Obodez greifen diese Methodik in ihrer 2007 veröffentlichten Arbeit [OB07] auf und passen ihren bisherigen Ansatz ebenfalls darauf an. Anhand kognitiver Studien, die Kopfdrehungen mit tatsächlichen Blickwinkeln zu Aufmerksamkeitszielen in einen Zusammenhang zu bringen versuchen, initialisieren die Autoren die Mittelwerte ihres Beobachtungsmodells dabei zunächst. Indem sie daraufhin ebenfalls über ein Maximum-A-Posteriori-Verfahren während des Verlaufs der Besprechungen adaptiert werden, können tatsächlich Verbesserungen der Klassifikationsgenauigkeit nachgewiesen werden.
Wo im Vergleich dazu bisher vorab eingelernte Modellen die individuellen Kopfdrehungsmuster der Personen bei Aufmerksamkeitszuwendungen zu beschreiben versuchten, besteht der Vorteil der Adaption darin, dass nun auf unterschiedliche Interaktionsmuster und Situationen während der Besprechung implizit Rücksicht genommen werden kann. Die Erhöhung der Klassifikationsgenauigkeit bestätigt dabei, dass der Kontext einer Szene offensichtlichen Einfluss auf die Verhaltensweisen nimmt, wie Personen ihren Kopf bei Aufmerksamkeitszuwendungen ausrichten.

Hinzunahme weiterer Kontextbeobachtungen

Die Kopfdrehung konnte bislang erfolgreich als Annäherung zur Blickrichtung eingesetzt werden - zumindest in statischen Szenen mit nicht-bewegten Zielen und, falls der Ansatz geometrischer Natur ist, bei disjunkten Zielanordnungen. Mit weiterführenden Experimenten wurde darüber hinaus jedoch nachgewiesen, dass die Ausrichtung der Kopfdrehungen neben individuellen Verhaltensmustern auch kontextbezogen geschieht und eine Adaption der Mittelwertverteilung der Kopfdrehwinkelballungen unterstützte die Klassifikationsgenauigkeit dabei.

Einen ersten Ausbau eines Systems auf weitere Merkmale als nur die Kopfdrehung, beschrieb Stiefelhagen in [SYW01a]. Stiefelhagen ergänzt seine bisher rein visuelle Schätzung des Fokusziels dabei um auditive Merkmale, indem mit Hilfe eines Neuronalen Netzes und der Historie, wer wann gesprochen hat, der sogenannten *Sprachaktivität* der Personen, Wahrscheinlichkeiten über die Menge der möglichen Aufmerksamkeitsziele ausgegeben werden. Damit kombinierte er die bisher rein visuellen Wahrscheinlichkeiten mit den jetzt zusätzlich auditiven zu einer fusionierten Wahrscheinlichkeitsfunktion. Wie Stiefelhagen dabei erstmalig zeigen konnte, erhöht die audio-visuelle Entscheidung die Genauigkeit und stabilisiert die Klassifikation der Aufmerksamkeitsziele im Vergleich zu rein unimodalen Referenzsystemen.

Ba und Obodez zeigten sich ferner dafür verantwortlich, die Verhaltensweisen der Teilnehmer mit Kontextbeobachtungen aus der Szene zu kombinieren [BO08b, BO08a, BO08c, BHO09, BO09]. Um eine umfassendere Beobachtung der Geschehnisse einzubeziehen, ergänzten sie dabei zunächst die Menge möglicher Aufmerksamkeitsziele um die in den Besprechungen ebenfalls genutzte Tafel und Projektionsleinwand [BO08b, BO08a]. Ferner erweiterten sie ihr Modell dabei äquivalent zu Stiefelhagen [SYW01a] und berücksichtigten fortan die Sprachaktivitäten der jeweiligen Teilnehmer. Ob eine Person spricht oder nicht wird hierbei mit Hilfe von Nahbesprechungsmikrofonen automatisch von der damit aufgezeichneten Energie des Sprachsignals über einen Schwellwert detektiert. Neben der Kopfdrehung, geben die binären Aktivitätszustände aller Teilnehmer als konkatenierter Vektor so einen zusätzlich einen Kontextbezug. Ferner fügten die Autoren die Zeitdauer zwischen den jeweiligen Folienwechseln auf der Leinwand als ergänzende Beobachtung hinzu, die im Vergleich zum Sprachbezug, der fehlenden Artikuliertheit der Leinwand gerecht werden sollte.
Mit [BO08c] gingen Ba et al. den Schritt weg von Hidden Markov Modellen hin zu Bayesnetzen, um der von ihnen beschriebenen Notwendigkeit die beobachtete Situation als Zufallsvariable modellieren zu können, Dienst zu leisten. Darin modellieren die Autoren den Wechsel der Aufmerksamkeit einer Person als Zustandswechsel innerhalb des Netzes, während Kopfdrehung, Sprachaktivität, Folienwechsel auf der Projektionsleinwand und die angenommene Situation jeweilige Zufallsvariablen im Entscheidungsprozess darstellen. Die dabei berücksichtigten Situationen werden mit Stille, Monolog, Dialog und Diskussion zusammengefasst. Mit jeder hiervon legen die Autoren eine A-Priori-Zähldichte über die Menge der Aufmerksamkeitsziele zugrunde, mit welcher Wahrscheinlichkeit einzelne Personen und Objekte unter den gegebenen Umständen betrachtet werden.
Diesen Ansatz vervollständigten sie in [BHO09], worin sie als zusätzliches Merkmal die *visuelle Aktivität* der Besprechungsteilnehmer hinziehen. Mit der visuellen Aktivität wird dabei ein Maß für die Bewegungsstärke, die eine Person während der Aufnahme aufweist, bezeichnet. Sie legen damit die Annahme zugrunde, dass sich eine Person stärker engagiert und in Besprechungen einbringt, wenn sie körperlich aktiver erscheint. Durch ihre verstärkten Bewegungen soll sie demzufolge eine größere Attraktivität darstellen, die Aufmerksamkeit anderer auf sich zu ziehen. Um als Merkmal rasch berechenbar zu sein, wurde dazu keine komplette Analyse der Körperpose aller Personen zu Rate gezogen, sondern lediglich die Differenz zweier aufeinanderfolgender Kameraaufnahmen als Indikator benutzt, wie stark die Erscheinung einer Person vom vorhergehenden Zeitpunkt abweicht.
Die Frage welche Merkmale die Aufmerksamkeit einer Person noch beeinflussen, ist allerdings durch die Beobachtung von menschlichen Verhaltensweisen nicht erschöpfend zu beantworten. Domänenspezifische Szenen, in denen die Aufmerksamkeit eines Menschen zur Ausrichtung eines Roboters oder anderweiligen Systems eingesetzt werden soll, sind oft starr und mensch-

liche Verhaltensmuster sind aufgrund der häufig rein objektbezogenen Aufmerksamkeitsziele nicht anwendbar.
In [HGSR06] gehen Hoffman et al. den Weg, aus der menschlichen Kognition bekannte, saliente Eigenschaften der Objekte, wie Kontrast und Farbe, Kantenstrukturen und Größe als Indikatoren hinzuzuziehen um die Blickrichtung für eine Person, die sich auf ein auf einem Tisch angeordneten Objekt bezieht, nachvollziehen und damit die Aufmerksamkeit eines Roboters steuern zu können. Die Erscheinungseigenschaften aller Gegenstände auf dem Tisch werden hierbei berücksichtigt, um eine Salienzkarte des Erscheinungsbilds des Tischs und der darauf abgelegten Gegenstände zu berechnen. An ihr wird bewertet, wie sich die einzelnen Objekte durch ihre Erscheinung im Umfeld hervorheben und so die Aufmerksamkeit eines Betrachters auf sich ziehen. Indem über Trainingsdaten eine A-Priori-Wahrscheinlichkeit für jedes Objekt eingelernt wird, wie häufig sich jeweilige Personen, je nach dahinterstehender Aufgabe, auf sie beziehen, gelingt es den Autoren den Aufmerksamkeitsfokus mit Hilfe der Kopfdrehung allein nachvollziehen zu können: Die Kopfdrehung spannt dabei das Blickfeld zum Tisch in Form eines Kegelstumpfs auf, der sich bei hinreichender Nähe auf eine nur kleine Schnittfläche mit dem Tisch konzentriert. Die A-Priori-Dichte über die Objekte auf dieser Schnittfläche, reicht schließlich aus, um jenes Objekt erkennen zu können, auf das sich eine Person am wahrscheinlichsten bezieht.
Aufgrund der Tatsache, dass solche Salienzen durch das Geschehen in einer Szene maßgeblich beeinflusst werden, übertragen Schauerte et al. sie in aufmerksame Umgebungen. Damit wollen die Autoren innen liegende Bereiche erkennen können, die für einen außenstehenden Beobachter ein hohes Maß an Information bereitstellen und so für detailliertere Systembeobachtungen vorgeschlagen werden können [SRF10].
Mit [OOF+05] gehen Ou et al. hingegen einen anderen Weg und untersuchen inwieweit die Aufmerksamkeitszuwendung einer Person bei Telepräsenzsystemen davon abhängt, was gerade gesprochen wird und welche Aktionen an den Bildschirmen vollzogen werden. Die Autoren interessiert dabei insbesondere, ob für ein Gegenüber vorhergesagt werden kann, welchen Teil seines Arbeitsbereichs dieser unter den gegebenen Umständen betrachtet, während er mit einer telepräsenten, weiteren Person eine Aufgabe kollaborativ lösen soll. Dabei berücksichtigen sie neben der Blickrichtung als eigentliche Referenz auch Sprachtranskribtionen. So soll aufgrund des Inhalts des Gesprochenen und beobachteter Mausaktivitäten an den Arbeitsrechnern, Rückschlüsse auf die Aktionen und damit auf den wahrscheinlichsten visuellen Aufmerksamkeitsfokus gezogen werden können.

Erste Schritte zu dynamisierten Szenen

Das Schließen der Aufmerksamkeitszuwendung einer Person setzt in allen bisher genannten Arbeiten stets die Kopfdrehung in einen Bezug zur eigentlichen Blickrichtung. Ein Zusammen-

hang konnte dabei insbesondere kognitiv begründet werden, was in statischen Zielpositionierungen zu Referenzwinkelannahmen der Kopfdrehungen bei einer jeweiligen Zuwendung führt. Eine Beantwortung der Fragestellung, inwiefern dieser Zusammenhang auch bei inkonstanten Zielanordnungen bestehen bleibt, steht bislang aber noch aus.
Bis dato ist nur eine Arbeit bekannt, die zum Nachvollziehen der Aufmerksamkeitszuwendung einer Person explizit Bewegungen innerhalb einer Szene zu berücksichtigen versucht. Darin gehen Smith et al. insbesondere der Problemstellung nach, für eine an einem Schaufenster vorbeigehende Person feststellen zu können, ob diese auf ein an der Scheibe angebrachtes Poster blickt [SBGPO06, SBGPO07]. Eine Kamera wurde hierzu hinter der Fensterscheibe platziert, welche Personen beobachtete, die sich außerhalb an der Scheibe vorbei bewegten. Die Autoren stützen sich mit ihrem System dabei auf Arbeiten von Ba et al. [BO06, BO07a] und erweitern deren jeweiligen Ansatz. Hierzu teilen sie die Wegstrecke vor dem Fenster in gleichgroße Segmente auf setzen pro Segment ein dediziertes Beobachtungsmodell der zu erwartenden Kopfdrehungen ein, wenn Personen in diesem Segment auf das Poster blicken. Damit konnte die Schwierigkeit umgangen werden, dass für bewegte Personen unter Umständen kein einheitliches Modell zugrunde gelegt werden kann. Indem stationäre Segmente die Wegstrecke unterteilen, konnte die Problemstellung so auf bewährte Annahmen vereinfacht werden.
Ein Berücksichtigen unrestriktiver Bewegungen und expliziter Dynamiken in einem probabilistischen Verfahrensansatz kam zuerst mit dieser Arbeit als grundlegende Problemstellung auf. Seither findet sich diese aber als wesentlicher Bestandteil in den Evaluationen weiterer Ansätze wieder [BHO09, GBBO09].

2.3. Die visuelle Aufmerksamkeit als Beobachtung in der Situationsanalyse

Aufmerksamkeitszuwendungen helfen beim Verstehen der Aktionen entsprechender Personen. Damit stellt das Wissen, wohin eine Person ihren Blick richtet, eine wichtige Beobachtung dar, um eine Situation im Ganzen verarbeiten und in Bezug auf vorgegebene Fragestellungen untersuchen und modellieren zu können. Im einfachsten Fall dient die visuelle Aufmerksamkeit dazu den Interaktionspartner einer Person feststellen und damit Bezug zu gemeinsamen Interaktionen und Aktivitäten nehmen zu können. Zum Beispiel interessiert in Dialogen die Frage wer der Angesprochene [vTTBE05, SMW+03], bei Vorträgen oder Präsentation hingegen wer der Vortragende und wer Publikum ist [GBBO09].
Anhand der Aufmerksamkeitszuwendungen kann damit nicht nur erkannt werden womit sich Personen beschäftigen; sukzessive kann damit Information zusammengetragen werden, wie die Rollenverteilung in einer Gruppe gegeben ist, wer eine Diskussion dominiert und wer wie starkes Interesse während einer Besprechung oder eines Vortrags bekundet. Quantitativ erfassbare Gütemaße aus soziologischen Forschungen geben dabei eine Möglichkeit, Interaktionsmuster oder Gruppenbildung messbar zu machen. In [HJB+08] stützen sich Hung et al. dabei auf das

Verhältnis wie lange zwei Personen sich gegenseitig anblicken, während sie jeweils selbst sprechen oder nur zuhören. Diese im Englischen genannte *visual dominance ratio*, also das Verhältnis der blickbezogenen Dominanz, gibt dabei Aufschluss wer in einem Dialog dominiert und eine Diskussion unter mehreren so maßgeblich steuert [DE82].

Ähnlich gehen Chen et al. in ihrer Arbeit in [CHF+05] vor, untersuchen dabei allerdings vorwiegend den bewussten Einsatz von Blicken während einer dialogartigen Konversation. Ebenso wie Sprecher und Zuhörer sowie Personen, die das Gespräch unterbrechen und damit der Diskussion beitragen, ihre visuelle Aufmerksamkeit einsetzen, um Gesprächs- oder Diskursabsichten zu offenbaren.

Wie interessiert Gesprächsteilnehmer in Besprechungen auf außenstehende Dritte wirken, untersuchten daneben Gattia-Perez et al. in [GPMB05]. Von Annotatoren holten die Autoren dabei zunächst die für sie offensichtlichen Merkmale ein, die hohes Interesse während bestimmter Situationen bedeuteten. Diese bildeten sie in einem automatischen Erkenner ab, der schließlich anhand der visuellen Zuwendung, Energie und Tonhöhe des Sprachsignals, Sprachaktivitäten aller Teilnehmer, visueller Bewegungsaktivität und Körperorientierung offenkundiges Interesse einer Person nachvollzog.

2.3.1. Weiterführende Kontextbeobachtungen in Besprechungen

Neben der Aufmerksamkeitszuwendung zeichnen sich weitere Kontextmerkmale als hilfreich dafür ab, Situationen einordnen zu können, in denen sich beobachtete Personen befinden. Obwohl sich die Arbeiten dabei nicht mit der eigentlichen Erkennung der Aufmerksamkeit auseinandersetzen, setzen sie diese jedoch gezielt zur Klassifikation von Situationen ein. Aus Gründen der Vollständigkeit soll an dieser Stelle deswegen ein kurzer Überblick gegeben werden.

Die Bewegungsaktivität einer Person während einer Besprechung, wurde bereits von Ba et al. in [BHO09] ausgenutzt, um einen expliziten Reiz zu modellieren, der die Aufmerksamkeit anwesender Personen auch sich zieht. Neben eines einfachen Attraktors bezüglich der Aufmerksamkeit, sehen andere Arbeiten darin einen Ausdruck der bestehenden Rollenverteilung innerhalb einer Gruppe. So setzen neben den genannten Verfahren von Hung et al., auch andere Ansätze auf die visuelle Aktivität der Protagonisten, jedoch um den hauptsächlichen Sprecher innerhalb einer Gruppe ausmachen zu können [HF08, FHY09]. In wiederum anderen Arbeiten wird darin eher dominantes Auftreten der Protagonisten interpretiert [HJY+07, JHYGP08, JHYGP09, HHGP08].

Dominanz in Besprechungen stellt prinzipiell ein wichtiges Merkmal dar, um auf die Rollenverteilung und hierarchische Ordnungen schließen zu können. Dass für hinreichende Detektionen nicht immer visuelle Beobachtungen notwendig sind, zeigen Arbeiten die aufgrund rein sprachbasierter Merkmale, wie Sprachhäufigkeit oder Menge der gestellten Fragen, Rückschlüsse zie-

hen [RH05, HGPHF08]. Einen umfassenden Überblick über verwandte Arbeiten geben dabei Hung et al. in [HGP08].

Neben der Fragestellung nach dem maßgeblichen Sprecher ist die Frage nach erkennbaren Aktivitäten und Handlungen im beobachteten Geschehen Grundlage vieler Veröffentlichungen [MBGP+03, MGPB+05, ZGPB+04, ZGPBM06, OHG02, BMR05].
Mit [MBGP+03, MGPB+05] gehen beispielsweise McCowan et al. der Aufgabe nach, zwischen Monologen, einer Präsentation an der Leinwand, einer Präsentation vor einer Tafel, einer Diskussion aller Besprechungsteilnehmer und einer Situation zu unterscheiden, in der die Besprechungsteilnehmer Notizen auf dem Tisch machen. Das Vorgehen der Autoren setzt dabei neben auditiven Merkmalen wie Sprachenergie und -höhe auch auf die visuell erfasste Position und Größe des Gesichts und der Hände der beteiligten Personen und benutzt diese als Beobachtungen um mit Hilfe eines Hidden Markov Modells über die verschiedenen Situationszustände zu evaluieren.
Mit ähnlichen Beobachtungen und vergleichbarem Ansatz gehen Zhang et al. in [ZGPB+04, ZGPBM06] vor, unterscheiden jedoch zwischen Aktionen auf Individuenebene (Person spricht, schreibt, ist inaktiv), die in einer ersten Schicht mit Hidden Markov Modellen erkannt werden und einer darauf aufbauenden Gruppenaktivität (Diskussion, Monolog, Präsentation), die die Gesamtsituation beschreibt und in einer zweiten Schicht von Hidden Markov Modellen darauf aufbaut.
Diesen Ansatz zweischichtiger Markovmodelle verfolgen auch Oliver et al. in [OHG02]. Die Autoren konzentrieren sich hingegen aber auf Büroaktivitäten, wie Telefongespräche und Dialoge zweier Personen. Statt mit den Modellschichten unterschiedliche Aktivitäten abzubilden, baut das Verfahren dabei auf unterschiedlich große Zeitfenster, über die die Merkmale evaluiert werden. Aktivitäten setzen sich dementsprechend sukzessive über die unterschiedlich andauernden Zeitspannen zusammen.
Mit [BMR05] gehen Brdiczka et al. im Gegensatz dazu lediglich der Fragestellung nach, wie Menschengruppen und deren Zusammensetzung in einer Szene rein auditiv erkannt werden können. Als Motivation sehen die Autoren, die durch Gruppenbildung erkennbaren Hinweise auf unterschiedliche, gleichzeitige Situationen in einer Szene. Personen, die gemeinsam an denselben Aktivitäten arbeiten werden darin zusammengefasst.
Einen ersten Versuch dabei den in einer Gruppe erkennbaren Zusammenhalt beziehungsweise das Ausmaß der Zusammenarbeit zwischen mehreren Menschen zu erkennen, stellt Hung in ihrer aktuellen Studie vor [HGP10].

Für einen hierüber hinausgehenden Überblick der Interaktions- und Situationserkennung, sei an dieser Stelle auf eine Veröffentlichung von Gattica-Perez et al. verwiesen [GPZB05, GP06]. Sie gibt dabei eine hinreichende Zusammenfassung über die in der Literatur gängigen Methoden

und benutzten Merkmalsbeobachtungen - auch unter Berücksichtigung der visuellen Aufmerksamkeit.

3. Bestimmen der Kopfdrehung

Um festzustellen wohin eine Person blickt, beobachtet man idealerweise ihre Pupillen und folgt der damit verbundenen Blickrichtung. Hat man diese Möglichkeit nicht, vermittelt zumindest die Kopfdrehung Aufschluss über die Ausrichtung des Blickfelds der Person. Das ist zum Beispiel immer dann der Fall, wenn die Sicht zu den Augen verdeckt wird, die Kamera so angebracht ist, dass sie nur die Hinterseite des Kopfes erfasst oder die Auflösung der Kamera zu gering und die Entfernung zum Gesicht zu groß ist, um eine detaillierte Aufnahme der Augen zu erhalten. Im Fall dieser Arbeit treffen alle genannten Möglichkeiten aufeinander: die uneingeschränkte Bewegungsfreiheit führt zu Verdeckungen durch andere Personen und Objekte, die distante Anbringung der Kameras zu nachteiligen, weil unscharfen Augenmotiven und die Benutzung mehrerer Kameraansichten zu gleichzeitigen Profil- und Hinterkopfansichten.
In diesem Kapitel wird das in dieser Arbeit entwickelte System beschrieben, das die Kopfdrehung einer Person mit Hilfe mehrerer Kameraansichten schätzt. Die Kopfdrehung soll zur Bestimmung des Blickfelds genutzt werden, das im Anschluss zur Deduktion des wahrscheinlichen Aufmerksamkeitsfokus eingesetzt wird. Weil zur Bestimmung der Ausrichtung des Blickfelds die Rotation um die eigene Sichtachse irrelevant ist, soll der Kopf lediglich in seiner dreidimensionalen Position und Maße sowie horizontaler und vertikaler Drehung erkannt und nachverfolgt werden. Die Motivation hinter dem Nutzen verschiedener Kameraansichten liegt darin, die distante Beobachtung und den damit verbundenen Auflösungsverlust auszugleichen und zum anderen mit Verdeckungen und Hinterkopfansichten umgehen zu können. Das System soll deswegen flexibel auf ein Hinzufügen weiterer oder eine Reduktion der verfügbaren Ansichten reagieren und infolgedessen kein erneutes Einlernen benötigen.

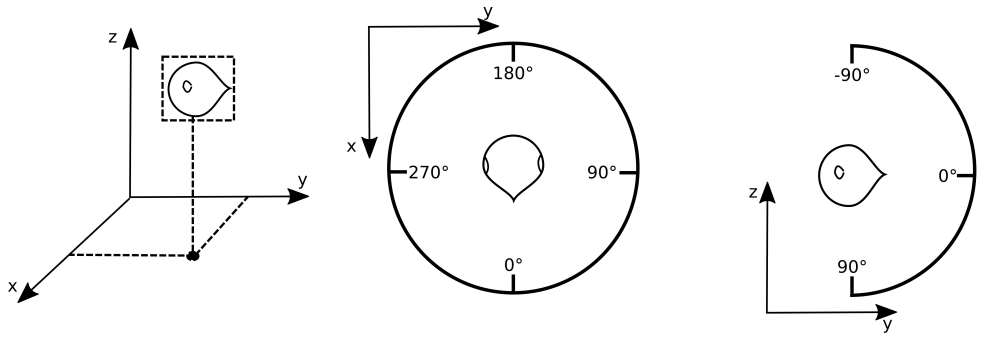


Abb. 3.1.: Schematischer Aufbau des Zustandsraums. Dargestellt ist die Position des Kopfs sowie dessen Größe (links), die horizontale (mittig) und vertikale Kopfdrehung (rechts).

3.1. Visuelles Tracking via Bayes'schem Filter

Unter *visuellem Tracking* versteht man das Erkennen und rekursive Verfolgen von Objekten beziehungsweise Motiven in Kamerabildsequenzen [PV05]. Das zeitliche Nachvollziehen der Kopfposition, der Kopfgröße und dessen Drehwinkel entspricht ebenfalls diesem Prinzip. Es handelt sich um ein dynamisches System, dessen Zustand sich stetig ändern kann, den wir aber zu diskreten Zeitpunkten messen möchten, jedoch nicht direkt können. Stattdessen liegen Beobachtungen vor, hier in Form verrauschter Kamerabilder, die Rückschlüsse auf diesen *verborgenen* Zustand erlauben. Der Beobachtungsvektor weist dabei allerdings häufig eine niedrigere Dimensionalität auf als der eigentliche Systemzustand [AMGC02].

Zum Verständnis soll in diesem Abschnitt die grundelegende Vorgehensweise visuellen Trackings unter den genannten Bedingungen beschrieben werden. Dabei stützt sich die Ausführung in großen Teilen auf Veröffentlichungen von Nickel [Nic08], Arulampalam et al. [AMGC02] und Canton-Ferrer [CF09].
In diesem Kontext sei mit $\mathbb{S} \subseteq \mathbb{R}^8$ der Zustandsraum des Systems definiert, welcher sich aus Position, Größe und Rotationswinkel des Kopfs zusammensetzt. Im Speziellen sei hierbei mit $\mathbf{s} = (\mathbf{x}, \mathbf{r}, \theta_{pan}, \theta_{tilt}) \in \mathbb{S}$ der Zustandsvektor des Kopfs gegeben, wobei mit $\mathbf{x} \in \mathbb{R}^3$ die Position, $\mathbf{r} \in \mathbb{R}^3$ die Größe und $\theta_{pan} \in [0°, 360°), \theta_{tilt} \in [-90°, 90°]$ die Drehwinkel eines Kopfkandidaten bezeichnet werden. Der Verlauf der Kopftrajektorie und -rotation kann hiermit durch eine Sequenz von Systemzuständen $\{\mathbf{s}_t, t = 1 \dots T\}$ beschrieben werden, deren Entwicklung durch eine nicht-lineare und zunächst unbekannte Funktion $f_t(\cdot, \cdot)$ zusammengefasst werden kann:

$$\mathbf{s}_t = f_t(\mathbf{s}_{t-1}, \mathbf{v}_{t-1}) \tag{3.1}$$

Ein aktueller Zustand ergibt sich damit aus seinem Vorgänger und einem anzuwendenden Rauschterm $\mathbf{v}_{t-1}$.
Im Gegenzug liegen Beobachtungen $\{\mathbf{o}^t, t = 1 \dots T\}$ vor, für die gilt, dass sie einer Abbildung $h_t(\cdot, \cdot)$ der jeweiligen Systemzustände entsprechen:

$$\mathbf{o}_t = h_t(\mathbf{s}_t, \mathbf{w}_t) \tag{3.2}$$

$\mathbf{w}_t$ beschreibt das den Beobachtungen unterliegende Rauschen. Ziel ist es, zum Zeitpunkt t Rückschlüsse auf den eigentlichen Zustand $\mathbf{s}_t$ zu ziehen, wenn die Beobachtungssequenz $\mathbf{o}_{1:t}$ seit Beginn vorliegt [AMGC02]. Die Bewertung eines Zustands kann so als rekursive Berechnung anhand aller vorliegenden Beobachtungen aufgefasst werden. Mit einer gegebenen, initialen A-Priori-Funktion $p(\mathbf{s}_0)$, ergibt sich die Wahrscheinlichkeitsfunktion $p(\mathbf{s}_t|\mathbf{o}_{1:t})$ aus einer Bayesschen Perspektive damit durch kontinuierliche Vorhersage der Zustandsbewertung und Anpassen anhand aktueller Beobachtungen. In der Vorhersage wird hierzu Wissen über die zu-

grunde liegende Prozessentwicklung aus Gleichung 3.1 ausgenutzt und damit eine A-Priori-Wahrscheinlichkeitsfunktion über den Systemzustand mit Hilfe des Chapman-Kolmogorov-Integrals gewonnen [AMGC02]:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \int p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})\mathrm{d}s_{t-1} \tag{3.3}$$

Dabei wurde dem Systemprozess die Markoveigenschaft zugrunde gelegt, dass der aktuelle Zustand lediglich vom Vorangehenden abhängt. Der zweite Integralterm $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$ bezeichnet die A-Posteriori-Dichtefunktion zum vorigen Zeitschritt, die sich rekursiv bis zu Beginn des Trackingverfahrens durchpropagiert und von der ausgegangen wird, dass sie zum aktuellen Zeitpunkt vorliegt.

Die Vorhersage erlaubt, aus dem mutmaßlichen Wissen über die Zustandsverteilung des Systems im vorigen Schritt, Rückschlüsse auf die aktuelle Verteilung zu ziehen. Mit einer neu eintreffenden Beobachtung $\mathbf{o}_t$, folgt im Aktualisierungsschritt so schließlich die Anpassung des Vorhergesagten [AMGC02]:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t}) = \frac{p(\mathbf{o}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{o}_{1:t-1})}{p(\mathbf{o}_t|\mathbf{o}_{1:t-1})} \tag{3.4}$$

Zu jedem beliebigen Zeitpunkt t wäre schließlich der Erwartungswert $\mathbb{E}[\mathbf{s}_t|\mathbf{o}_{1:t}]$ momentaner Schätzwert des Systemzustands [CF09].

Die rekursive Propagierung der A-Posteriori-Wahrscheinlichkeitsfunktion bildet das Gerüst einer optimalen Bayesschen Herangehensweise, erlaubt aber nur in einzelnen Fällen eine geschlossene Darstellung und analytische Berechenbarkeit [AMGC02]. Legt man dem Systemprozess in Gleichung 3.1 Linearität und der A-Posteriori-Funktion eine Normalverteilung zugrunde, stellt der Kalman-Filter hierzu eine bewährte Lösungsmethode dar [AMGC02]. Mit Kopfbewegungen und -drehungen in einer realen Umgebung, kann dem Systemprozess allerdings keine faktische Linearität zugrunde gelegt werden. Das resultiert in einer A-Posteriori-Funktion, die sich nicht ohne weiteres in geschlossener Form angeben lässt. Sequentielle Monte-Carlo-Methoden legen keine einschränkende Funktionsfamilie zugrunde und umgehen solche Einschränkungen, indem sie die Funktion nicht berechnen sondern lediglich approximieren. Dabei wird die Funktion durch eine Menge gewichteter Diracimpulse beschrieben, die bei ausreichender Anzahl eine hinreichend genaue Annäherung der eigentlichen Funktion erlauben. Im Folgenden soll deshalb näher auf die in der Arbeit eingesetzte Monte-Carlo-Methodik eingegangen werden.

3.1.1. Bayesian Sequential Importance Sampling

Sequential Importance Sampling ist ein Monte-Carlo-Verfahren, das eine numerische Annäherung für die in Gleichung 3.3 beziehungsweise 3.4 beschriebene Rekursion nach Bayes beschreibt. Mit einer solchen Annäherung muss kein Wissen über die eigentliche Funktionsfamilie der Wahrscheinlichkeitsfunktion vorliegen, hingegen kann ihr stochastisches Mittel bestimmt werden [Nic08]. Dabei wird die Dichtefunktion durch eine Menge mit $N_{\mathscr{P}}$ gewichteten, sogenannten *Partikeln* beschrieben: $\mathscr{P} = \{(\mathbf{s}_t^i, \pi_t^i), i = 1 \dots N_{\mathscr{P}}\}$ - die Funktion wird damit abgetastet und in ihrer Form erfasst. Für die Gewichte gilt dabei aus Normalisierungs- und Konvergenzgründen $\sum_i^{|\mathscr{P}|} \pi^i = 1$. Nach dem Gesetz der großen Zahlen erlaubt eine solche Approximation eine hinreichend genaue Beschreibung jeder Funktionsform. Insbesondere entspricht der durch die Gewichtung erhaltene Erwartungswert im Mittel dem der eigentlich anzunähernden A-Posteriori-Wahrscheinlichkeit. Die Partikel können dabei als Diracimpulse interpretiert werden, deren Position auf die des entsprechenden Partikels verschoben wird. Die Gewichtung π_t^i gibt dabei die Amplitude des Impuls wieder und beziffert damit die Ausprägung der Funktion an der jeweiligen Stützstelle $\mathbf{s}_t^i$.

Entsprechend [CF09] und [AMGC02] sei im Folgenden die in Gleichung 3.4 beschriebene A-Posteriori-Dichte durch die genannte Menge an Diracimpulsen gegeben:

$$p(\mathbf{s}_{1:t}|\mathbf{o}_{1:t}) \approx \sum_{i=1}^{|\mathscr{P}|} \pi_t^i \delta(\mathbf{s}_{1:t} - \mathbf{s}_{1:t}^i) \tag{3.5}$$

Die Bewertung der Stützstellen geschieht über das Prinzip des *Importance Samplings* [DGA00]: Die Stützstellenauswahl geschieht dabei entsprechend $\mathbf{s}^i \sim q(\mathbf{s})$, wobei mit $q(\cdot)$ eine A-Priori-Dichtefunktion bezeichnet wird, die Abtaststellen im durch $\mathbf{s}$ aufgespannten Zustandsraum probabilistisch vorschlägt. Darauf bezugnehmend, erfolgt die Berechnung der Partikelgüte π_t^i proportional zum Verhältnis des Dichtewerts und der Stützstellenrelevanz nach $q(\cdot)$ [AMGC02]:

$$\pi^i \propto \frac{p(\mathbf{s}_{1:t}|\mathbf{o}_{1:t})}{q(\mathbf{s}_{1:t}^i|\mathbf{o}_{1:t})} \tag{3.6}$$

Die Partikeldichte $p(\mathbf{s}_{1:t}|\mathbf{o}_{1:t})$ aus Gleichung 3.5 lässt sich damit wie folgt vereinfachen:

$$p(\mathbf{s}_{1:t}|\mathbf{o}_{1:t}) \approx \sum_{i=1}^{|\mathscr{P}|} \pi^i \delta(\mathbf{s}_{1:t} - \mathbf{s}_{1:t}^i) \tag{3.7}$$

Aufgrund der Markov-Annahme und geeigneter Wahl von $q(\cdot)$ kann davon ausgegangen werden, dass sich $q(\mathbf{s}_{1:t}^i|\mathbf{o}_{1:t})$ aus Gleichung 3.6 durch

$$q(\mathbf{s}_{1:t}|\mathbf{o}_{1:t}) = q(\mathbf{s}_t|\mathbf{s}_{1:t-1}, \mathbf{o}_{1:t}) q(\mathbf{s}_{1:t-1}|\mathbf{o}_{1:t-1}) \tag{3.8}$$

faktorisieren lässt und die Partikelgewichtung aus Gleichung 3.6 dadurch schließlich auflöst zu:

$$\pi^i \propto \frac{p(\mathbf{s}_{1:t}|\mathbf{o}_{1:t})}{q(\mathbf{s}_t|\mathbf{s}_{1:t-1},\mathbf{o}_{1:t})q(\mathbf{s}_{1:t-1}|\mathbf{o}_{1:t-1})} \tag{3.9}$$

Formuliert man den Zähler letztlich auch nach Bayes um, ergibt sich aus

$$p(\mathbf{s}_{1:t}|\mathbf{o}_{1:t}) \propto p(\mathbf{o}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{1:t-1}|\mathbf{o}_{1:t-1}) \tag{3.10}$$

und Gleichung 3.9 schließlich ein iteratives Aktualisieren der Partikelgewichte zu jedem Zeitschritt anhand [AMGC02]:

$$\begin{aligned}\pi_t^i &\propto \frac{p(\mathbf{o}_t|\mathbf{s}_t^i)p(\mathbf{s}_t^i|\mathbf{s}_{t-1}^i)}{q(\mathbf{s}_t^i|\mathbf{s}_{1:t-1}^i,\mathbf{o}_{1:t})}\frac{p(\mathbf{s}_{1:t-1}^i|\mathbf{o}_{1:t-1})}{q(\mathbf{s}_{1:t-1}^i|\mathbf{o}_{1:t-1})} \\ &= \frac{p(\mathbf{o}_t|\mathbf{s}_t^i)p(\mathbf{s}_t^i|\mathbf{s}_{t-1}^i)}{q(\mathbf{s}_t^i|\mathbf{s}_{1:t-1}^i,\mathbf{o}_{1:t})}\pi_{t-1}^i\end{aligned} \tag{3.11}$$

Ein konsequentes Durchpropagieren der Markovannahme führt ähnlich zu Gleichung 3.3 und 3.8 schließlich zur Reduktion der Historie auf den Vorgängerzustand, so dass abschließend die Gewichtung aus Gleichung 3.11 zusammengefasst werden kann zu

$$\pi_t^i \propto \pi_{t-1}^i \frac{p(\mathbf{o}_t|\mathbf{s}_t^i)p(\mathbf{s}_t^i|\mathbf{s}_{t-1}^i)}{q(t_t^i|\mathbf{s}_{t-1}^i,\mathbf{o}_t)} \tag{3.12}$$

3.1.2. Sampling Importance Resampling

Mit dem Sampling Importance-Vorgehen werden die den Partikel entsprechenden Stützstellen bewertet. Die daraus erhaltene Partikelgewichtung approximiert die anzunähernde A-Posteriori-Dichtefunktion des Systemzustands nach Gleichung 3.4. Wie Doucet et al. angeben [DGA00], kann jedoch nachgewiesen werden, dass das kontinuierliche Aktualisieren der Partikelgewichte dazu führt, dass die Varianz der Gewichte stetig ansteigt. Infolgedessen konvergieren die normalisierten Partikelgewichte gegen 0 und alle bis auf sehr wenige Partikel tragen fortan fast nichts mehr zur Approximation der eigentlichen Funktion bei. Parallel aber wird ein Großteil der Rechenzeit dazu eingesetzt ineffiziente Samples zu aktualisieren.
Umgangen werden kann das Problem durch Sampling Importance *Resampling*. Dabei werden Partikel, die der Funktionsapproximation wenig beisteuern und damit degenerierenden Samples entsprechen, verworfen. Aus den übrig gebliebenen wird ein neuer Satz Partikel generiert, der als Vorschlagfunktion $q(\cdot)$ die gegenwärtige A-Priori-Dichte $p(\mathbf{s}_t|\mathbf{s}_{t-1}^i)$ zugrunde legt. Für den neuen Satz Partikel gilt somit also $\mathbf{s}^i \sim p(\mathbf{s}_t|\mathbf{s}_{t-1}^i)$. Indem die A-Priori-Dichte der Vorschlagfunktion gleichgesetzt wird, werden relevante Stützstellen wegen ihrer bereits hohen Gewichtung häufiger ausgewählt, während Zustandsannahmen mit niedriger Gewichtung seltener zum

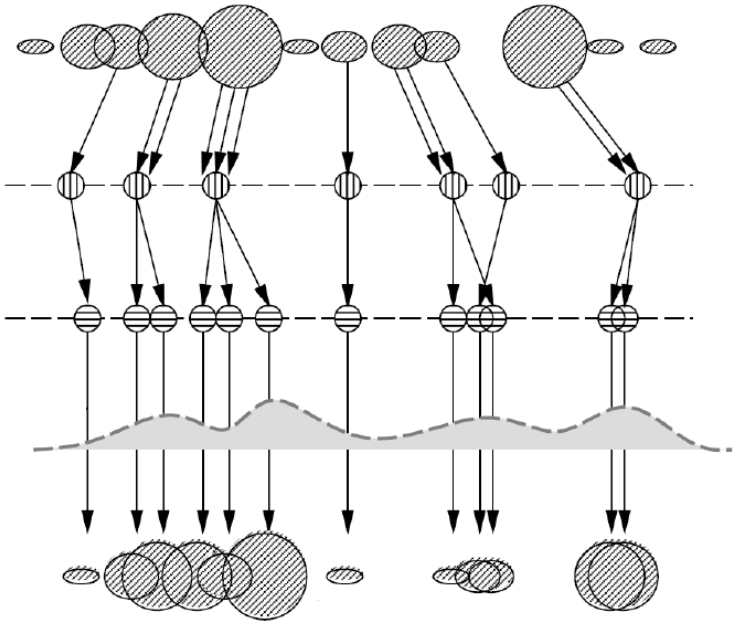


Abb. 3.2.: Schematischer Ablauf des Vorgehens beim Sampling Importance Resampling (Quelle: [CF09]): Aus der Partikelverteilung zum Zeitpunkt $t-1$ wird eine neue Menge gezogen. Diese wird anschließend entsprechend der zugrunde gelegten Prozessdynamik propagiert und mit durch einen additiven Rauschterm diffundiert. Mit gegebenen Beobachtungen werden die Zustandshypothesen bewertet, womit die finalen Partikelgewichtungen der gezogenen Menge berechnet werden.

Zuge kommen. Damit werden diejenigen Partikel beibehalten, die die Form der zu approximierenden Funktion deutlich unterstützen. Wird somit in jedem Sequenzschritt der Satz Partikel neu erstellt, können die entsprechenden Gewichtungen gleichverteilt initialisiert werden: $\pi^i_{t-1} = |\mathscr{P}|^{-1}$. Gleichung 3.11 kann damit weiter reduziert werden zu

$$\begin{aligned} \pi^i_t &\propto \pi^i_{t-1} \frac{p(\mathbf{o}_t|\mathbf{s}^i_t)p(\mathbf{s}^i_t|\mathbf{s}^i_{t-1})}{q(\mathbf{s}^i_t|\mathbf{s}^i_{t-1},\mathbf{o}_t)} \\ &= \frac{1}{|\mathscr{P}|} \frac{p(\mathbf{o}_t|\mathbf{s}^i_t)p(\mathbf{s}^i_t|\mathbf{s}^i_{t-1})}{p(\mathbf{s}_t|\mathbf{s}^i_{t-1})} \\ &= p(\mathbf{o}_t|\mathbf{s}^i_t) \end{aligned} \tag{3.13}$$

Letztendlich lässt sich das Aktualisieren der Partikelgewichte damit auf ein rein proportionales Anpassen hinsichtlich der aktuellen und vorliegenden Beobachtung $p(\mathbf{o}_t|\mathbf{s}^i_t)$ vereinfachen [AMGC02].

3.1.3. Propagierung der Prozessdynamik

Im Vorhersageschritt des beschriebenen Verfahrens, werden die gezogenen Partikel der zugrunde liegenden Prozessdynamik unterzogen. Die Zustandshypothesen, die den Abtastungen ent-

sprechen, werden entlang des angenommenen Prozessverhaltens propagiert. Ein sich geradeaus bewegendes Objekt mit gleichbleibender Geschwindigkeit würde sich somit linear weiter fortbewegen und seine bisherige Trajektorie beibehalten. Partikel, die diese Zustandshypothese unterstützen, würden im Resampling-Schritt weitergeführt werden und ihre jeweilige Zustandsbeschreibung entlang ihrer bisherigen Annahme durch den beigefügten Rauschterm diffundiert werden. Das Verhaltensmuster des zu modellierenden Objekts ist jedoch vom jeweiligen Anwendungsfall abhängig. Häufig kann eine einfache Prozessdynamik zugrunde gelegt werden, dies schränkt aber gleichzeitig die Abdeckung des Zustandsraums auf die zu erwartenden Hypothesen ein und erlaubt keine flexible Anpassung an unerwartete Beobachtungen.
Wie Menschen Kopfrotationen einsetzen, hängt stark vom Kontext der Situation und des Interaktionsverhaltens der Person ab. Diese Faktoren erschweren das Einbeziehen einer vorgegebenen und festen Prozessdynamik. Die Stärke von Partikelfiltern zeigt sich jedoch gerade in ihrer Flexibilität sowohl multimodale A-Posteriori-Dichtefunktionen approximieren zu können als auch damit keiner Einschränkung hinsichtlich einer vordefinierten Zustandsentwicklung unterliegen zu müssen. In Anlehnung an [CF09] soll deswegen im Rahmen dieser Arbeit auf diesen Schritt verzichtet werden. Stattdessen werden die Zustandshypothesen lediglich mit normalverteiltem Rauschen diversifiziert, um explizit Varianz in die Abdeckung des Zustandsraums einfließen zu lassen. Mit gegebener Kovarianzmatrix $\mathbf{C}_{\mathscr{D}}$ gilt somit:

$$\mathbf{s}_t^i = \mathbf{s}_{t-1}^i + n, \qquad n \sim \mathscr{N}(0, \mathbf{C}_{\mathscr{D}}) \tag{3.14}$$

3.1.4. Das Beobachtungsmodell

Mit der in Gleichung 3.13 aufgeführten Partikelgewichtung, werden die Zustandshypothesen anhand einer Wahrscheinlichkeitsfunktion $p(\mathbf{o}_t|\mathbf{s}_t^i)$ bewertet. Sie nähern so die Ausprägung der A-Posteriori-Dichte an ihrer jeweiligen Zustandshypothese an und bilden in der Menge deren komplexe, nicht berechenbare Form nach. Damit wird ein Zusammenhang zwischen einer vorhandenen Beobachtung und der Zustandsannahme hergestellt. Die darauf bezogene Wahrscheinlichkeitsfunktion ist jedoch in der Regel unbekannt oder, falls doch gegeben, oft nur komplex zu berechnen. Deshalb wird häufig eine Annäherung $w(\mathbf{o}_t, \mathbf{s}_t^i) : \mathbb{S} \mapsto [0,1]$ vorgenommen, in der ein Beobachtungsmodell zugrunde gelegt wird das die Partikel proportional zu ihr gewichtet, insbesondere aber schnell berechnet werden kann [CF09]:

$$\pi_t^i := w(\mathbf{o}_t, \mathbf{s}_t^i) \tag{3.15}$$

Unter der Beobachtung $\mathbf{o}_t$ versteht man das aktuelle Kamerabild. Auf diesem werden Merkmale $m(\mathbf{o}_t)$ berechnet, welche anschließend zur Bewertung der Stützstellenhypothesen zugrunde gelegt werden. Mit verschiedenen solcher Merkmale erhält man die Menge $\{(m^j(\mathbf{o}_t), w^j(\cdot,\cdot)), j \in \mathbb{N}\}$, wobei die Tupel jeweils aus den Merkmalen $m^j(\mathbf{o}_t)$ und ihnen zugehörigen Bewertungs-

funktionen $w^j(m^j(\mathbf{o}_t), \mathbf{s}_t^i)$ bestehen. Setzt man die Unabhängigkeit der Merkmale voraus, ergibt sich die endgültige Bewertung eines Zustands damit aus der Multiplikation aller Einzelbewertungen:

$$\pi_t^i = \prod_j w^j(m^j(\mathbf{o}_t), \mathbf{s}_t^i) \tag{3.16}$$

Mit der Multiplikation wird die Eigenschaften eingebracht, dass der Ausfall eines einzelnen Merkmals, die gesamte Bewertung gegen 0 gehen lässt, auch wenn die übrigen Merkmale gegensätzlich klassifizieren. Nach [KHDM98] darf die Gesamtgewichtung stattdessen auch als Summenregel implementiert werden, wenn sich die A-Posteriori-Wahrscheinlichkeit nur marginal von der A-Priori-Wahrscheinlichkeit unterscheidet. Im Rahmen des Bayesschen Trackings kann diese Annahme angewandt werden, weil mit genügend hoher Framerate aufeinanderfolgender Kamerabilder die Zustandsveränderung des Systems nur geringfügig ausfällt [Nic08]. Prinzipiell geschieht die Akkumulierung der Einzelbewertungen

$$\pi_t^i = \sum_j \lambda^j w^j(m^j(\mathbf{o}_t), \mathbf{s}_t^i), \qquad \sum_j \lambda^j = 1 \tag{3.17}$$

dabei statisch, Arbeiten wie [Nic08] setzen sich jedoch damit auseinander die komponentenweise Gewichtung der Merkmale in der Gesamtbewertung dynamisch anzupassen und dabei mit Hilfe vordefinierter Qualitätsmetriken Einzelmerkmale stärker oder schwächer einzubeziehen. Für eine genauere Betrachtung dieser sogenannten *Demokratischen Integration* unterschiedlicher Merkmale, sei an dieser Stelle auf die genannte Arbeit verwiesen. Im Rahmen dieser Arbeit sollen die Bewertungsstrategien 3.16 und 3.17 lediglich bei fest vorgegebener Gewichtung gegenüber gestellt und evaluiert werden. Um dabei korrektweise auch bei der multiplikativen Fusion eine unterschiedliche Gewichtung der Merkmale zu erlauben, wird Gleichung 3.16 hierzu um eine exponentielle Gewichtung der Faktoren erweitert:

$$\pi_t^i = \prod_j (w^j)^{\lambda^j}(m^j(\mathbf{o}_t), \mathbf{s}_t^i), \qquad \sum_j \lambda^j = 1 \tag{3.18}$$

Kamerabezug

Da im Rahmen dieser Arbeit nicht nur eine, sondern mehrere Kameraansichten zur Verfügung stehen, muss die Partikelevaluation alle zur Verfügung stehenden Ansichten gleichermaßen miteinbeziehen. Für $N_{\mathcal{O}}$ Kamerabilder ergibt sich so die Menge $\mathcal{O}_t = \{\mathbf{o}_t^c, c \in N_{\mathcal{O}}\}$ vorhandener Beobachtungen. Damit wird ein Anpassen der Fusionsstrategien erforderlich, so dass schließlich

über die erweiterte Menge der Merkmale $\{m^j(\mathscr{O}_t)\}$ in der Stützstellenbewertung akkumuliert wird. Für die multiplikative Merkmalsfusion aus Gleichung 3.18 gilt damit:

$$\pi_t^i = \prod_j (w^j)^{\lambda^j}(m^j(\mathscr{O}_t), \mathbf{s}_t^i) = \prod_j \sum_c^{N_\mathscr{O}} (w^j)^{\lambda^j}(m^j(\mathbf{o}_t^c, \mathbf{s}_t^i)) \tag{3.19}$$

Äquivalent ist für die additive Strategie vorzugehen.
Für eine Anwendung desselben Bewertungsschemas $w^j(m^j(\mathbf{o}_t^c, \mathbf{s}_t^i))$ in verschiedenen Kameraansichten, müssen die Zustandsbeschreibungen vom Weltbezug in das jeweilige Kamerasystem abgebildet werden. Mit der in Anhang A.1 beschriebenen Kalibrierungsmatrix $\mathbf{K}^c$, lässt sich damit eine Abbildungsfunktion $g^c(\cdot) : \mathbb{S} \mapsto \mathbb{N}^2 \times \mathbb{N}^2$ definieren, durch die die raumbezogene Position x und Maße r des Kopfs auf eine lokale Rechteckregion in der jeweiligen Bildebene von Kamera c projiziert wird, die durch ihre zweidimensionale Position und Größe angegeben wird. Die Drehwinkelkomponente bleibt durch die Projektion dabei aber zunächst unberücksichtigt. Im Folgenden soll deshalb zusätzlich für die horizontale Winkelkomponente θ_{pan} eine Berücksichtigung zum Bezugsystem der Kamera beispielhaft beschrieben werden. Für die vertikale kann im Anschluß analog vorgegangen werden.

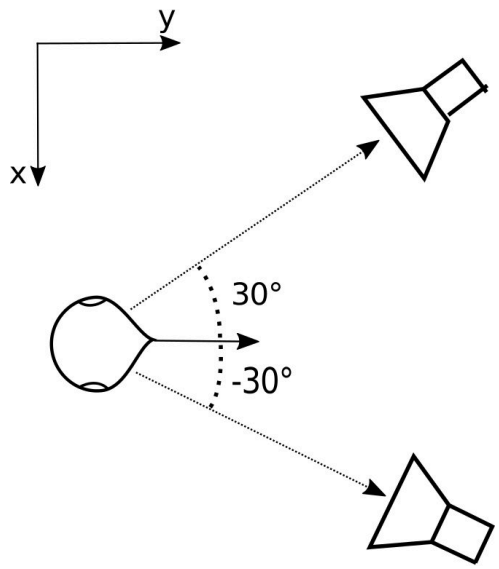


Abb. 3.3.: Beispielhafte Darstellung einer horizontalen Kopfdrehung. Aus verschiedenen Blickwinkeln erscheint der Kopfausschnitt trotz desselben Drehwinkels verschieden orientiert: Während die obere Kamera Drehung von 30° beobachten kann, nimmt die untere ein um $-30°$ gedrehtes Kopfmotiv auf.

Dieselbe Kopfdrehung erscheint aus unterschiedlichen Blickwinkeln wie verschiedene Drehmotive: Während eine Kamera einen ins Profil gedrehten Kopf beobachtet, erscheint derselbe aus einer anderen Ansicht eventuell frontal zugedreht. Derselbe Drehwinkel führt so je nach Ansicht zu unterschiedlichen Darstellungen und erfordert verschiedenartige Winkelbeschrei-

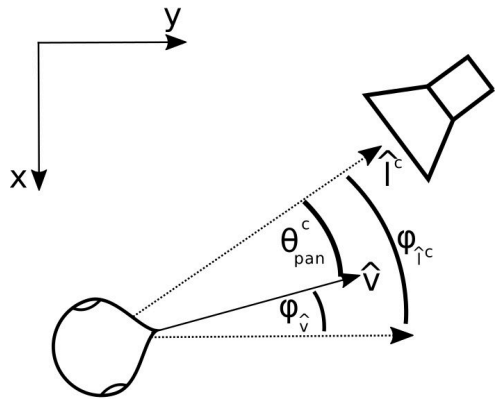


Abb. 3.4.: Die horizontale Kopfdrehung wird im Bezug zum Kamerasystem bestimmt. Damit kann Rücksicht auf die verschiedenen Ansichten genommen werden, in denen dieselbe Kopfdrehung unterschiedlich beobachtet wird.

bungen. Sei im Folgenden, wie in Abbildung 3.4 dargestellt, $\mathbf{v} \in \mathbb{R}^3$ der Richtungsvektor des Blickfelds im Weltbezug sowie $\mathbf{l}^c \in \mathbb{R}^3$ der Vektor der Sichtachse von Kamera c zum dreidimensionalen Kopfmittelpunkt. Ferner sei mit $\hat{\mathbf{v}}$ bzw. $\hat{\mathbf{l}}^c$ die orthogonale Projektion des Vektors $\mathbf{v}$ beziehungsweise $\mathbf{l}^c$ auf die Bodenebene des Raums gegeben. Die horizontalen Drehwinkel der beiden Vektoren $\mathbf{v}$ und $\mathbf{l}^c$ lassen sich jeweils durch die Winkelkomponente ihrer Polarkoordinatendarstellung $(r_{\hat{\mathbf{v}}}, \varphi_{\hat{\mathbf{v}}})$ und $(r_{\hat{\mathbf{l}}^c}, \varphi_{\hat{\mathbf{l}}^c})$ ausdrücken:

$$\varphi_{\hat{\mathbf{v}}} = \begin{cases} \arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right) & \text{für } \hat{v}_1 > 0 \text{ und } \hat{v}_2 > 0 \\ \arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right) + 2\pi & \text{für } \hat{v}_1 > 0 \text{ und } \hat{v}_2 < 0 \\ \arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right) + \pi & \text{für } \hat{v}_1 < 0 \\ \frac{\pi}{2} & \text{für } \hat{v}_1 = 0 \text{ und } \hat{v}_2 > 0 \\ \frac{3\pi}{2} & \text{für } \hat{v}_1 = 0 \text{ und } \hat{v}_2 < 0 \end{cases} \tag{3.20}$$

Äquivalent ist für den Vektor $\hat{\mathbf{l}}_c$ vorzugehen.
Die Differenz der beiden Winkelwerte ergibt die gewünschte relative Beschreibung zur Sichtachse der Kamera c:

$$\theta^c_{pan} = \begin{cases} \varphi_{\hat{\mathbf{v}}} - \varphi_{\hat{l}^c} & \text{für } \mid \varphi_{\hat{\mathbf{v}}} - \varphi_{\hat{\mathbf{l}}^c} \mid \leq \pi \\ 2\pi + (\varphi_{\hat{\mathbf{v}}} - \varphi_{\hat{\mathbf{l}}^c}) & \text{für } \varphi_{\hat{\mathbf{v}}} - \varphi_{\hat{\mathbf{l}}^c} < -\pi \\ -(2\pi - (\varphi_{\hat{\mathbf{v}}} - \varphi_{\hat{\mathbf{l}}^c})) & \text{für } \varphi_{\hat{\mathbf{v}}} - \varphi_{\hat{\mathbf{l}}^c} > \pi \end{cases} \tag{3.21}$$

Geht man entsprechend für die vertikale Winkelkomponente vor, erhält man eine vollständige Zustandshypothese im Bezug zum Kamerasystem.

3.2. Merkmale

Die Merkmale in deren Kontext die Partikelbewertung nach Gleichung 3.19 müssen die Eigenschaften aufweisen, dass sie Rückschlüsse auf die Kopfposition, -maße und Drehwinkel geben. Im hiesigen Anwendungsfall bedeutet das, dass mit Hinterkopfansichten gleichermaßen umgegangen werden muss, wie auch mit Vorderkopf- und Profilmotiven. Die im Folgenden beschriebenen Merkmalsbeobachtungen wurden infolgedessen deshalb ausgewählt, weil sie ausreichende Generalisierungsfähigkeiten vorweisen konnten, mit diesen Herausforderungen umgehen zu können.

3.2.1. Detektion von Kopfdrehungen mit Hilfe Neuronaler Netze

Als robuste und gut generalisierende Schätzer, haben sich Neuronale Netze für die Kopfdrehung etablieren können [KBS00, BAMPS98, ZPC02, BT02, TBC+03, SYW98]. Ebenso für eine Anwendung auf niedrig aufgelösten Motivbildern [BT02, TBC+03]. Aus diesen Gründen sollen Neuronale Netze auch hier zur Drehwinkelbewertung eingesetzt werden.

Neuronale Netze stellen miteinander verbundene Neuronenzellen dar, die als individuelle Einheiten die Reaktion ihrer Vorgänger aufnehmen, verarbeiten und an alle verbundenen Nachfolgezellen weiterleiten - wo wiederum auf die dort anliegende Eingabe reagiert wird. Ein angelegtes Signal wird damit durch das Netz durchgereicht und in allen durchlaufenen Neuronen entsprechend verarbeitet. Die Funktionalität der Neuronenmodelle entspricht dabei vage dem Verständnis von Nervenzellen im tierischen Gehirn [Gur97].

Die Ausgabe des Netzes wird im wesentlichen durch zwei Faktoren beeinflusst: Zum einen wie die Zellen selbst auf an ihnen angelegte Signale reagieren und zum anderen durch die Verbindungen zwischen den Zellen im Netz. Je nachdem wie diese beiden Faktoren ausgeprägt sind, ergeben sich vielfältige Netztopologien, die jeweils für unterschiedliche Anwendungsbereiche eingesetzt werden können.

Populär sind Neuronale Netze insbesondere in Form einer vorwärtsgerichteten Topologie, in der eine dedizierte Eingabeschicht ein angelegtes Signal empfängt und an nachfolgende Schichten von Neuronen weiterreicht. Das Signal wird so bis zu einer Ausgabeschicht durchgereicht und entsprechend verarbeitet und beeinflusst. Die Form und Reaktionen dieser Neuronen geben schließlich die abschließende Ausgabe wieder. Der Vorteil solcher Topologien liegt insbesondere darin, dass bereits mit drei Schichten ein universeller Approximator jeder stetigen Funktion eingelernt werden kann [Hor91, Zha07]. Mit einem Ausgabeneuron und kontinuierlichen Werten kann so ein Regressor eingelernt werden, der bei angelegten Motiven eine stetige Ausgabe erzeugt. Mit mehreren Neuronen hingegen kann eine Wahrscheinlichkeitsfunktion über mehrere Klassen geschätzt werden, womit jedes Neuron in der Ausgabeschicht respektive die Wahrscheinlichkeit seiner zugehörigen Klasse ausgibt. Aus Gründen die im Folgenden noch erläutert

werden, soll zuletzt genannte Topologie in dieser Arbeit beibehalten und ein Netz zur Winkelschätzung horizontaler, ein zweites für vertikale Winkel eingesetzt werden. Für eine detaillierte Beschreibung des Einlernens und hierüber hinausgehenden Anwendungsmöglichkeiten sei der Leser im Speziellen auf Sekundärliteratur, wie zum Beispiel [Gur97] verwiesen.

Im Folgenden wird der Ansatz von Stiefelhagen et al. [Sti02] weiterverfolgt.
Wie in Abschnitt 3.1.4 beschrieben wurde, projiziert zunächst die Funktion $g^c(\mathbf{s}^i)$ eine Zustandshypothese $\mathbf{s}^i$ in das Kamerabild der Kamera c. Damit wird eine Bildregion der Kopfhypothese erhalten, die im Kamerabild ein Rechteck darstellt. Dieses wird im Anschluß pixelweise ausgelesen und zu einem zusammenhängenden Vektor konkateniert. Weil die Drehschätzung unabhängig von der Farbinformation im Bild geschehen kann und die drei Farbkanäle des Bildes damit redundant sind, wird der Kopfausschnitt zunächst in ein reines Intensitätenbild umgerechnet. Ein Vorteil dabei ist, dass damit die Dimensionalität des entstandenen Merkmalsvektors zunächst um zwei Drittel reduziert werden kann, weil statt drei Komponenten pro Pixel nur noch eine in den Vektor einfließt. Als Zusatzinformation wird zum Grauwert- das Gradientenbild der hypothetisierten Kopfregion pixelweise an den Vektor angefügt. Die Konkatenation beider Bilder ergibt schließlich den Vektor, der als Signal an das Neuronale Netz angelegt und damit verarbeitet wird. Wegen der durch die Netztopologie vorgegebenen Anzahl an Eingabeneuronen, muss der berechnete Merkmalsvektor dieser in seiner Dimensionalität entsprechen. Weil sich der Vektor aus der Konkatenation der Pixelwerte im Intensitäten- und Gradientenbild ergibt, entspricht das einer notwendigen Skalierung des unterschiedlich groß erscheinenden Motivs auf eine vorgegebene Größe. Mit einer Skalierungsgröße von 32×32 Pixeln, ergibt das 1024 konkatenierte Vektorkomponenten für das Grauwertbild, 2048 für den endgültigen Vektor inklusive Gradienteninformation. Die Eingabeschicht des Netzes muss demnach aus 2048 Neuronenzellen bestehen. Die Skalierungsgröße kann dabei beliebig aber fest geschehen. Neben der Neuronenzahl in der zweiten Ebene, soll sie deswegen Teil der nachfolgenden Evaluationen sein.

Im Gegensatz zu Stiefelhagens Ansatz, wird in dieser Arbeit bewusst auf den Einsatz eines Regressors verzichtet. Interpretiert man die Ausgabe eines Regressors in Form einer Wahrscheinlichkeitsfunktion, so entspricht eine regressive Ausgabe im wesentlichen einer unimodalen Schätzung, deren globales Maximum an derjenigen Stelle der gemachten Winkelausgabe ist. Aufgrund der herausfordernden Beobachtungsbedingungen, mit denen im Rahmen dieser Arbeit umgegangen wird, soll dem Netz jedoch bewusst die Möglichkeit gegeben werden, mit mehreren Modi eine Multihypothesenschätzung auszugeben. Dies wird erreicht, indem über diskrete Winkelklassen eine Wahrscheinlichkeitsfunktion geschätzt wird, für die jedes Neuron der Ausgabeschicht die Wahrscheinlichkeit seiner ihm entsprechenden Klasse ausgibt. Wohingegen regressive, unimodale Schätzungen in Einzelkameraanwendungen sinnvollen Einsatz finden, weil hier nur innerhalb eines umrahmenden Trackingverfahrens mit Multihypothesen um-

gegangen werden kann, erlaubt eine multimodale Hypothese im Fusionsprozess das Einbringen von Unsicherheiten, die durch komplementäre Blickwinkel jeweils von den Schätzungen in den anderen Ansichten unterstützt oder abgewertet werden können. Wie sich in den nachfolgenden Evaluationen zeigen wird, führt die niedrig aufgelöste Motivqualität der Kopfkandidaten und der große Winkelwertebereich von 360° bei horizontalen Kopfdrehungen insbesondere dazu, dass bei einer hinreichend hochauflösenden Winkeldiskretisierung globale Maxima der Schätzungen zunehmend weiter in die Nachbarschaft der eigentlich annotierten Winkelklasse diffundieren: statt dass das Maximum in der korrekten Klasse auftritt, erscheint es im umliegenden Nachbarbereich; insbesondere desto weiter entfernt je feiner die Diskretisierung erfolgt. Darüberhinaus treten weitere lokale Modi auf, die sich sowohl in der korrekten Klasse als auch in weiteren Nachbarschaften des Winkelwertebereichs ansiedeln. Erkennbar ist damit, dass bei feiner Winkelauflösung die Anzahl der Modalstellen zunimmt und diese insbesondere konzentriert um die korrekte Winkelklasse auftreten. Damit wird die Schätzung unscharf und eine einfache Klassifikation sinnlos. In Kombination mit mehreren Ansichten kann die Form der fusionierten Wahrscheinlichkeitsfunktion jedoch ausreichend beeinflusst werden, um lokale Modalstellen abzuschwächen und die Schätzung stärker auf die eigentlich korrekte Winkelklasse zu konzentrieren.

Mit der beschriebenen Topologie und Ausgabe, wäre eine intuitive Anwendung der Netze die direkte Übertragung der geschätzten Wahrscheinlichkeitswerte in die Bewertungsfunktion der Partikelstützstellen: So würde für eine gegebene Winkelhypothese die vom jeweiligen Neuron gemachte Wahrscheinlichkeitsaussage als Bewertungsergebnis interpretiert werden. Nach [Zha07] entspricht die Wahrscheinlichkeitsfunktion jedoch der A-Posteriori-Dichte $p(\theta|A)$ über den Winkel θ, wenn dem Netz ein Motiv A angelegt wird. Weil das Netzverhalten im Vorfeld auf einem Datensatz eingelernt wird, steht die Wahrscheinlichkeitsschätzung damit insbesondere in keinem Verhältnis zur Partikelbewertung nach Gleichung 3.13, in der die klassenbedingte Wahrscheinlichkeit $p(A|\theta)$ im aktuellen Zeitschritt erforderlich ist.
Um dieses Problem zu umgehen, soll deswegen im Folgenden eine Metrik eingesetzt werden, die sich an die Arbeit von Nickel et al. in [Nic08] anlehnt: darin wurden lokale Detektionen von Gesichtern in einem Bildbereich und deren jeweilige Distanz zur Position der eigentlichen Hypothese als Merkmal und Bewertungsstrategie im Beobachtungsmodell eingesetzt. Das Neuronale Netz wird auf Suchfenstern angewandt, die ein vorliegendes Kamerabild in verschiedenen Größenstufen vollständig abtasten. Zur Unterscheidung sollen diese Suchfenster nachfolgend mit $\hat{\mathbf{o}}^j \in \mathbf{o}$ bezeichnet werden, wobei sich $\mathbf{o}$ auf das ursprüngliche Kamerabild bezieht. Als Beobachtungsmerkmal kann damit die Menge aller Rechtregionen $\hat{\mathbf{o}}^j$ sowie deren zugeordneter Netzausgaben $p(\theta|\hat{\mathbf{o}}^j)$ aufgefasst werden. In diesem Kontext sollen nun Zustandshypothesen $\mathbf{s}^i$ bewertet werden, die - projiziert in die jeweilige Kameraansicht - jeweils Rechteckregionen aufspannen, in denen Kopfausschnitte mit ihnen entsprechenden Drehwinkeln erwartet werden.

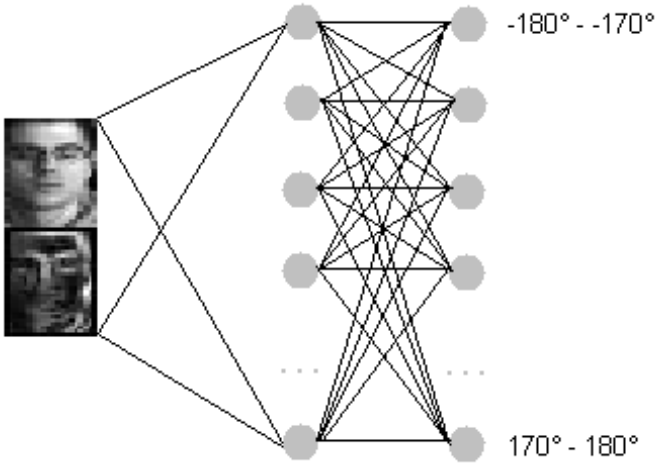


Abb. 3.5.: Schematische Darstellung des Neuronalen Netzes für die horizontale Drehwinkelschätzung. Das Netzwerk besteht aus drei Schichten: an die erste Schicht wird das ausgeschnittene und skalierte Intensitätenbild des Kopfausschnitts mitsamt seiner Kantendarstellung angelegt. Die Ausgabeschicht besteht aus sovielen Neuronen wie Winkelklassen vorliegen und jedes Neuron gibt die Wahrscheinlichkeit seiner zugehörigen Klasse für das angelegte Motiv aus.

Überträgt man die Metrik, die Nickel in [Nic08] vorschlägt, auf den hiesigen Anwendungsfall, dann bezieht sie sich im lokalen Umfeld auf jenes Suchfenster, dessen zugehörige Klasse für die entsprechende Winkelhypothese die maximale Wahrscheinlichkeit ausgegeben hat. Dieser stellt sie die Distanz gegenüber, mit der dieses Suchfenster von der eigentlichen Position der projizierten Zustandshypothese geometrisch entfernt liegt. Ein Zustand wird demnach abhängig davon bewertet, mit welcher Konfidenz und Entfernung zur projizierten Position im Kamerabild ein Kopfmotiv detektiert wurde, das dem jeweilig angenommenem Drehwinkel entspricht. Formal lässt sich die Metrik hierzu wie folgt zusammenfassen:
Sei mit $m^{NN}(\mathbf{o}^c)$ die Menge der Suchfenster und zugehörigen Netzausgaben bezeichnet, die als Merkmal in diesem Bewertungsschritt auf einem Kamerabild $\mathbf{o}^c$ berechnet werden. Ferner sei mit $\mathbf{s}^i$ die Partikelhypothese gegeben, die im jeweiligen Kamerabild c eine Rechteckregion und Winkelhypothese vorgibt. Dann erfolgt die Bewertung der Partikelstützstelle auf dem Kamerabild in Abhängigkeit von der Distanz derjenigen Suchfensterregion $\hat{\mathbf{o}}^j \in \mathbf{o}^c$, die dem Drehwinkel entsprechend die maximale Wahrscheinlichkeit $p(\theta|\hat{\mathbf{o}}^j)$ darstellt:

$$w^{NN}(m^{NN}(\mathbf{o}^c), \mathbf{s}^i) = \max_{\hat{\mathbf{o}}^j \in \mathbf{o}} \{p(\theta|\hat{\mathbf{o}}^j) d(\hat{\mathbf{o}}^j, g^c(\mathbf{s}^i))\} \tag{3.22}$$

Das Distanzmaß $d(\cdot,\cdot)$ vergleicht die beiden Bildausschnitte $\hat{\mathbf{o}}^j$ und $g^c(\mathbf{s}^i)$ auf ihre Überlappung. Theoretisch werden damit Bildregionen im Kamerabild vernachlässigbar, in denen keine Köpfe dargestellt sind, weil dort die Distanz zu nächstgelegenen Kopfmotiven zu groß und die Bewertung infolgedessen gegen 0 streben würde. Darüberhinaus kann davon ausgegangen werden, dass durch den eingesetzten Partikelfilter genau derjenige Bildbereich ausreichend ab-

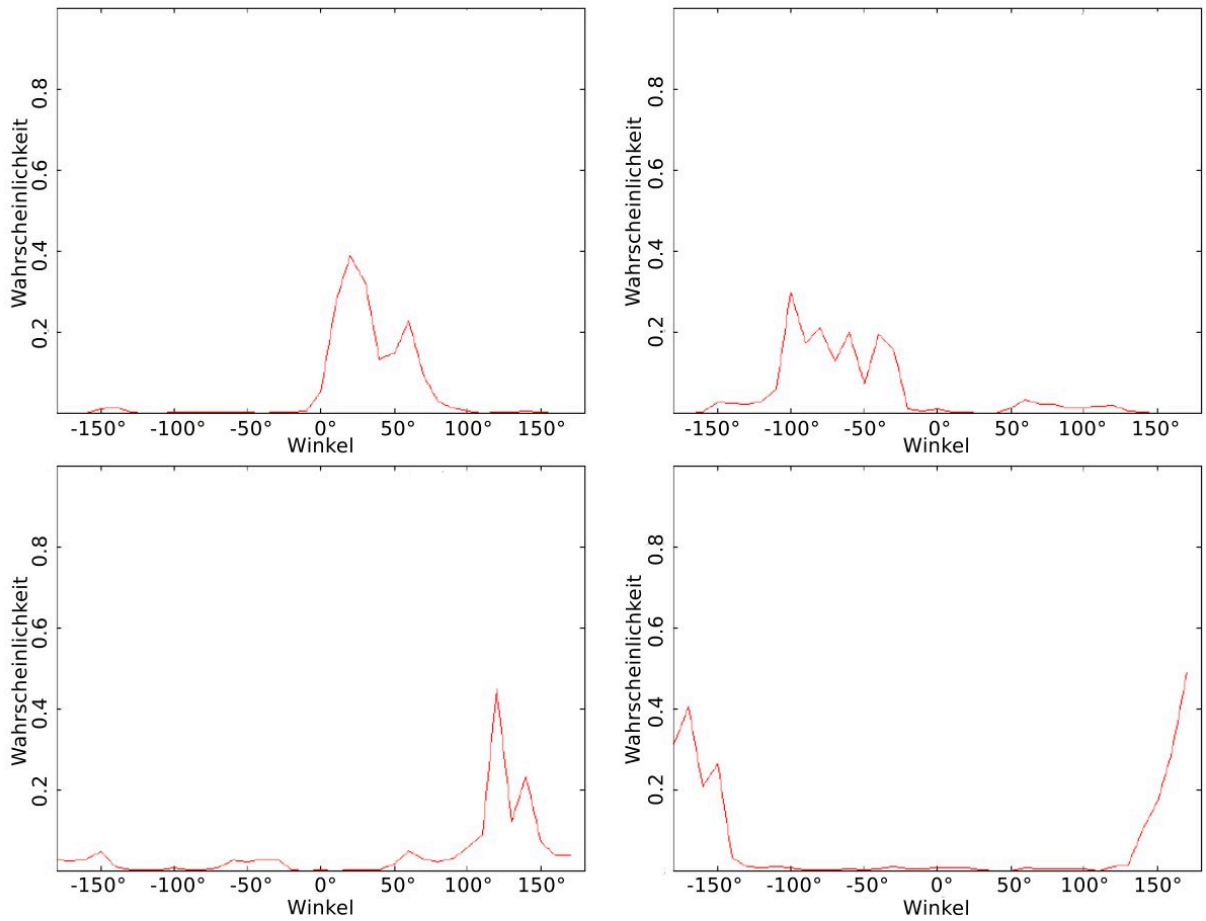


Abb. 3.6.: Beispiel der einzelansichtsbasierten Schätzungen: Pro Kamera wird eine Wahrscheinlichkeitsfunktion über den diskretisierten Winkelwertebereich geschätzt. Die Ausgaben werden anschließend akkumuliert um eine gemeinsame Wahrscheinlichkeitsfunktion zu erzeugen. Durch Hinzunahme mehrerer Kameraansichten, werden einzelne Modi entsprechend unterstützt und bestätigt oder in ihrem Ausmaß reduziert. Indem statt einer Regression die Wahrscheinlichkeitsfunktion pro Ansicht geschätzt wird, können Multihypothesen in einzelnen Ansichten zugelassen werden, die im Zuge der Fusion mit den übrigen Schätzungen dann entsprechend unterstützt oder abgeschwächt werden.

getastet wird, in dem ein tatsächliches Kopfmotiv erwartet werden kann. Das Neuronale Netz muss also nicht über das gesamte Kamerabild hinweg angewandt werden, sondern lediglich auf jenen Bildausschnitten, die durch alle Partikel zum Zeitpunkt t vorgegeben werden, was den Rechenaufwand schließlich deutlich senkt.

3.2.2. Gradientenhistogramme

In der beschriebenen Implementierung, brauchen die Neuronalen Netze eine definierte Rechteckregion, in der das Kopfmotiv eingefasst wird, um die dargestellte Drehung auf dem Motivbild zu schätzen. Lokalisierungsfehler wirken sich dahingehend aus, dass statt des eigentlichen Motivs Hintergrundrauschen mit einbezogen wird oder die Region zu klein ausfällt, was zu ei-

nem Wegfall etwaiger Merkmale im Motiv führen kann. Brown et al. untersuchten in [BT02] den Einfluss von Lokalisierungsfehlern auf die Korrektklassifikationsrate Neuronaler Netze mit äquivalenter Topologie und unterstrichen dabei die Sensitivät hinsichtlich einer konsistenten Motivlokalisierung. Mit der alleinigen Schätzung einer Wahrscheinlichkeitsfunktion, fehlt den Neuronalen Netzen ein Konfidenzmaß über die Lokalisierungsgüte. Erscheint die vorgegebene Rechteckregion inkonsistent zu den eingelernten Motivregionen, äußert sich das in zusätzlichem Rauschen in der gemachten Hypothese. Für eine ausgleichende Stabilisierung soll deswegen ein zweites Merkmal hinzugezogen werden, das die Lokalisierung des Kopfmotivs beisteuernd bewertet.

Eine ansichtsinvariante Bewertung der Kopflokalisierung muss sich dabei auf Merkmale stützen, die unabhängig von der jeweiligen Kamerasicht verfügbar sind. Ein solches ist hierzu die Kopfform, die sich mit ihrer ellipsoiden Erscheinung bei ausreichend Kontrast in der Bildebene deutlich vom Hintergrund und Rest des Körpers der Person abhebt. Eine einfache Bewertung kann infolgedessen so geschehen, dass die projizierte, den Kopfkandidaten umfassende Rechteckregion in der Bildebene dahingehend evaluiert wird, wie gut sie eine Ellipsenform heuristischer Kopfmaße umschließt. Die Bewertung der Ellipsenform geschieht dabei auf dem Gradientenbild der Kameraansicht: Ellipsen verschiedener Maße werden in die Region gelegt und entlang ihres Umfangs mit den darunterliegenden Gradientenwerten verglichen. Mit hinreichender Übereinstimmung kann so eine Pixelfolge gefunden werden, die annähernd ellipsoid erscheint und damit einer Kopfhypothese entspricht [Bir97]. Mit dieser intuitiven Vorgehensweise stellen sich jedoch zwei Probleme: (1) Bei einem Hintergrund mit hohem Anteil an Kantenstrukturen ist es nicht auszuschließen, dass die punktweise Auswertung der Ellipsenform auch Hintergrundpixel miteinbezieht. In einem solchen Fall würde sich die Ellipse hauptsächlich an dominanten Kanten im Hintergrund ausrichten, statt der eigentlichen Form zu entsprechen. (2) Die Anbringung der Kameras im hiesigen Anwendungsfall hat zur Folge, dass die Kopfmotive von oben betrachtet, geneigt erscheinen. Mit der Varianz der Frisuren weicht die Kopfform dabei nicht unerheblich von der Form einer Ellipse ab. Segmentweise kann zwar eine Bogenform nachvollzogen werden, eine konkatenierte Pixelfolge in Ellipsenform ist jedoch in der Regel nicht vorhanden.
Darauf eingehend, kann die Ellipsensuche wie in [Bir98] dahingehend erweitert werden, dass sie nicht nur die Gradientenpixel zur Bewertung berücksichtigt, sondern stattdessen auch die Normalen der Pixelgradienten hinzuzieht, so dass Hintergrundrauschen anhand der jeweiligen Kantenausrichtung von der ellipsenförmigen Kopfform differenziert werden kann. Mit der jedoch geforderten, pixelweisen Überdeckung reagiert das Verfahren nicht flexibel auf abweichende Formmotive. Durch den teilweise fehlenden Kontrast bei sehr hellen Haarfarben, beziehungsweise einem zu Haarfarben ähnlichen Hintergrund, und der nicht deckungsgleichen Annäherung an die Ellipsenform durch Frisuren und Haarmustern muss die Bewertung noch

immer zu einem gewissen Grad invariant gegenüber Translationen in einzelnen Formabschnitten bleiben. Ein solches Herausstellungsmerkmal bietet der Ansatz der sogenannten *Histograms of Oriented Gradients* von Dalal und Triggs [DT05], der im Folgenden mit Gradientenhistogramme bezeichnet und beschrieben werden soll.

Gradientenhistogramme sind globale Merkmalsdeskriptoren, die auf dem Gradientenbild die Orientierung der Kantenstrukturen lokalen Histogrammen zuordnen. Dabei wird empfohlen das Bild zunächst in seinem Gammawert zu normalisieren, um das Kontrastverhältnis zu erhöhen. Wie Dalal und Triggs in ihrer ursprünglichen Arbeit untersucht haben, nimmt die Methode zur Gradientenberechnung nur marginalen Einfluss auf die Genauigkeit: Filterkerne kleiner Größe sind deswegen schon allein aufgrund ihrer schnelleren Berechenbarkeit zu bevorzugen. Im weiteren Verlauf dieser Arbeit wird deswegen eine Faltung des Kopfmotivs mit einem Sobelkern der Größe 3×3 Pixel eingesetzt: Das Kopfmotiv wird zunächst in ein Intensitätenbild umgerechnet und in seinem Kontrast normalisiert. Die Magnituden aus den Faltungen mit einer Sobelmatrix für Gradienten in horizontaler und einer Sobelmatrix für Gradienten in vertikaler Richtung dienen dann im weiteren Verlauf zur Berechnung des Merkmalsvektors.

Das Gradientenbild wird in lokale Bereiche der Größe 8×8 Pixel unterteilt, nach Dalal und Triggs sogenannten *Zellen*. Das sind nicht-überlappende Fenster mit gleicher Breite, innerhalb der jeweils alle Pixelwerte einem lokalen Histogramm zugeordnet werden. Die Histogrammtöpfe entsprechen dabei Orientierungsdiskretisierungen der Pixelgradienten. Die Granularität der Diskretisierung nimmt wesentlichen Einfluss auf die Genauigkeit der Modellbeschreibung und späteren Lokalisierung. Das überrascht nicht, wenn man bedenkt, dass durch eine zu grobe Klassenaufteilung der Winkel der Diskretisierungsfehler wächst und der Deskriptor damit weniger diskriminativ wird. Den Autoren nach sollen jedoch neun Diskretisierungsintervalle beziehungsweise Histogrammtöpfe zu je 40° ausreichend sein.

Das Ausmaß der Gradientenmagnituden hängt zu einem großen Teil von Kontrast und Beleuchtung in den jeweiligen Zellen ab. Eine Normalisierung über mehrere Zellen hinweg ist deswegen wichtig um lokale Varianzen zu minimieren. Dazu werden Histogramme benachbarter Zellen konkateniert und deren Konkatenation jeweils normalisiert. Das Zusammenfügen geschieht dabei überlappend zu *Blöcken* von je 3×3 Zellen. Durch die Überlappung werden Varianzen zwischen Zellen in all denjenigen Blöcken berücksichtigt, die diese Zellen miteinbeziehen und die Normalisierung so auf diese durchpropagiert. Die Abbildungen 3.7 und 3.8 stellen hierzu eine schrittweise Verdeutlichung der gesamten Vorgehensweise dar. Konkateniert man die Blöcke, erhält man den endgültigen Merkmalsvektor. Dieser beschreibt schließlich aufgrund seines Aufbaus durch lokale Zellbereiche die örtliche Anordnung der Gradienten zueinander. Damit wird die ellipsenförmige Annäherung des Kopfes erfasst, unabhängig von der übrigen Kopftextur oder anderweitigen Hintergrunderscheinungen.

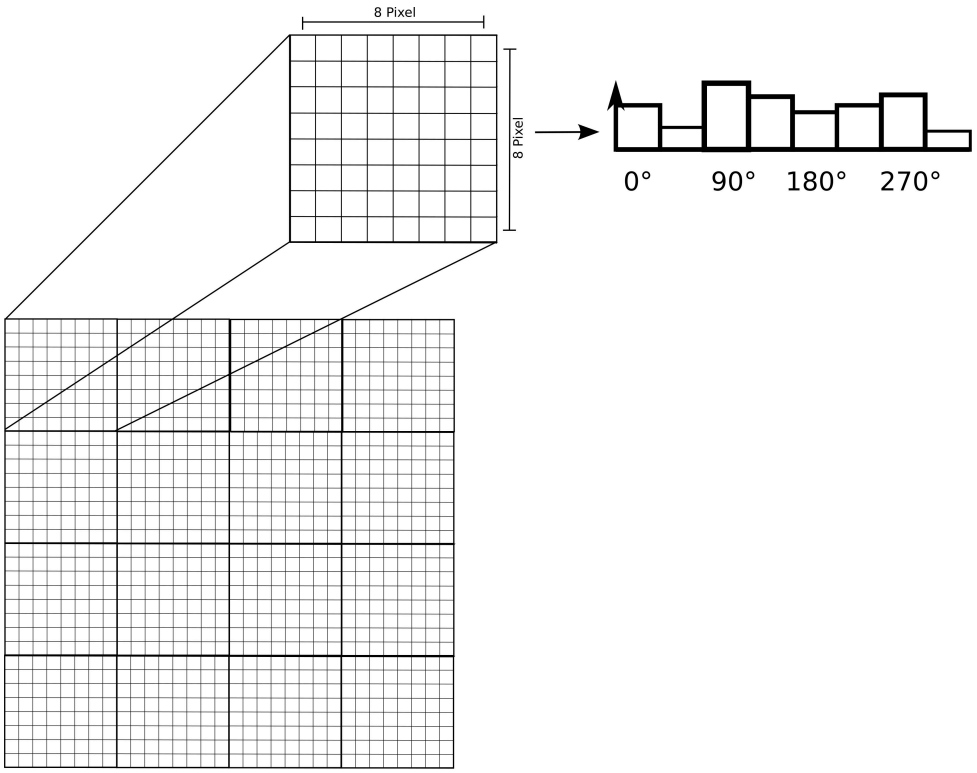


Abb. 3.7.: Berechnung der Gradientenhistogramme: Im ersten Schritt werden die Motive in äquidistant große Zellen zerlegt (8×8 Pixel). Für jede Zelle wird anschließend ein Histogramm über die darin beinhalteten Gradienten berechnet (bei einer Diskretisierung von 45°).

Dalal und Triggs setzten den Deskriptor dazu ein, mit Hilfe einer Support Vektor Maschine Suchfenster in einem Kamerabild daraufhin zu klassifizieren, ob eine Personensilhouette darin dargestellt war oder nicht. Die Histogramme kodierten hierzu die eindeutige Form eingelernter Personen - in Abbildung 3.9 wird diese Eigenschaft dadurch hervorgehoben, dass die Stützvektoren der Support Vektor Maschine im Gradientenbild eingezeichnet wurden. Man sieht, dass insbesondere die Gradientenausprägungen im Silhouettenbereich jene sind, die dominant zur Kennzeichnung beitragen. Innere Zellen, die den übrigen Bereich des Motivs abdecken, tragen hingegen nur marginal zur Klassifikation bei. Dem entsprechend kann die ellipsenartige Kopfform in den hiesigen Gradientenhistogrammen aufgefasst werden: Diejenigen Gradienten im Motiv, die die Silhouette des Kopfs nachbilden, treten dominant auf, während die Gradienten in den übrigen Zellen, verrauscht durch die vielen unterschiedlichen Motive im Lernprozess, eher uniform verteilt auftreten.
Durch die Gradientenauswertung innerhalb lokaler Zellen, statt pixelweise wie bei einem direkten Ellipsenvergleich, wirkt der Deskriptor darüber hinaus aber auch zu einem gewissen Maße translationsinvariant. Kopfmotive müssen somit nicht mehr pixelgenau deckungsgleich mit der angenommenen Ellipsenform sein. Durch die Zellbildung ist es ausreichend, wenn innerhalb der Zellregion ein lokaler Ellipsenbogenabschnitt auftritt. Die Invarianz zeigt damit nicht nur

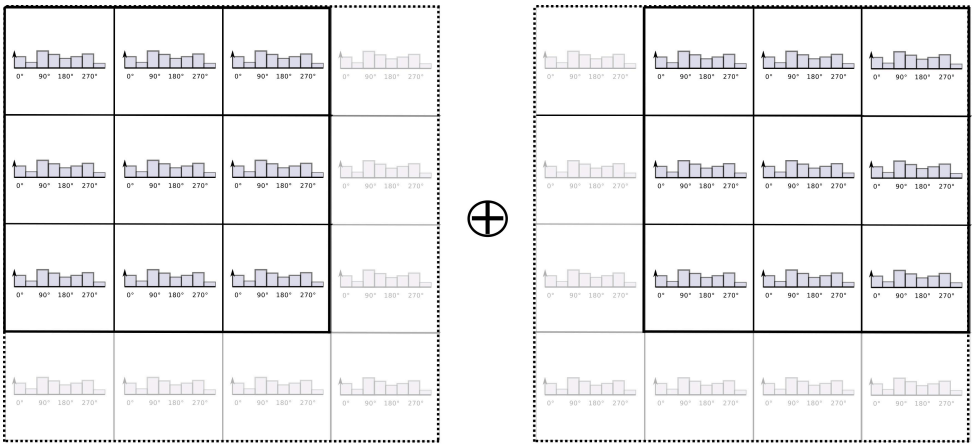


Abb. 3.8.: Blockbildung bei Berechnung der Gradientenhistogramme: Im zweiten Schritt werden die berechneten Zellenhistogramme zu Blöcken zusammengefasst und die Histogramme jeweils konkateniert. Dabei wird ein Fenster über die Zellen verschoben, das jeweils 3×3 Zellen umfasst.

Robustheit gegenüber Translationen in der Lokalisierung, sondern auch hinsichtlich verschiedener Frisuren.

Im Gegensatz zur originalen Arbeit von Dalal und Triggs, soll kein Klassifikationsschritt anhand des Deskriptors eingesetzt werden. Stattdessen wird ein einzelner Repräsentant einer Kopfsilhouette über verschiedene Kopfdarstellungen im Datensatz gebildet und Bildregionen mit diesem Referenzmodell verglichen und im Hinblick auf ihre Ähnlichkeit bewertet. Wie gut eine Bildregion einem Referenzmodell einer Kopfsilhouette dabei entspricht, kann daran gemessen werden, wie ähnlich die Gradientenausprägung in den Zellen denen des Referenzmodells gleicht. Mit der Blockbildung werden die Zellhistogramme zwar kombinatorisch permutiert, der fertige Vektor bleibt aber durch die Zellduplikation in den Überlappungen komponentenweise vergleichbar. Als vergleichendes Gütekriterium soll deshalb die L2-Norm $\|\cdot\|_2$ als Distanzmaß vorgeschlagen werden, nach der ein Kopfkandidat im Vergleich zum Referenzmodell nach Gleichung 3.16 bewertet wird. Die Repräsentation des Kopfkandidaten stellt dabei insbesondere die Merkmalsbeobachtung $m^{GH}(g^c(\mathbf{s}^i))$ dar. Bezeichnet $\hat{m}^{GH}$ im Zuge dessen die Repräsentation des Referenzmodells, so lässt sich nach [IB98, CFCP07, BO05] eine Stützstellenbewertung exponentialverteilt probabilistisch durchführen:

$$w^{GH}(m^{GH}(\mathbf{o}^c), \mathbf{s}^i)) = \mathrm{e}^{-\lambda^{GH}\|m^{GH}(g^c(\mathbf{s}^i))-\hat{m}^{GH}\|_2} \tag{3.23}$$

Der empirisch festgelegte Skalierungsfaktor λ^{GH} orientiert sich an der Standardabweichung des angenommenen Fehlermodells und skaliert den Abfall der Bewertung zu weit distanten Beobachtungen. Die Verwendung von Exponentialfunktionen hat dabei insbesondere den Vorteil, dass selbst schwach bewertete Stützstellenhypothesen mit endlicher und berechenbarer Wahrscheinlichkeit ausgestattet werden. Das hat insbesondere zur Folge, dass auch in spärlich ab-

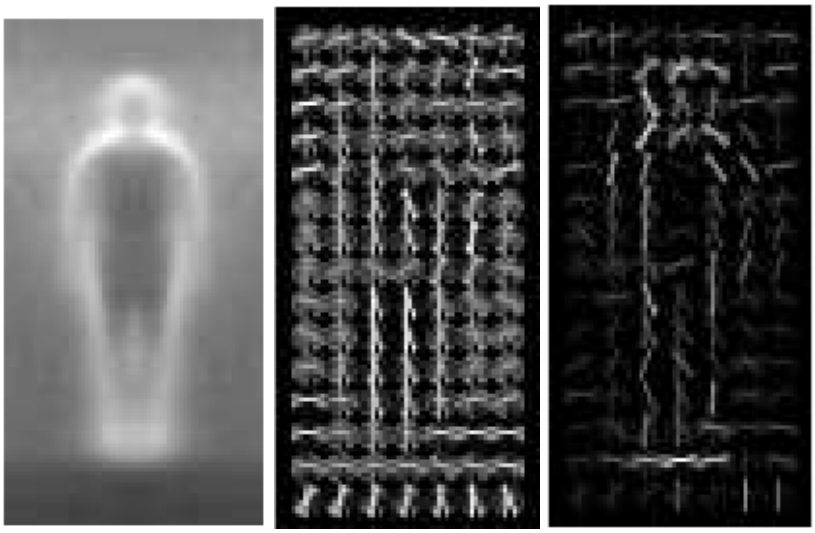


Abb. 3.9.: Darstellung einer Personensilhouette nach Dalal und Triggs [DT05]. Links abgebildet ist das gemittelte Gradientenbild verschiedener Silhouetten im Datensatz. In der Mitte die zugehörigen Gradientenhistogrammzellen. Rechts sind diejenigen Gradienten hervorgehoben, die nach Einlernen der Support Vektor Maschine zum Klassifizieren den Stützvektoren entsprechen. Zu erkennen ist hierbei, dass insbesondere diejenigen Gradienten relevant sind, die die Silhouette deutlich nachbilden.

getasteten Zustandsräumen, Stützstellen in Randbereichen weiterhin einbezogen werden können [CF09].

3.3. Evaluationen

In diesem Abschnitt sollen die beschriebenen Komponenten auf Parameterbelegung und Fehler beziehungsweise Korrektklassifikationsrate untersucht werden. Wie bereits in Kapitel 2.1.3 erwähnt wurde, stellt der im Zuge dieser Arbeit entstandene Datensatz mit seiner Veröffentlichung und Nutzung während der CLEAR-Evaluationen 2007 [SBB+07] bis dato die alleinige Referenz hierfür dar. Die folgenden Ergebnisse stehen so in direktem Vergleich mit den wenigen verwandten Systemen, die in der Literatur vorzufinden sind. Deswegen soll zunächst der Datensatz im einzelnen vorgestellt und beschrieben werden. Die im Anschluss folgenden Analysen der Systemkomponenten geschehen daraufhin zunächst auf einzelkamerabasierten Auswertungen. Damit soll die Leistungsfähigkeit der einzelnen Verfahren auf den hiesigen Aufnahmebedingungen untersucht werden und insbesondere der zusätzliche Nutzen einer Mehrkameraumgebung herausgearbeitet werden.
Hierfür wird zunächst die Lokalisierung von Kopfkandidaten evaluiert, in dem die in Abschnitt 3.2.2 beschriebenen Gradientenhistogramme in einem rein detektorbasierten Vorgehen auf dem Datensatz angewandt und die hypothetisierten Kopfkandidaten mit den im Datensatz beinhalteten Annotationen der tatsächlichen Kopfregion verglichen werden.

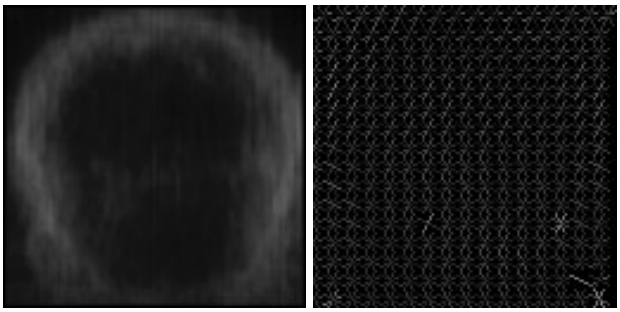


Abb. 3.10.: Darstellung eines Kopfrepräsentanten für die Lokalisierung. Links abgebildet ist ein durchschnittliches Gradientenbild über 100 verschieden gedrehte Köpfe im benutzten Datensatz. Deutlich erkennbar ist die stark hervortretende Silhouette im Gegensatz zum inneren Texturbereich des Kopfs. Rechts abgebildet ist eine Visualisierung der entsprechenden Gradientenhistogrammzellen: eingezeichnet sind pro Zellfenster (Fenster von je 8×8 Pixel Größe) die Ausprägungen der einkodierten Gradientenrichtungen. Erkennbar sind die stärkeren Gradienten im Silhouettenbereich, während die inneren Gradienten gleichverteilt wirken und keine ausreichende Unterscheidungsfähigkeit für ein Motiv bedeuten.

Daneben wird die in Abschnitt 3.2.1 dargestellte Drehwinkelschätzung direkt auf den annotierten Kopfregionen im Datensatz angewandt, um die zu erwartende Genauigkeit bei der Winkelschätzung, aber auch die tatsächliche Form der geschätzten Wahrscheinlichkeitsfunktion näher zu untersuchen. Wie gezeigt werden wird, stellt gerade im Hinblick auf die nachgestellte Entscheidungsfusion die mögliche Multimodalität der ansichtsbasierten Wahrscheinlichkeitsschätzung einen wesentlichen Einfluss auf den Mehrwert durch verschiedene, gleichzeitige Kameraansichten dar. Als weiteren Faktor soll aber auch der Einfluss des möglichen Lokalisierungsfehlers hierbei nicht unberücksichtigt bleiben.
Im Zuge der Entscheidungsfusion werden die Zustandsbewertungen aller Kamerraansichten zusammengeführt und der daraus resultierende Erwartungswert im Zustandsraum als endgültige Schätzung des Systems ausgegeben. Die Genauigkeit hängt dabei von den aufgeführten Faktoren und jeweiligen Parameterbelegungen ab. Die Evaluation der gesamtheitlichen Systemschätzung basiert deswegen auf den optimalen Parameterwerten der zunächst folgenden Teilkomponentenauswertung.

3.3.1. Der CLEAR'07 Datensatz

Die CLEAR-Evaluationen [SBB+07] sind zwei jeweils in den Jahren 2006 und 2007 durchgeführte, internationale Workshops, die maßgeblich aus dem EU-geförderten Forschungsprojekt CHIL [WS09] hervorgegangen sind. Schwerpunkt des Projekts lag auf intelligenten Mehrkameraumgebungen. Demzufolge war das automatische Schätzen der Kopfdrehung unter den dieser Arbeit entsprechenden Bedingungen Bestandteil der Workshopevaluationen.
Der Datensatz, der dabei im Rahmen der Workshops veröffentlicht wurde, sollte die bestehende Lücke füllen und fortwährend eine Referenz für die bis dahin weitgehend unbeachtete Pro-

blemstellung bieten. Der während dieser Arbeit aufgenommene und -bereitete Datensatz war dabei Grundlage für diese Publikation und dient bis heute zum Vergleich verwandter Ansätze [ZHLH06, CFCP06, PZ07, LB07, BO07b, CFCP07, YZF+08].

Umfang

Der CLEAR'07-Datensatz zur multikamera-basierten Kopfdrehungsschätzung mit niedrig aufgelösten Kopfausschnitten beinhaltet 15 Videos unterschiedlicher Personen in einem mit mehreren Kameras ausgestattetem Labor des Instituts für Theoretische Informatik der Universität Karlsruhe in Deutschland. Darin beobachten vier Kameras vom Typ Sony DFW500 die Person aus unterschiedlichen Blickwinkeln: die Kameras sind hierfür unter der Raumdecke, in den jeweiligen Ecken des Raums angebracht. Dabei handelt es sich um Digitalkameras, die mit einer Auflösung von 640×480 Pixeln Einzelbildfolgen mit 15 Hz im JPEG-Format aufnahmen. Jede der 15 Personen wurde drei Minuten lang mit 15 Frames pro Sekunde pro Kamera aufgezeichnet. Während der Aufnahmen wurden die Personen dabei aufgefordert, ihren Kopf wiederholt den Kameras zu und weg zu drehen um so aus jeder Ansicht den gesamten Winkelwertebereich abzudecken.
Die Kameras wurden untereinander nicht synchronisiert. Die Zeitstempel der Digitalisierungen dienen dazu, die Videoströme einander auszurichten. Daraus ergeben sich verschiedene Anzahlen an Einzelbildern in den Kameraaufnahmen einer Person. Tabelle 3.1 gibt eine detaillierte Übersicht über die sich daraus ergebenen Einzelbildmengen der jeweiligen Videos.

Annotationen

Während der Aufnahmen trugen alle Probanden einen magnetischen Bewegungssensor auf dem Kopf (Flock of Birds der Firma Ascension Technology Corporation), der die tatsächliche Kopfdrehung mit einer Frequenz von $30Hz$ protokollierte. Dabei wurde die relative Orientierung des auf dem Kopf getragenen Sensors zu einem vermessenen Transmitter ermittelt, welcher entlang der positiven X-Achse des Raumkoordinatensystems ausgerichtet war, so dass die aufgenommenen Winkelangaben den Winkelwerten im Weltbezugssystem entsprachen.
Der getragene Sensor auf dem Kopf war dabei auf einem gefärbten Haarreif befestigt, um dessen Präsenz in den Kameraaufnahmen zu reduzieren.
Zusätzlich zu den automatisch protokollierten Drehwinkeln, wurde die Kopfregion im Anschluss jeder Aufnahme in jedem Kameraeinzelbild manuell annotiert. Damit wurde für Evaluationszwecke schließlich für jede Person das den Kopf umfassende Rechteck sowie die Bildposition des Kopfmittelpunkts vorgegeben. Die annotierten Kopfgrößen schwanken dabei von 24×16 Pixel bis zu 60×40 Pixel, abhängig von Person und Entfernung zu den jeweiligen Kameras.

Video	Anz. Frames pro Kamera	min. Kopfgröße [Px]	max. Kopfgröße [Px]
1	535	34×40	52×58
2	533	30×32	52×48
3	533	26×32	54×52
4	533	24×28	50×50
5	268	28×32	48×50
6	535	36×40	58×60
7	535	28×36	46×54
8	528	32×22	48×56
9	535	36×36	**56×62**
10	535	30×40	58×54
11	535	30×36	54×48
12	535	28×34	50×50
13	536	**24×24**	52×54
14	535	26×36	50×56
15	534	26×30	52×50

Tab. 3.1.: Umfang des CLEAR'07-Datensatzes für die Kopfdrehungsschätzung. Zusätzlich sind die minimalen und maximalen Kopfgrößen in den Videodaten angegeben (fett hervorgehoben: die minimale und maximale Kopfgröße im gesamten Datensatz).

3.3.2. Parameterevaluation einzelkamerabasierter Teilkomponenten

In diesem Abschnitt sollen die Einflüsse verschiedener Parameterbelegungen auf die Teilkomponenten untersucht werden. Die Evaluation erfolgt dabei auf Einzelkamerabildern ohne weitergehende Fusion. Die Komponenten sollen untersucht werden, um optimale Parameterwerte als Grundlage für die Gesamtsystemevaluation anschließend anwenden zu können.

Im Einzelnen handelt es sich bei den zu untersuchenden Parametern unter anderem um die Skalierungsgröße des zu bewertenden Motivausschnitts: Mit wachsender Distanz zu einer Kamera erscheint eine Person, insbesondere ihr Kopf, stetig kleiner werdend. Im Rahmen der Gradientenhistogramme muss deswegen eine Skalierung des Kopfkandidatens erfolgen, damit der daraus gewonnene Merkmalsvektor dieselbe Dimensionalität aufweist wie der zum Vergleich vorliegende Referenzvektor. Im Fall der Neuronalen Netze ist die Dimensionalität des zu berechnenden Merkmalsvektors durch die Anzahl der Eingangsneuronen festgelegt.

Weil die Topologie der Neuronalen Netze bereits in Abschnitt 3.2.1 vorgegeben wurde, liegen weitere Parameter in der Anzahl der Neuronen in der zweiten und schließlich dritten Schicht vor. Mit der Neuronenmenge in der Ausgabeschicht wird ferner die Diskretisierungsgranularität des zu berücksichtigenden Winkelwertebereichs vordefiniert: Eine Zerlegung der zum Beispiel horizontalen Winkel im Bereich $[-180°, 180°)$ führt so bei wachsender Anzahl an Ausgabeneuronen zu einer feineren Winkelklassenunterscheidung. Unter Berücksichtigung der Qualität der Kopfmotive ist dabei allerdings zu erwarten, dass eine hochauflösende Klassenaufteilung

Abb. 3.11.: Beispielaufnahme aus dem CLEAR'07 Datensatz zum Evaluieren von Kopfdrehungshypothesen in niedrig aufgelösten Mehrkameraumgebungen. Personen wurden darin von vier Kameras aus den jeweiligen Ecken des Raums aufgezeichnet. Zur Protokollierung trugen sie einen Magnetsensor, der die tatsächliche Kopfdrehung im Weltkoordinatensystem annotierte (als Pfeil eingezeichnet). Dargestellt sind ferner die nachträglich manuell erstellten Annotationen der jeweiligen Kopfmotivregion in den Bildebenen.

zwar eine prinzipiell detailreiche Winkelschätzung zulässt, ein eventueller Modalwert in der Wahrscheinlichkeitsfunktion aber zunehmend in Nachbarbereiche der eigentlichen Winkelklasse divergiert und eine naive Klassifikation damit fehlerbehafteter ausfällt. In Anbetracht der späteren Akkumulierung der Dichteschätzungen, stellt sich darüber hinaus die Frage, welchen Einfluss die Winkelgranularität auf die Fusion nehmen wird.

Bewertung der Lokalisierungskomponente

Wie in Abschnitt 3.1.4 dargestellt wurde, geschieht die Repräsentation eines Kopfkandidaten durch sein entsprechendes Gradientenhistogramm, indem die Zustandshypothese $vektors^i$ eines Partikels, insbesondere Position und Maße des Kopfs, zunächst in die Bildebene einer Kamera projiziert wird. Die aufgespannte Bildregion darin entspricht der, in der jeweiligen Kameraansicht erwarteten, Lage des Kopfs. Dieses Motiv wird in seine Gradientenhistogrammdarstellung

überführt und das Ergebnis schließlich mit einem entsprechenden Referenzmodell eines Kopfs verglichen.
Die Repräsentation des Kopfkandidaten muss in ihrer Dimensionalität dabei der des Referenzvektors entsprechen. Weil die projizierte Motivgröße eines distanzierten Kopfs proportional sinkt, beziehungsweise bei einem näher gelegenen Kopf entsprechend wächst, muss der Kopfkandidat auf die jeweilige Motivgröße des Referenzmodells skaliert werden. Die Segmentierung des Motivs in Zellregionen, wie in Abschnitt 3.2.2 beschrieben wird, verlangt dabei eine minimale Skalierung auf 24×24 Pixel. Die Größe wird von der eingesetzten Blockbildung im Anschluss vorgegeben, weil sonst keine 3×3 Zellnachbarschaften zu je 8×8 Pixel pro Zelle erreicht werden kann. Im Folgenden sollen deshalb zwei unterschiedliche Skalierungsgrößen evaluiert werden: 24×24, sowie 32×32 Pixel. Größere Werte können aufgrund der kleinen Kopfmotive in den Kamerabildern vernachlässigt werden und würden darüber hinaus zu erhöhtem Rechenaufwand führen.

Im Unterschied zur ursprünglich diskriminativen Anwendung der Gradientenhistogramme mit einem nachträglichen Klassifikator, wird im Zuge der Bewertungsstrategie dieser Arbeit die Motivrepräsentation mit dem Referenzmodell geometrisch verglichen. Hierfür wird die L2-Norm als Distanzmaß eingesetzt, um die Äquivalenz eines Kandidaten zur zu erwartenden Erscheinung zu bewerten. Für eine Evaluation der Bewertungsstrategie per se wird hier deshalb explizit ein Suchverfahren angewandt, das in einem gegebenen Fenster fester Größe nach dem globalen Minimum der Distanzwerte sucht. Für ein solches Suchfenster kann argumentiert werden, dass die Streuung der Partikel um das stochastische Mittel der Wahrscheinlichkeitsfunktion im Zustandsraum diffundiert. Dieser Mittelwert entspricht der eigentlichen Position der zu beobachtenden Person. Da im Datensatz keine weiteren Personen unterschieden werden müssen und die Position so singulär ist, ist ein Suchfenster mit hinreichender Größe ausreichend, um den Zustandsraum faktisch begrenzen zu können. Heuristisch wurde hierfür eine dreidimensionale Würfelregion um den Kopfmittelpunkt, mit einer Kantenlänge von $1m$ in die jeweilige Kamerabildebene projiziert. In Abbildung 3.12 ist die mittlere Distanz und Standardabweichung der Detektionen dargestellt. Die Evaluation erfolgte in Form einer Leave-One-Out-Kreuzvalidierung - die Referenzmodellbildung geschah exklusiv der Person im Datensatz, auf der jeweils evaluiert wurde.
Zur Vollständigkeit wurde ein kameraabhängiges als auch kameraunabhängiges Modell berechnet: im kameraabhängigen Fall wurde die Referenz nur auf Beispielen der jeweiligen Kamera erstellt und auch angewandt. Im kameraunabhängigen Fall geschah die Referenzbildung über alle Ansichten gleichzeitig und das Modell wurde auch auf allen Kameras zur Evaluation benutzt. Dabei kann gesehen werden, dass sich die Abweichung des unabhängigen Modells nur marginal von den abhängigen Resultaten unterscheidet. Ebenso scheint die Skalierungsgröße nur wenig Einfluss auf die Lokalisierung zu nehmen.

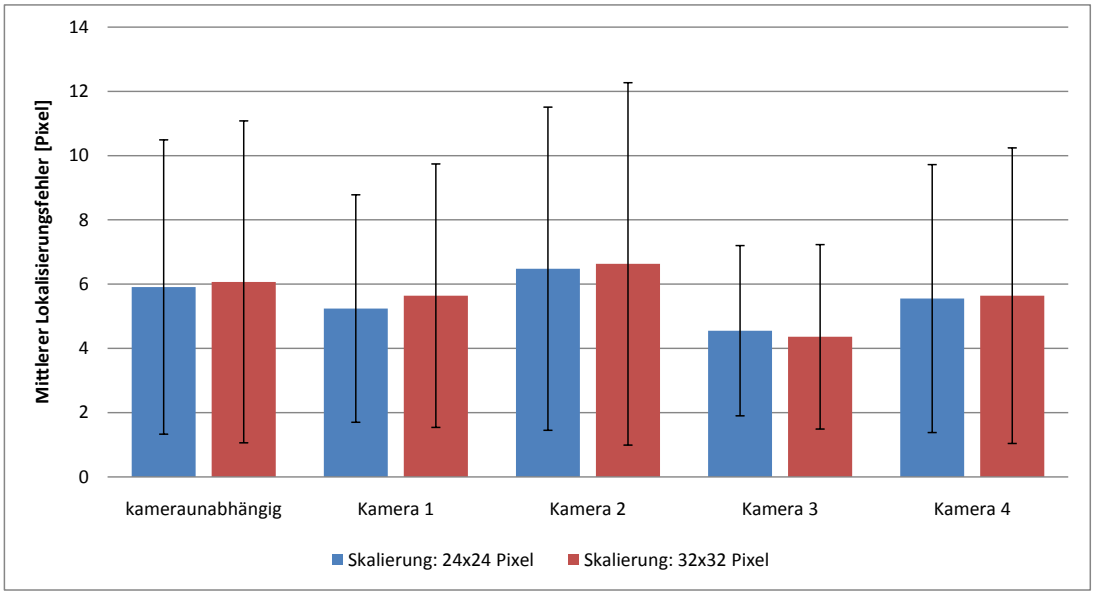


Abb. 3.12.: Fehler der Kopflokalisierung bei unterschiedlichen Skalierungsgrößen. Dargestellt ist die mittlere Distanz und Standardabweichung der Detektionen. Unterschieden wird dabei zwischen dem Einsatz eines kameraunabhängigen Referenzmodells auf allen Kameraansichten sowie der jeweiligen Erstellung und alleinigen Anwendung eines Modells auf seiner entsprechenden Ansicht.

Die größten Unterschiede treten in den Standardabweichungen auf, die auf Kamera 1 und 3 geringer, auf den Kameras 2 und 4 aber größer ausfällt. Der detektierte Kopfmittelpunkt weicht damit in den Kameras 2 und 4 stärker von den manuellen Annotationen ab, als in den übrigen Ansichten. Unter Berücksichtigung der Verteilung der Kopforientierungen im Datensatz wird dabei deutlich, dass mit Ansicht 1 und 3 Profilaufnahmen erfolgen, wenn der Kopf in seine Ursprungslage orientiert ist (0°). Ein solcher Bias ist in den Daten zu erkennen. Die Kameras 2 und 4 nehmen infolgedessen Hinterkopfansichten der Personen in solchen Momenten auf. Der Unterschied in der Standardabweichung kann also damit begründet werden, dass durch diesen Hinterkopfbias in den Daten in der Ruheposition eine größere Streuung der Detektionen stattfindet als auf den gegensätzlichen Vorderkopf- beziehungsweise Profilansichten. Diese Lokalisierungsdivergenzen sollten jedoch in einem gemeinsamen Trackingverfahren, das alle Ansichten gleichermaßen einbezieht, weitgehend reduziert werden können. Zusammenfassend erscheint die kameraunabhängige Evaluation damit durchaus zu rechtfertigen. Damit kann insbesondere für die Erstellung eines ansichtsunabhängigen Modells argumentiert werden, das die Flexibilität des Systems in Bezug auf ein Hinzufügen weiterer Ansichten beibehält. In den nachfolgenden Evaluationen des Gesamtsystems wurde deswegen ein solches Modell bei einer Skalierungsgröße von 24×24 Pixel eingesetzt.

Topologieevaluation der Neuronalen Netze zur Drehwinkelbewertung

Die Faktoren, die Einfluss auf die Drehwinkelbewertung nehmen, sind die Größe der Motivskalierung zur Merkmalsvektorberechnung, die Granularität der Winkeldiskretisierung sowie Anzahl der Neuronen in der zweiten Schicht. Mit der Größe der Motivskalierung ist die Anzahl notwendiger Eingangsneuronen vorgegeben, die Winkeldiskretisierung hingegen definiert die Anzahl der Neuronen in der Ausgabeschicht. In den folgenden Evaluationen wurde deswegen insbesondere der Einfluss dieser Parameter auf die Korrektklassifikationsrate der Winkelbewertung untersucht. Dabei wurde direkt jene Winkelklasse als Hypothese verwendet, in der das Maximum der vom Netz geschätzten Wahrscheinlichkeitsfunktion auftrat. Die Kopfmotive zur Evaluation wurden aus den manuellen Annotationen im Datensatz übernommen. Ebenso wurden diese zum Einlernen der Netze benutzt. Wie auch in der Evaluation der Gradientenhistogramme wurde dabei in einer Leave-One-Out-Kreuzvalidierung evaluiert: Die Netze wurden dabei mit allen Kopfannotationen im Datensatz, exklusiv der zu evaluierenden Person, eingelernt und schließlich auf letzterer angewandt. Abbildung 3.13 stellt hierfür die Ergebnisse der horizontalen Winkelschätzung dar. Abbildung 3.14 entsprechend für die vertikale. In den Abbildungen ist die Korrektklassifikationsrate der genannten Klassifikation bei verschiedenen Netztopologien eingezeichnet. Die Abbildungen unterscheiden dabei zwischen den jeweiligen Skalierungsgrößen der Kopfmotive. Die verschiedenfarbigen Kurven in den Abbildungen entsprechen jeweils einer unterschiedlichen Anzahl Neuronen in der zweiten Schicht. Die Ergebnisse sind dabei über die verschiedenen Winkelgranularitäten aufgeführt. Die berücksichtigten Diskretisierungsgrößen waren für die horizontale Schätzung {1°, 3°, 5°, 10°, 15°, 20°, 30°, 45°, 60°, 90°, 120°, 180°} sowie {1°, 3°, 5°, 10°, 15°, 20°, 30°, 45°, 60°, 90°} für die vertikale Ausgabe. Als Vergleichsgrundlage stellt die schwarze Gerade in den Abbildungen die zu erwartende Korrektklassifikationsrate dar, wenn die Winkelhypothese aus der Menge vorhandener Klassen zufällig gewählt worden wäre.

Wie man deutlich erkennen kann, sind die Unterschiede in allen Skalierungsgrößen nur marginal. Die Anzahl der versteckten Neuronen nimmt damit nur unwesentlich Einfluss auf die Lage eines Modalwerts in der Wahrscheinlichkeitsfunktion über die Klassen. Deutlicher hervorzuheben ist die Abhängigkeit von der Winkeldiskretisierung. Unschwer zu sehen ist, dass bei feinerer Granularität die Anzahl der korrekt klassifizierten Beispiele gegen 0 konvergiert. Wie zu Beginn des Abschnitts bereits hervorgehoben wurde, ist dieses Verhalten zu erwarten: Dabei steigt zwar die Auflösung des modellierten Winkelwertebereichs und Schätzungen geschehen im einstelligen Gradbereich, die Qualität der Kopfmotive ist jedoch zu gering um die Leistungsfähigkeit der Komponente auf die entsprechende Granularität übertragen zu können. Je feiner die Winkelaufteilung geschieht, desto eher divergiert der Modalwert in Nachbarbereiche der eigentlich zu erkennenden Winkelklasse.

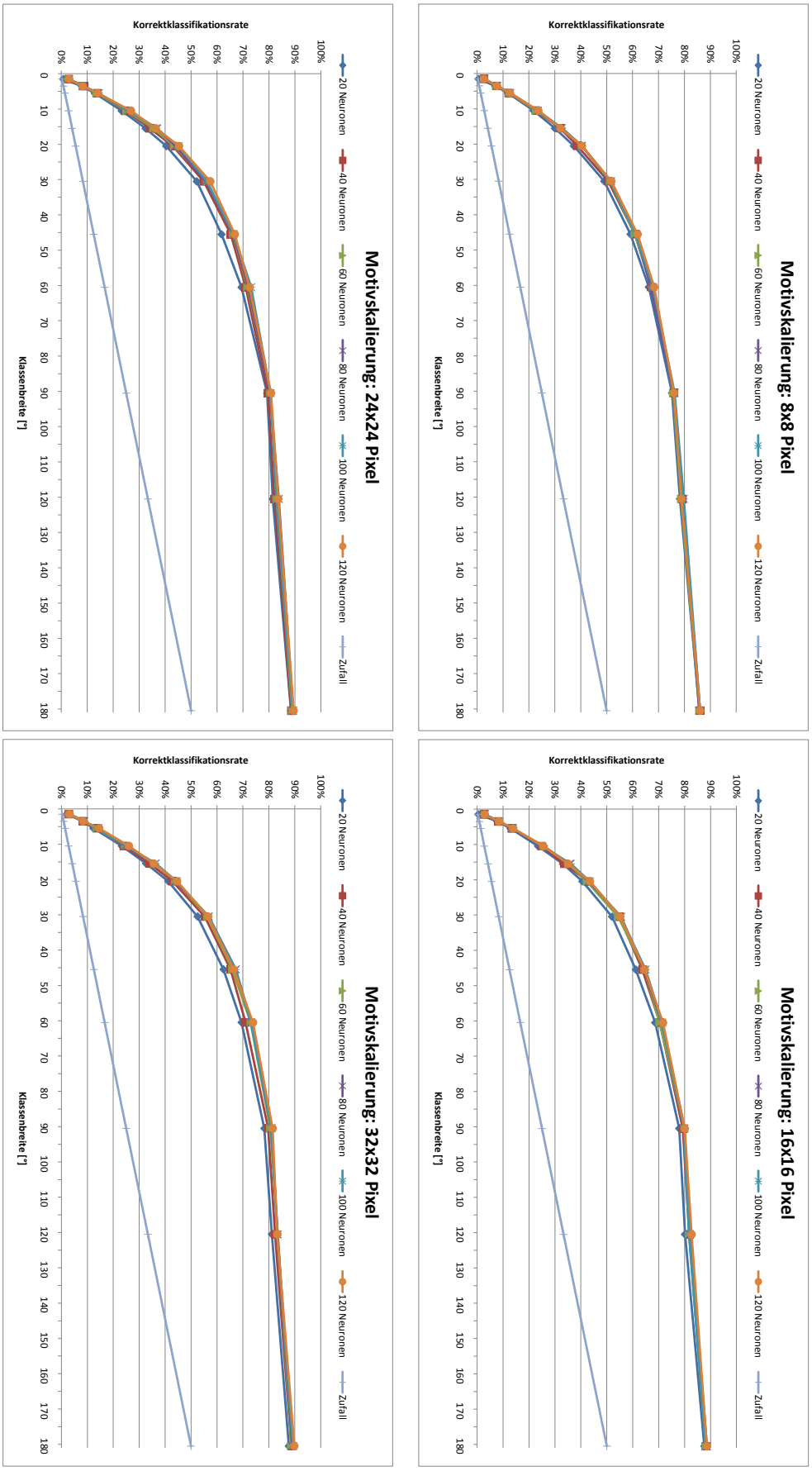


Abb. 3.13.: Korrektklassifikationsrate bei einzelkamerabasierter horizontaler Kopfdrehungsschätzung.

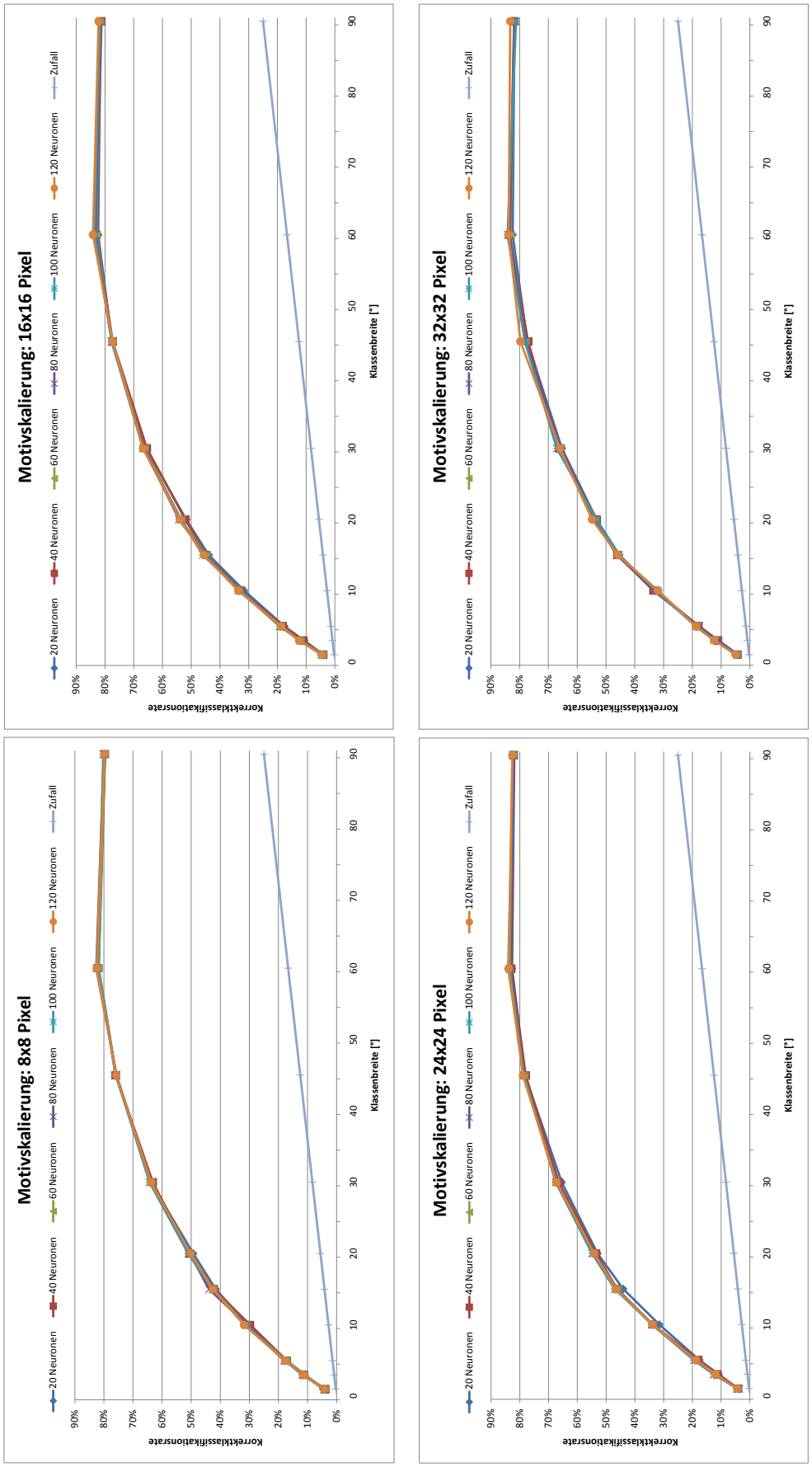


Abb. 3.14.: Korrektklassifikationsrate bei einzelkamerabasierter vertikaler Kopfdrehungsschätzung.

Klassenbreite [°]	KKR [%]	#verst. Neuronen	Skalierung [Pixel]
1	3,2	120	32×32
3	8,8	100	32×32
5	14,5	120	32×32
10	26,8	120	24×24
15	36,9	80	24×24
20	45,3	120	24×24
30	57,4	120	24×24
45	67,3	80	32×32
60	73,9	120	32×32
90	81,4	120	32×32
120	83,8	80	24×24
180	89,8	120	32×32

Tab. 3.2.: Parameterauswahl und jeweilige Korrektklassifikationsrate (KKR) für einzelkamerabasierte, horizontale Kopfdrehungsschätzung.

Klassenbreite [°]	KKR [%]	#verst. Neuronen	Skalierung [Pixel]
1	4,9	120	32×32
3	12,5	100	24×24
5	19,5	100	24×24
10	34,1	80	24×24
15	46,9	100	24×24
20	54,9	100	24×24
30	67,6	100	24×24
45	79,7	120	32×32
60	84,2	100	32×32
90	83,3	120	32×32

Tab. 3.3.: Parameterauswahl und jeweilige Korrektklassifikationsrate (KKR) für einzelkamerabasierte, vertikale Kopfdrehungsschätzung.

Die beste Parameterbelegung für die unterschiedlichen Diskretisierungsgrößen sind in den Tabellen 3.2 und 3.3 aufgeführt. Obwohl unwesentlich, dominieren dabei insbesondere Neuronenmengen ≥ 80 in der zweiten Schicht bei Skalierungsgrößen $\geq 24 \times 24$ Pixel. Die Tabellen geben so eine obere Grenze der theoretisch zu erwartenden Korrektklassifikationsrate bei den unterschiedlichen Topologien vor. Um insbesondere näher auf die Form der geschätzten Wahrscheinlichkeitsfunktion einzugehen, ist in Abbildung 3.15 die Verteilung uni-, bi- und multimodaler Schätzungen bei der horizontalen Anwendung dargestellt. Pro Diskretisierungsgranularität wurden die optimalen Parameterbelegungen aus Tabelle 3.2 übernommen (Skalierungsgröße, Anzahl Neuronen in versteckter Schicht). Dabei wird insbesondere deutlich, dass bis zu einer Klassengröße von einschließlich 15° fast ausnahmslos multimodale Hypothesen ausgegeben werden. Erst ab einer Granularität von 30° halten sich multimodale und bimodale

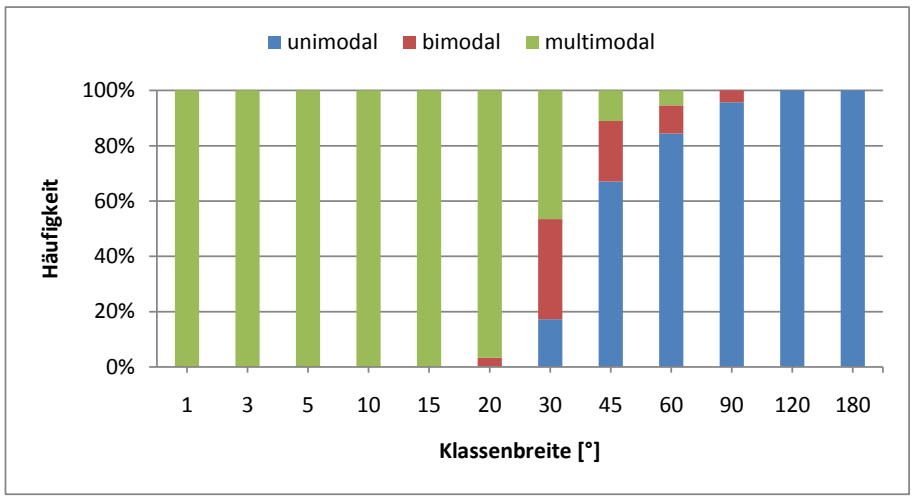


Abb. 3.15.: Häufigkeiten uni-, bi- und multimodaler Hypothesen bei horizontalen Drehwinkelschätzungen bei einer Skalierung von 32×32 Pixel und 120 versteckten Neuronen.

Schätzungen die Waage, ab 45° überwiegen erst unimodale Ausgaben. Das unterstreicht die vorigen Erkenntnisse, dass die Motivqualität per se keine detaillierte Winkelaufteilung unterstützt. Insbesondere scheint eine zu erwartende Winkelgenauigkeit bei 45° zu liegen, womit zumindest nach Tabelle 3.2 67% aller Motive korrekt klassifiziert werden. Dieses Resultat entspricht den wenigen in der Literatur beschriebenen Versuchen, unter ähnlichen Qualitätsbedingungen die Kopfdrehung zu schätzen [KBS00, NF96, RR98, WT00, ZPC02, OGX09, BT02]. Auch hier konnte bei Einzelansichten in der Regel keine höhere Winkelgenauigkeit erreicht werden. Eine multimodale Hypothese verdeutlicht die Unsicherheit der Schätzung, ein Kopfmotiv einer Drehwinkelklasse eindeutig zuzuordnen. Wie in den Ergebnissen gezeigt wurde, betrifft das im wesentlichen Winkelklassengrößen kleiner 15°. Hier verschiebt sich das Maximum der Wahrscheinlichkeitsfunktion zunehmend in Nachbarklasenbereiche oder wird von weiteren Modi dominiert. In Abbildung 3.16 ist die Verteilung der globalen Extremstellen in Form einer Konfusionsmatrix beschrieben. Zugrunde gelegt wurde hierbei die horizontale Schätzung mit einer Topologie von 10°-Winkelklassen, 120 Neuronen in der zweiten Schicht und einer Vorskalierung von 24×24 Pixel.
Aus Gründen der Übersichtlichkeit bei einer so hohen Anzahl vorhandener Klassen, wurde eine Hinton-Darstellung gewählt. Diese stellt ausgeprägtere Matrixkomponenten als größeres Rechteck dar, während Elemente mit niedrigem Betrag nur klein eingezeichnet sind. Die dominante Diagonale entspricht der verwechslungsfreien Klassifikation in die korrekte Drehwinkelklasse. Man sieht jedoch, dass die Klassifikation in die Nachbarbereiche der Diagonale diffundiert. Im Vergleich dazu treten nur wenige Verwechslungen in entfernt gelegene Winkelklassen auf.
Daran kann gesehen werden, dass die Problematik bei einer höheren Winkelgenauigkeit maßgeblich in der Verschiebung in direkte Nachbarbereiche liegt. Anstatt nur auf Einzelbildern zu schätzen, wäre in Videosequenzen ein aufgesetztes Trackingverfahren in der Lage diese Proble-

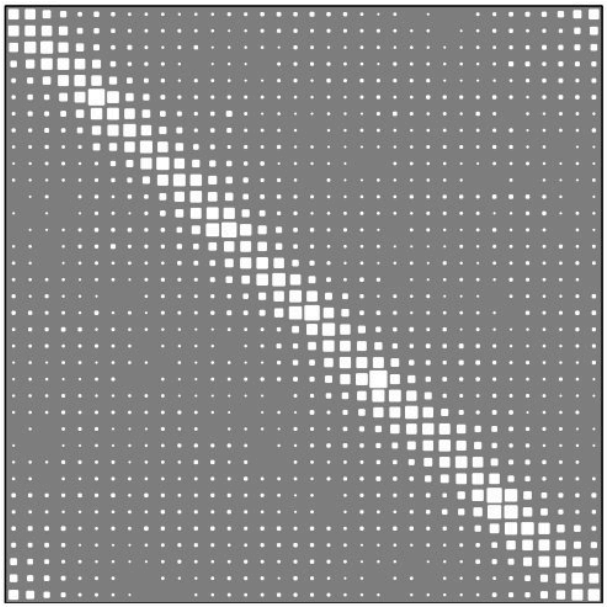


Abb. 3.16.: Hinton-Diagramm der Konfusionsmatrix bei einzelkamerabasierter Kopfdrehungsschätzung für horizontale Drehwinkel und einer Winkeldiskretisierung in 36 Klassen der Größe 10° (mit der in Tabelle 3.2 aufgeführten Skalierung in 24×24 Pixel große Eingabebilder und 120 Neuronen in der versteckten Schicht).

matik einzudämmen. Daneben stellt sich jedoch die Frage, inwieweit das gleichzeitige Berücksichtigen weiterer Kameraansichten die Dichtehypothesen beeinflusst. Dieser Fragestellung soll dabei insbesondere in den folgenden Abschnitten nachgegangen werden.

Einfluss von Lokalisierungsfehlern auf die Drehwinkelbewertung

In den bisherigen Evaluationen dieser Arbeit blieb der Einfluss der Lokalisierungskomponente auf die Drehwinkelbewertung unberücksichtigt. Den Neuronalen Netzen wurden exklusiv die manuellen Annotationen der Kopfregion im Datensatz vorgegeben. Mit dem Einlernen auf dieselben hat sich diese Komponente dabei sensibel gegenüber Translationen oder anderweitigen, fehlerbehafteten Lokalisierungen gezeigt. Der Einfluss suboptimal ausgerichteter Kopfmotive darf insofern nicht unberücksichtigt bleiben und soll im Folgenden untersucht werden.

Um die Neuronalen Netze mit fehlerbehafteten Lokalisierungen zu konfrontieren, wurde zunächst die Basis möglicher Fehlertypen ähnlich zu der Arbeit von Brown et al. [BT02] erfasst. Dabei handelt es sich im Einzelnen um die folgenden Quellen:

- **Asymmetrische Breitenskalierung**
  Der Fehler erfolgt in der Breite des Kopfmotivs. Die dem Gesicht zugewandte Seitenkante wird verschoben. Bei einer Reduktion gehen Gesichtsmerkmale verloren, bei einer Vergrößerung umfasst das Motiv zusätzlich zum Kopf auch den Hintergrund.

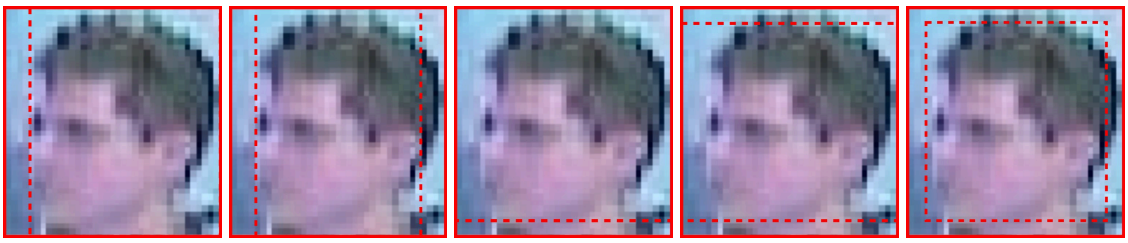
- **Symmetrische Breitenskalierung**
  Ähnlich zur asymmetrischen Skalierung, der Fehler betrifft jedoch sowohl die dem Hinterkopf als auch die dem Vorderkopf zugewandte Kante. Das Motiv wird auf beiden Seiten breiter oder schmaler.

- **Asymmetrische Höhenskalierung**
  Der Fehler erfolgt in der Höhe des Kopfmotivs. Die dem Kinn zugewandte, untere Kante des Motivs wird verschoben. Wie auch bei der Asymmetrischen Breitenskalierung gehen bei einer Reduktion der Motivhöhe Gesichtsmerkmale verloren, bei einer Vergrößerung wird zusätzliche Hintergrundinformation einbezogen.

- **Symmetrische Höhenskalierung**
  Sowohl die obere als auch untere Kante des Kopfmotivs werden verschoben.

- **Symmetrische Breiten- und Höhenskalierung**
  Alle Kanten des Motivs werden verschoben, das Motiv wird im Gesamten kleiner oder größer.

Kombinatorisch lassen sich aus diesen Fehlertypen alle weiteren Fehlerquellen nachbilden. Die Auswirkungen der Fehlerquellen auf die Drehwinkelschätzung ist dabei in den Abbildungen 3.18 und 3.19 dargestellt. Darin abgebildet sind die beobachteten Korrektklassifikationsraten verschiedener Winkelgranularitäten in Abhängigkeit vom Ausmaß des zugrunde gelegten Fehlertyps. Der Fehler wurde hierzu prozentual von der eigentlichen Breite beziehungsweise Höhe des Motivs angewandt: So entspricht der asymmetrische Breitenfehler bei $+5\%$ einer Verschiebung der dem Gesicht zugewandten Seitenkante um 5% der ursprünglichen Motivbreite. Eine Kopfbreite von 20 Pixel würde so um 1 Pixel seitlich vergrößert werden.
Mit der prinzipiell geringen Korrektklassifikationsrate feiner Winkelgranularitäten wirken sich die Fehlertypen entsprechend gering aus. Erst in den gröberen Diskretisierungen sind die Ein-



(a) (b) (c) (d) (e)

Abb. 3.17.: Die fünf evaluierten Lokalisierungsfehler: In den asymmetrischen Skalierungen ((a), (c)) wird nur eine Kante der Rechteckregion skaliert, in den symmetrischen Skalierungen ((b), (d), (e)) hingegen auch die gegenüberliegende. Kombinatorisch lassen sich aus diesen alle weiteren Fehlerquellen nachbilden.

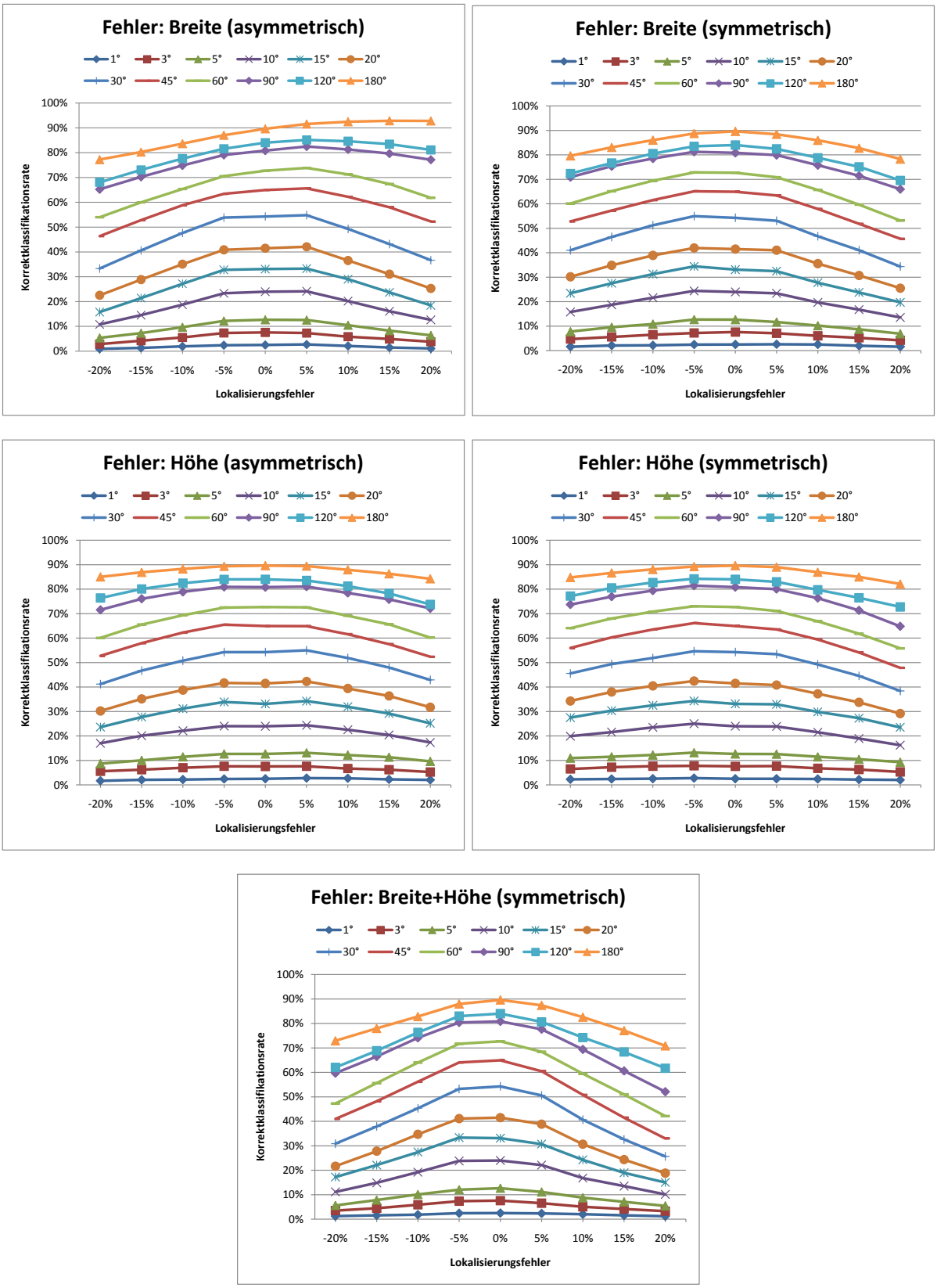


Abb. 3.18.: Korrektklassifikationsrate für die horizontale Kopfrehungsschätzung in Abhängigkeit von unterschiedlichen Lokalisierungsfehlern.

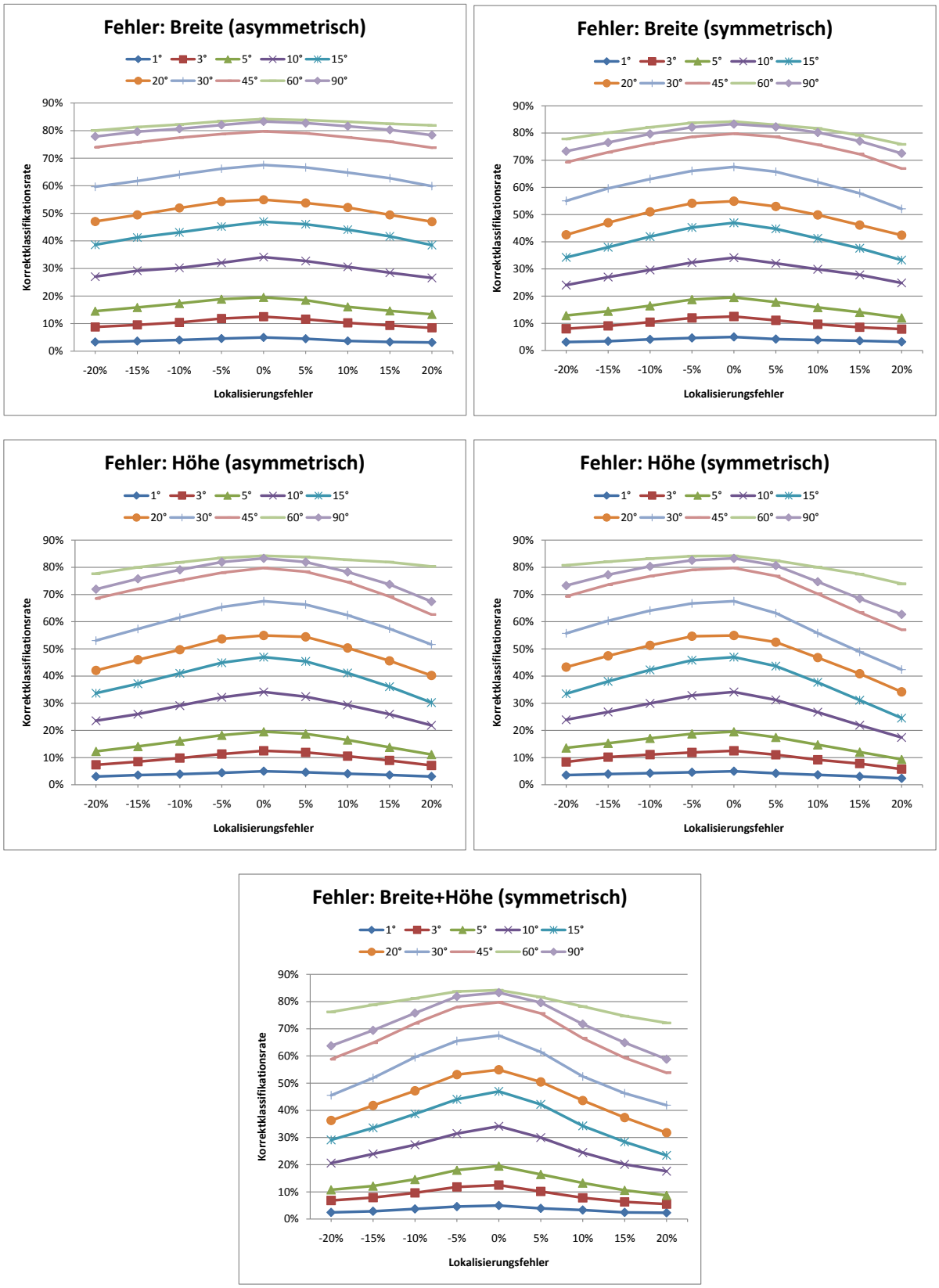


Abb. 3.19.: Korrektklassifikationsrate für die vertikale Kopfrehungsschätzung in Abhängigkeit von unterschiedlichen Lokalisierungsfehlern.

flüsse deutlich zu erkennen. Insbesondere zeigen sich die Netze bei niedrigen Skalierungsfehlern noch robust genug, jenseits der 5% jedoch deutlich beeinflusst. Bezogen auf die Kopfmotivgrößen im Datensatz entspricht das einer Fehlertoleranz von bis zu 3 Pixel. Hinsichtlich der zuvor evaluierten Lokalisierungsgenauigkeit der Gradientenhistogramme lässt sich damit zusammenfassen, dass die genannte Detektionsgenauigkeit von 6 Pixel mittlerer Abweichung bereits einen deutlichen Einfluss auf die Drehwinkelbewertung nehmen wird. Unter dem Aspekt der diversifizierten Zustandsraumabdeckung durch die Partikel, ist in der Praxis somit eine deutlich unzuverlässige Drehwinkelbewertung zu erwarten, die ihresgleichen durch die Lokalisierungsbewertung in ihrer Konfidenz reduziert werden muss. Zumindest in einem einzelkamerabasierten Anwendungsfall, kann jedoch durch eine gröbere Diskretisierung des Winkelwertebereichs gegengesteuert werden, um von robusteren und fehlertoleranteren Schätzungen ausgehen zu können.

3.3.3. Evaluation von Multikamera-Hypothesen

In Abschnitt 3.3.2 wurden die Teilkomponenten auf ihre Leistungsfähigkeit und eine optimale Parameterbelegung für den weiteren Verlauf untersucht. Dabei wurden sie jeweils nur auf Einzelbildern angewandt. Der sich eventuell resultierende Mehrwert bei einem gleichzeitigen Einbeziehen aller Kameraansichten, blieb dabei unberücksichtigt.

In diesem Abschnitt soll der Frage nachgegangen werden, ob der Nutzen weiterer Kameraansichten die Korrektklassifikationsrate der Teilkomponenten verbessern kann. Im besonderen interessiert dabei die Drehwinkelschätzung, die unter den gegebenen Bedingungen auf Einzelbildern bisher keine ausreichende Korrektklassifikationsrate für eine feingranulare Winkelhypothesenbildung erlaubt. Diese ist jedoch für ein zielgerichtetes Nachverfolgen der Aufmerksamkeitszuwendung im späteren Verlauf notwendig.

Rückführung der Winkel zum Raumbezugssystem

Die im System gemachten Zustandshypothesen, beschreiben die Kopforientierung im Bezug zum Weltkoordinatensystem. Wie in Abschnitt 3.1.4 dargestellt wurde, ist dabei eine Abbildung der jeweiligen Winkelkomponente in das Bezugssystem der Kamera erforderlich, um das System flexibel auf die Anzahl verfügbarer Kameraansichten reagieren lassen zu können. Die vorgeschlagene Abbildung ermöglicht dieselbe Bewertungsstrategie auf allen Ansichten gleichermaßen anwenden zu können.
Die Topologie der Neuronalen Netze verlangt jedoch eine Diskretisierung des Winkelwertebereichs in äquidistant große Klassen. Weil die Diskretisierung hierbei erst nach der Abbildung in das Kamerabezugssystem geschieht und damit auf dem von der jeweiligen Kamera beobachteten, relativen Drehwinkel aufsetzt, ist die Verteilung der Klassen über den Winkelwertebereich

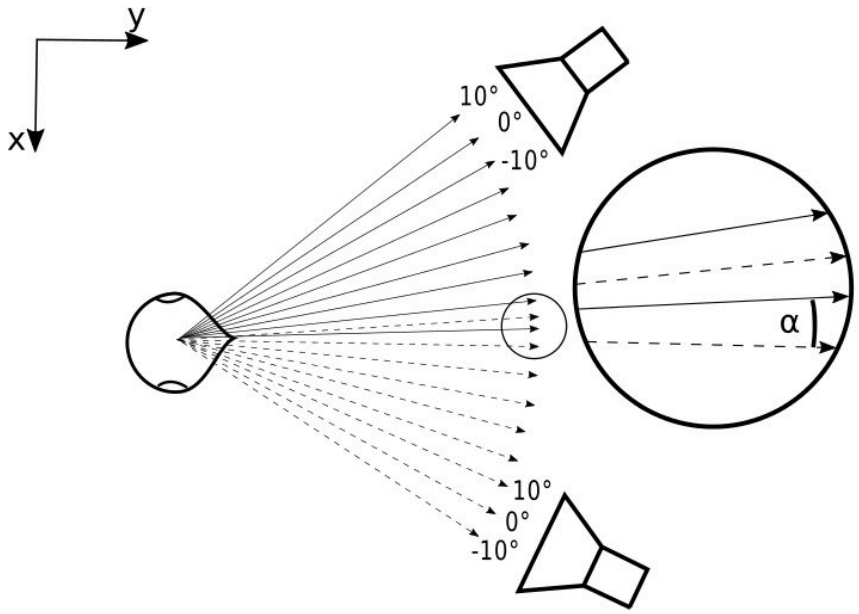


Abb. 3.20.: Schematische Darstellung der Winkeldiskretisierung in zwei verschiedenen Kameraansichten in Winkelklassen von 10° Größe. Die durchgezogenen Pfeile symbolisieren die Winkelklassen der oberen Kamera, die gestrichelten Pfeile die Klassen der unteren Kamera. Je nach Position des Kopfmittelpunkts erscheinen die Ränder der jeweiligen Winkelklassen deckungsgleich oder gegeneinander verschoben, wie hier im gezeigten Beispiel. α bezeichnet im Beispiel die Randungleichheit der beiden kameraabhängigen Winkeldiskretisierungen. Der gezeigte Effekt tritt ebenfalls für die weiteren Kameras im Raum auf; die Ausprägung von α tritt dabei variabel, abhängig von der Lage der Kameras und der Position des Kopfmittelpunkts auf und führt zu einer nicht konsistenten Diskretisierung des Winkelwertebereichs, wenn die Dichteschätzungen zurück in das Weltbezugssystem abgebildet werden.

sowohl abhängig von der Position der beobachteten Person als auch von der beobachtenden Kamerasicht. Die kamerarelativen Winkelklassen erscheinen bei einer inversen Abbildung zurück zum Weltbezug nicht mehr zwangsläufig deckungsgleich.
Abbildung 3.20 verdeutlicht diesen Sachverhalt noch einmal im Einzelnen. Für zwei Kameras ist darin die Aufteilung des horizontalen Winkelwertebereichs $[-180°, 180°)$ in diskrete Klassenabschnitte teilweise dargestellt. Zur Unterscheidung der kamerabezogenen Klassenbereiche, wurden die zur oberen Kamera gehörenden Abschnitte mit durchgezogener Linie, die zur unteren Kamera gehörenden mit gestrichelter Linie symbolisiert. Aus Gründen der Übersichtlichkeit ist die Überlappung der jeweiligen Winkelklassen nur in einem Teilsegment hervorgehoben. Der Winkel α bezeichnet darin die Verschiebung der Klassenränder, die zwischen den beiden Kamerabezügen auftritt. Eine entsprechende Verschiebung wäre darüber hinaus für weitere Kameras vorzufinden, ist jedoch zur besseren Übersicht im Beispiel vernachlässigt worden. Je nach Position des Kopfmittelpunkts und Lage der Kameras, ändert sich die Ausprägung des Winkels α deutlich. Die Diskretisierung erscheint demzufolge in Einzelansichten sinnvoll, verhindert jedoch eine allgemeingültige Übertragbarkeit zurück auf die Zustandshypothesen

im Weltbezug: Die Deckungsungleichheit der Winkelklassen führt dazu, dass bei inkonstantem α die für das System fusionierte Winkelgranularität ebenfalls inkonstant erscheint. Die finale Zustandsbewertung geschieht folglich nicht mit derselben Auflösung wie in den jeweiligen Einzelansichten, sondern muss hinreichend gering ausfallen, um den zurück abgebildeten Klassenverschiebungen der Kamerabezüge gerecht zu werden. Im Folgenden soll daher von einer minimalen Winkelzustandsdiskretisierung in 1°-Schritten ausgegangen werden. Die Klassenwerte der Wahrscheinlichkeitsschätzungen werden so in die ihnen entsprechenden Klassen im Weltbezugssystem akkumluiert.

Informationszugewinn durch komplementäre Ansichten

Der durch den gleichzeitigen Nutzen verschiedener Kameraansichten gewünschte Effekt ist, dass durch komplementäre Blickwinkel auf den Kopf, der insgesamt vorliegende Informationsgehalt steigt. Damit sollen fusionierte Schätzungen robuster werden, während die Winkelauflösung feiner gewählt werden kann.
Wie in den bisherigen Evaluationen der Winkelbewertungen festgehalten werden konnte, kann die Klassengröße der Winkeldiskretisierung beliebig vordefiniert werden. Bei zu hoher Auflösung aber steigt die Anzahl der Fehlklassifikationen in benachbarte Winkelklassen. Ebenfalls wächst die Anzahl multimodaler Schätzungen, was zu einem erhöhten Rauschen um den eigentlich zu erkennenden Winkelwert führt. In den Abbildungen 3.21 und 3.22 sind die Zuwächse der Korrektklassifikationsrate dargestellt, wenn verschiedene Kameraschätzungen akkumuliert werden. Dabei wurde entsprechend der vorigen Ergebnisse auf den manuell anno-

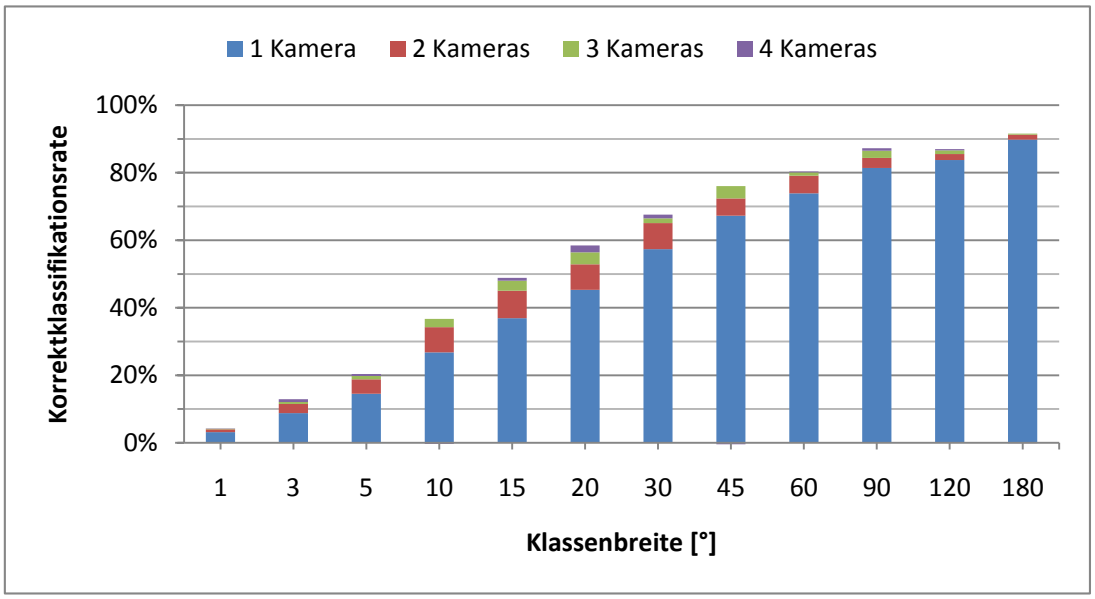


Abb. 3.21.: Steigerung der Korrektklassifikationsrate horizontaler Kopfdrehung bei Hinzunahme weiterer Kameraansichten.

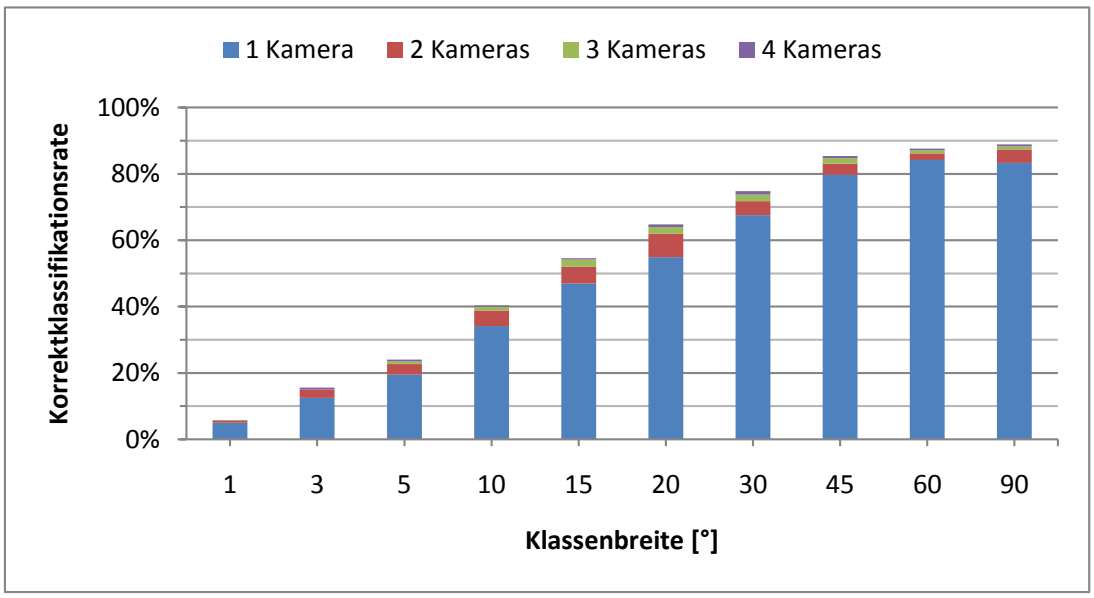


Abb. 3.22.: Steigerung der Korrektklassifikationsrate vertikaler Kopfdrehung bei Hinzunahme weiterer Kameraansichten.

tierten Kopfregionen im Datensatz getestet. Für die jeweiligen Diskretisierungsgranularitäten wurden die optimalen Parameterbelegungen aus den Tabellen 3.2 und 3.3 benutzt. Damit die maximal zu erwartende Steigerung der Korrektklassifikationsrate sichtbar wird, wurden für die jeweilige Anzahl zu fusionierender Ansichten alle Fusionspermutationen getestet und die jeweils beste ausgewählt. Für die Fusion zweier Ansichten wurden demnach die Kamerakombinationen $\{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$ betrachtet und nur diejenige berücksichtigt, die den höchsten Zuwachs der Korrektklassifikationsrate vorwies. Entsprechend den vorherigen Evaluationen wurde hierzu auch in Form der Leave-One-Out-Kreuzvalidierung das eingesetzte Neuronale Netz nicht auf der zu evaluierenden Person eingelernt. Wie man sehen kann, sind Verbesserungen mit bereits einer weiteren Kameraansicht ab einer Winkelklassengröße von 3° erkennbar. Mit einer dritten Kamera wird ein Mehrwert ab einer Größe von 10° deutlich. Daran werden für die unterschiedlichen Mengen vereinter Ansichten untere Grenzen sichtbar, ab denen kein tatsächlicher Zuwachs an Information mehr zu erwarten ist. Der Grund liegt in den nicht deckungsgleichen Modalstellen der geschätzten Wahrscheinlichkeitsfunktionen. Weil auch insbesondere diejenigen Kamerapermutationen keine Verbesserungen bringen, die die übrigen Ansichten berücksichtigen, kann daraus geschlossen werden, dass bei zu feiner Diskretisierung ein zu hohes Rauschen verursacht wird, das der erhofften Erhöhung der Korrektklassifikationsrate entgegenwirkt. Der maximale Zuwachs bei bis zu vier Kameras ist bei der horizontalen Komponentenschätzung bei einer Klassengröße von 20° erkennbar. Dabei stellt der Nutzen aller Ansichten gemeinsam die deutlichste Steigerung der Korrektklassifikationsrate dar. Für Klassengrößen jenseits 20° nimmt der Zuwachs bei vier Ansichten wiederum ab - sogar der Mehrwert bei einer Fusion mit drei Kameras ist sichtbar reduziert. Hier liegt die Be-

gründung in der sowieso groben Winkelklassengröße, die Fehlklassifikation in Nachbarklassen an sich bereits verhindert und damit auch bei Hinzunahme weiterer Ansichten keinen weiteren Zugewinn ermöglicht.

Es lässt sich feststellen, dass grobe Winkelklassen zu genaueren Einzelkameraschätzungen führen und damit allenfalls durch eine weitere Ansicht Fehlklassifikationen reduziert werden können. Mit kleiner werdender Klassengröße werden die geschätzten Hypothesen allerdings so stark verrauscht, dass sie voneinander grundsätzlich unterschiedliche Modalwerte besitzen. Das führt dazu, dass auch bei einer Fusion das Rauschen entweder nur unwesentlich reduziert werden kann oder ein gemeinsamer, globaler Extremwert auch übereinstimmend noch immer in den Nachbarklassen zu liegen kommt. Das bestätigt auch die in Abbildung 3.23 dargestellte Konfusionsmatrix nach der Fusion: generelles Rauschen in weit entfernten Winkelklassen wird deutlich reduziert. Die Varianz um die eigentliche Diagonale wird darüber hinaus zwar geringer, bleibt jedoch bestehen. Bei der der Matrix zugrunde gelegten Winkelauflösung von 10° großen Klassen kann so allein durch eine Fusion der Schätzungen eine Verbesserung von $\sim 10\%$ erwartet werden, die aber maßgeblich durch die Reduktion der Fehlklassifikationen im weiten Umfeld der Diagonalen erreicht wird. Prinzipiell bleibt also je nach Auflösung der Diskretisierung, trotz Fusion, eine unvermeidbare Diffusion des Modalwerts in Nachbarbereiche der eigentlich korrekten Winkelklasse bestehen. Die Fusion führt hingegen maßgeblich zu einer Reduktion des Rauschens über den gesamten Winkelwertebereich. Um fusionierten Fehlklassifikationen entgegenwirken zu können, stellt sich nun die Frage, wie die im System beschriebene Filterung in Form des Zustandtrackings beitragen wird. Bildlich ausgedrückt, soll damit die Streuung um

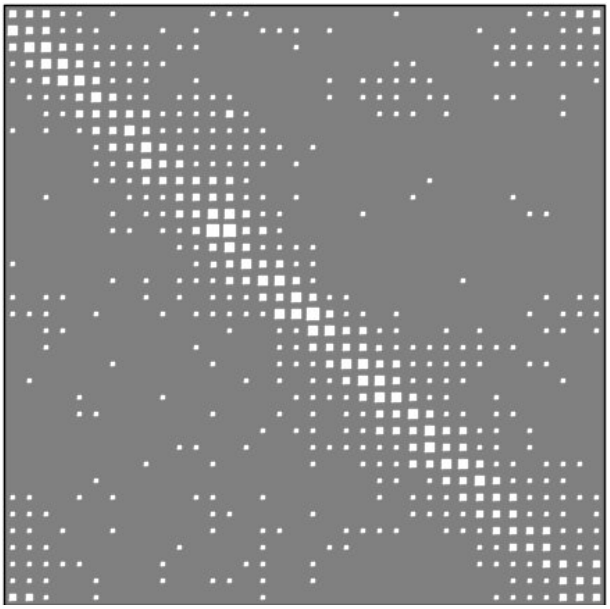


Abb. 3.23.: Hinton-Diagramm der Konfusionsmatrix bei mehrkamerabasierter Kopfdrehungsschätzung für horizontale Drehwinkel.

die Diagonale der Konfusionsmatrix weiter eingeschränkt werden. Infolgedessen würde dies auch zu einer Erhöhung der Korrektklassifikationsrate führen.

Filterung der Fehlklassifikationen durch Zustandstracking und Zusammenführen der Komponenten

Die Ergebnisse bei Zusammenführung aller Teilkomponenten und Einsetzen des beschriebenen Zustandstrackings sind in den Abbildungen 3.24 und 3.25 dargestellt. Aus Gründen der Übersichtlichkeit sind die detaillierten Parameterbelegungen und Fehlerergebnisse in den Tabellen C.1, C.2, C.3, C.4 und C.5 (ab Seite 176) in Anhang C aufgeführt und in den Abbildungen nur der mittlere Fehler dargestellt, der minimal bei den darin angegebenen Parametereinflüssen beobachtet wurde. Die dabei zum Einsatz gekommenen Netztopologien orientierten sich an den entsprechenden Parameterbelegungen aus Tabelle 3.2 und 3.3.
Für die Evaluation wurde von einer maximalen Winkeldiskretisierung von 15° ausgegangen, um eine hinreichend feine Erfassung der Blickfeldorientierung für die nachfolgende Aufmerksamkeitszuwendung sicherstellen zu können. Neben der Winkelgranularität wurde auch die Menge eingesetzter Partikel als Stützstellenabtastungen im Zustandsraum parametrisiert und der Einfluß der Standardabweichung des normalverteilten Rauschterms, der auf die Partikeldiffusion aufaddiert wird. Insgesamt wurde mit 250, 500 und 1000 Partikel evaluiert und Rauschen mit den Standardabweichungen $(\sigma_{pan}, \sigma_{tilt})$=$\{(10°, 3°), (15°, 5°), (20°, 10°), (25°, 15°), (30°, 20°)\}$ berücksichtigt. Erkennbar ist, dass der niedrigste Fehler bei einer Winkeldiskretisierung von 10° und 1000 Partikeln für die horizontale Schätzung und 5° und 1000 Partikeln für die vertikale Schätzung beobachtet werden konnte. Er liegt jeweils bei $5,3°$ beziehungsweise $8,69°$. Weil die horizontale Genauigkeit für die Zielunterscheidung relevanter ist, soll für

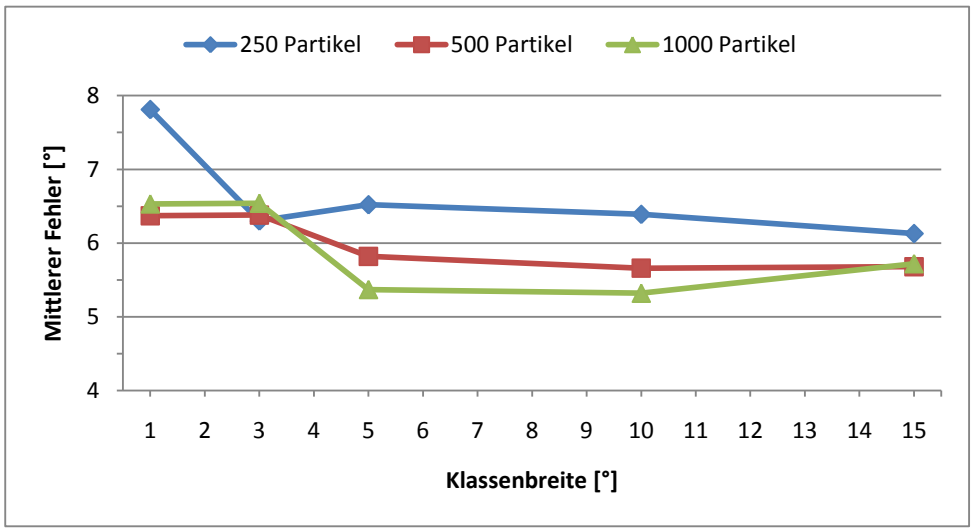


Abb. 3.24.: Mittlerer horizontaler Fehler des Systems bei verschiedener Winkeldiskretisierung und Partikelanzahl.

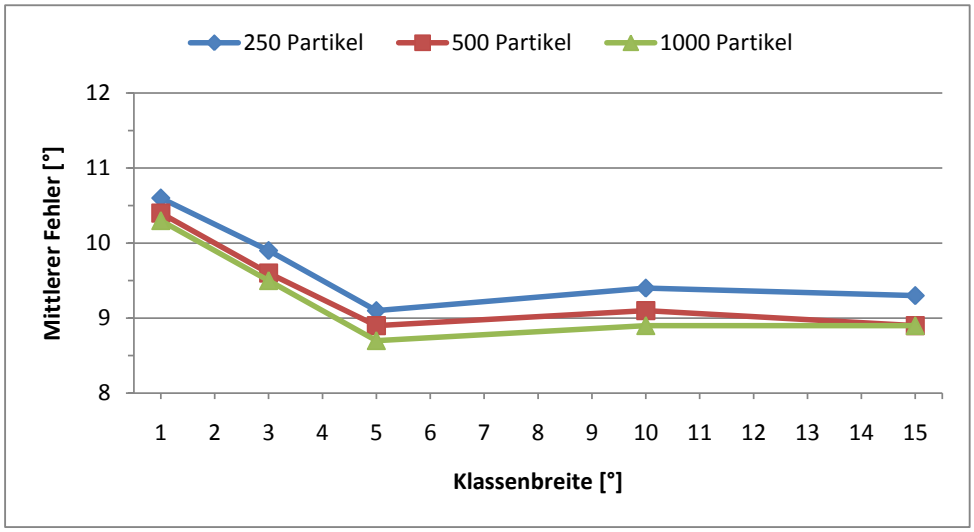


Abb. 3.25.: Mittlerer vertikaler Fehler des Systems bei verschiedener Winkeldiskretisierung und Partikelanzahl.

die im Folgenden beschriebene Zuwendungsschätzung der Aufmerksamkeit daher das System eingesetzt werden, das für den minimalen horizontalen Fehler verantwortlich ist.

Gegenüberstellung additiver und multiplikativer Merkmalsfusion

Die in Abschnitt 3.1.4 beschriebene Merkmalsfusion umfasste die multiplikative und damit gleichwertige Einbeziehung der beiden Merkmale (gemeinhin Lokalisierung und Drehwinkelschätzung), stellte aber auch die Möglichkeit einer additiven Fusion dar, in der die beiden Merkmalsbewertungen gewichtet gemittelt werden. Wie diskutiert wurde, ist die gewichtete Addition in der Bildverarbeitung etabliert und unter gewissen Bedingungen, die in diesem Rahmen als gegeben angesehen werden können, auch annähernd korrekt. Im Unterschied zur multiplikativen Fusion sichert die additive, dass nicht die gesamte Bewertung wegfällt, wenn ein einzelnes Merkmal ausfällt.
Durch die vorzugebene Gewichtung ist die Kombination der Merkmale statisch und baut prinzipiell auf Heuristiken auf, welches Merkmal letztendlich dominanter einzubeziehen ist. Es existieren zwar Verfahrensansätze, in denen die Gewichtung dynamisch zur Laufzeit adaptiert wird (siehe zum Beispiel [Nic08]), deren Verwendung würde jedoch den Rahmen dieser Arbeit sprengen. Aus Gründen der Vollständigkeit soll an dieser Stelle deswegen lediglich der Einfluss des Fusionsfaktors bei einer statischen additiven Merkmalsfusion und einer statischen multiplikativen evaluiert werden. Wie in den Abbildungen 3.26 und 3.27 dargestellt ist, liegt der mittlere Fehler bei additiver Merkmalsfusion sowohl für die horizontale als auch für die vertikale Drehwinkelschätzung über dem der multiplikativen. Evaluiert wurden die jeweiligen Fusionsgewichte $\lambda^{NN} = \{0,1,0,3,0,5,0,7,0,9\}$ für die Bewertungsstrategie der Neuronalen Netze - das Gewicht für die Gradientenhistogramme wurde respektive auf $(1-\lambda^{NN})$ gesetzt.

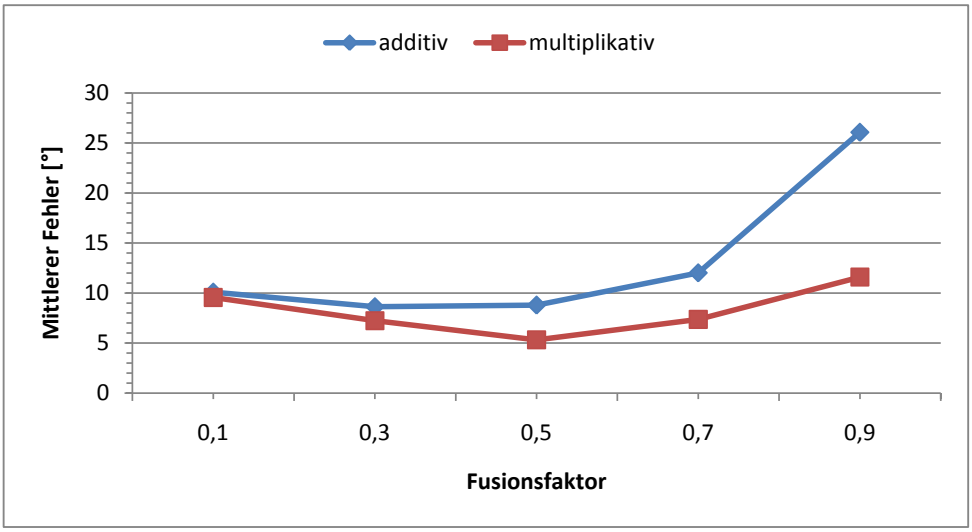


Abb. 3.26.: Gegenüberstellung des horizontalen Fehlers bei additiver und multiplikativer Merkmalsfusion, bei einer Winkeldiskretisierung von 10° und einer Tabelle 3.2 entsprechenden Netztopologie.

Ein Minimum des Fehlers konnte bei der horizontalen Drehwinkelschätzung bei $0,3$ festgestellt werden, bei der vertikalen bei $0,5$. Im Fall der horizontalen lässt sich das damit interpretieren, dass eine stärkere statische Dominierung der Gradientenhistogrammgewichtung (in diesem Fall $0,7$) zu einer verbesserten Schätzung führte, was grundlegend die Notwendigkeit einer konsistenten Lokalisierung der Kopfregion unterstreicht. Die Streuung der Partikelpositionen wog folglich schwerer als die Streuung der Winkelhypothesen, weil suboptimale Rechteckregionen um den Kopf zu einem starken Rauschen der Neuronalen Netze führte und die Winkelbewertung damit obsolet werden ließ.

Dass der Fehler im Vergleich zur multiplikativen Merkmalsfusion dennoch größer ausfällt liegt daran, dass nicht mit einem Ausfall eines der beiden Merkmale zu rechnen ist - eine Tatsache, die dem Datensatz angerechnet werden muss weil dies darin vorgesehen war: Der Datensatz wurde so aufgenommen, dass immer verdeckungsfreie Ansichten zu den beobachteten Personen gewährleistet sind und keine Veränderung der Beleuchtung oder anderer optischen Einflüsse auftritt. Damit kann zu jedem Zeitpunkt eine Lokalisierungsbewertung mit den Gradientenhistogrammen durchgeführt und das Kopfmotiv in das Neuronale Netz eingespeist werden. Ferner sind beide Merkmale prinzipiell äquivalent zu gewichten, weil sie stochastisch unabhängig voneinander sind. Verdeutlicht wird das durch den minimalen Fehler bei der multiplikativen Fusion mit identischer Gewichtung beider Komponenten.

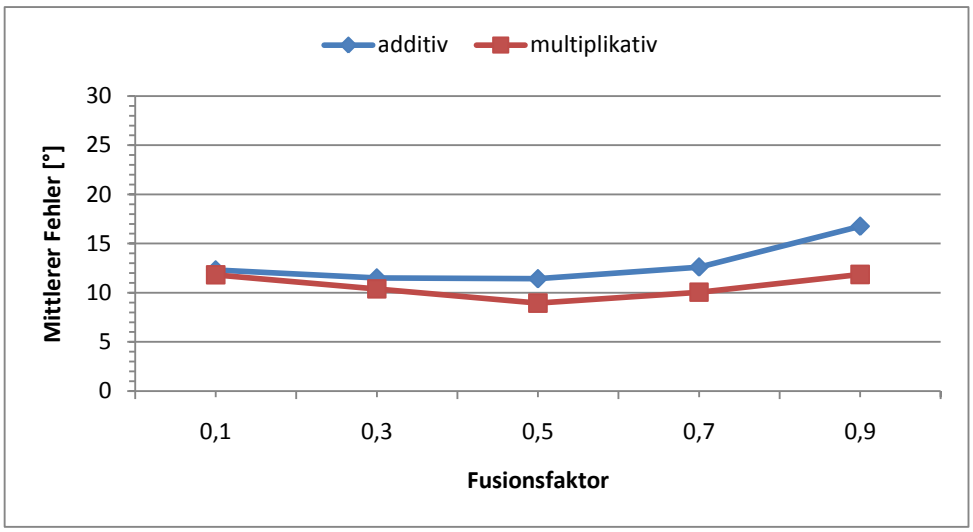


Abb. 3.27.: Gegenüberstellung des vertikalen Fehlers bei additiver und multiplikativer Merkmalsfusion, bei einer Winkeldiskretisierung von 10° und einer Tabelle 3.3 entsprechenden Netztopologie.

3.3.4. Alternativer Referenzansatz: Winkelregression nach vorangehender Kameraselektion

Mit dem entworfenen Ansatz werden alle Kameraansichten gleichermaßen in die Schätzung einbezogen. Das erfordert eine erhöhte Rechenleistung, weil die vierfache Menge an Merkmalen berechnet werden muss. Um dem entgegenzutreten liegt ein Referenzansatz darin, eine Kameraselektion voranzustellen, die die tatsächlich zu benutzende Menge an Ansichten auf Vorderkopf- oder Profilaufnahmen beschränkt. Ein solcher Ansatz wurde zu Vergleichszwecken implementiert und soll an dieser Stelle als Referenzsystem den Ergebnissen gegenübergestellt werden.

Detektion von Vorderköpfen

Weil Hinterkopfansichten nicht genügend Merkmale für eine detaillierte Drehwinkelschätzung darbieten, sollen Kopfmotive detektiert werden, die bis maximal ins Profil gedreht erscheinen. Hierzu werden Detektorkaskaden auf Basis von Haar-Merkmalen implementiert, die auf den Arbeiten von Viola und Jones aufsetzen [VJ01]. In der Bildverarbeitung ist dieses Vefahren mitunter durch seine Echtzeitfähigkeit populär geworden, weil durch die kaskadenartige Klassifikationsstruktur schnell negative Beispiele im Bild verworfen werden können und zusätzlich durch die Merkmalsberechnung ein gesamtes Kamerabild in hoher Geschwindigkeit abgetastet werden kann. Der Detektoransatz ist darüberhinaus in der weitverbreiteten Softwarebibliothek *OpenCV* [Bra00] in der Bildverarbeitung implementiert und steht damit einer breiten Masse von Anwendern leicht zugänglich zur Verfügung. Aus diesem Grund wird jene Implementierung herangezogen um einen Detektor einzulernen, der Vorderkopf- und Profilansichten in einem

Kamerabild detektiert. Hierzu wird die Standardparameterbelegungen der Implementierung belassen und pro Personenvideo im Datensatz - wie auch zuvor - anhand einer Kreuzvalidierung jeweils ein Detektor auf den übrigen Videos eingelernt.

Assoziation der Detektorergebnisse und dreidimensionaler Bezug

Der eingelernte Detektor wird auf jedem Kamerabild angewandt und liefert hypothetisierte Rechteckregionen erkannter Kopfansichten zurück. Weil die relative Drehwinkelschätzung im Anschluss die dreidimensionale Position des Kopfs benötigt, um den Winkel in einen Bezug zum Raumkoordinatensystem stellen zu können, müssen die Detektionen von mindestens zwei Kameraansichten miteinander assoziiert werden, um durch eine Triangulation auf die dreidimensionale Position des Kopfs schließen zu können. Die Anbringung der Kameras erlaubt ein solches Vorgehen, weil durch deren Blickwinkel mindestens in einer weiteren Ansicht stets eine Profilaufnahme zu erwarten ist, wenn der Kopf einer Kamera frontal zugedreht erscheint. Hierzu werden die Detektionen paarweise miteinander assoziiert, die von benachbarten Kamerapaaren stammen, und deren jeweiliger Mittelpunkt aus den beiden Ansichten trianguliert. Der Abweichungsfehler der Triangulation wird schließlich als Indikator benutzt um nicht übereinstimmende Detektionen auszusortieren. Jene Rechteckregion mit dem niedrigsten Abweichungsfehler wird schließlich als Kopfhypothese zur Winkelschätzung weitergereicht und die berechnete Position im Raumkoordinatensystem als Position des Kopfs interpretiert.

Drehwinkelschätzung nach vorangehender Kameraselektion

Die assoziierten Detektionen bestehen aus Rechteckregionen, die denselben Bereich im beobachteten Raum aus unterschiedlichen Kamerablickwinkeln beschreiben und in der ein bis maximal ins Profil gedrehter Kopf aus den beiden Ansichten detektiert werden konnte. Die Rechteckregionen werden deswegen als Lokalisierung des Kopfs herangezogen und in ein Neuronales Netz eingespeist, das den Drehwinkel des beobachteten Kopfs schätzt. Hierzu wurde jeweils für die zu schätzende Winkelkomponente ein Regressor eingelernt, der mit einem Ausgabeneuron den kontinuierlichen Winkelwert des angelegten Kopfmotivs hypothetisiert. Der an dieser Stelle zum Kamerasystem bezogene Winkel wird schließlich in den Bezug zum Raumkoordinatensystem gebracht und die Schätzungen aller Ansichten, die im Vorfeld selektiert wurden, für eine endgültige Hypothese gemittelt. Weil die Hypothesen zu diesem Zeitpunkt noch ungeglättet sind, wurde eine ebenfalls aus der OpenCV-Bibliothek stammende Implementierung eines Kalmanfilters darauf angewandt.
Der schließlich zu beobachtende mittlere Fehler beider Referenzsysteme ist in Tabelle 3.4 aufgeführt. Wie dabei unschwer zu erkennen ist, liegt der Fehler weitaus höher als im eigentlichen Partikelfilteransatz - jenem System das dem im Rahmen dieser Arbeit vorgeschlagenen Verfahren entspricht. Die Gründe dafür liegen vor allem in einer inkonsisten Kopflokalisierung, die

	Referenzsystem	Referenz (Kalmanfilter)	Partikelfilter
Fehler horizontal [°]	48,3	53,9	5,3
Fehler vertikal [°]	30,3	30,1	8,9

Tab. 3.4.: Gegenüberstellung des Systemfehlers zum implementierten Referenzansatzes. Detaillierte Auflistung des Fehlers über alle einzelnen Personenvideos: siehe Tabelle C.6 (Seite: 179).

aufgrund der Sensibilität der Neuronalen Netze zu erhöhtem Rauschen führt und einer fehlerbehafteten Kameraselektion, so dass hier Hinterkopfansichten einfließen, obwohl Vorderkopf- oder Profilaufnahmen erwartet werden.
Die inkonsistente Kopflokalisierung resultiert daher, dass die Detektorkaskaden mit derselben Motivgröße eingelernt wurden, wie die Bildskalierung für die Neuronalen Netze vorgibt. Das resultiert in Detektionen, die vorrangig der vorgegebenen Breite und Höhe entsprechen und damit vermehrt Hintergrund einbeziehen als notwendig, statt die detektierte Region eng um den Kopf anzupassen und erst im Anschluß durch die Skalierung auf die vorgegebenen Bildmaße zu verändern. Entsprechend den Experimenten in Abschnitt 3.3.2, geben die Regressoren dadurch Schätzungen aus, die einem erhöhtem Rauschen unterliegen, was in folge zu starken Abweichungen führt und kameraübergreifend inkonsistente Winkelwerte verursacht.
Der zweite Fehlergrund, der die eigentliche Kameraselektion betrifft, resultiert von Hinterkopfansichten, die fälschlicherweise als positive Beispiele klassifiziert wurden. In den Experimenten war dies mitunter darin begründet, dass die Detektorkaskaden sich vorrangig auf die dominanten Kanten zwischen Haar- und Hautfarbenbereich sowie zwischen Haarbereich und dem hierzu heller erscheinenden Hintergrund konzentrierten und einlernten. In der Schlußfolge waren Hinterkopfansichten durch die dominante Kantendarstellung des Haarbereichs und des Hintergrunds bevorzugt worden, wenn keine weniger dominant auftretenden Gesichtsmerkmale aufgrund der niedrigen Auflösung detektiert werden konnten. Lagen dazu die in den übrigen Ansichten detektierten Kopfregionen inkonsistent zueinander, weil der Triangulationsfehler wegen Verschiebungen oder Größenunterschiede in den Ergebnissen zu hoch ausfiel, dann wurden fälschlicherweise Kameraansichten ausgewählt, die in der Tat Hinterköpfe einfingen und folglich ungeeignet für eine nachfolgende Drehwinkelschätzung waren. Auch hier führte das schließlich dazu, dass die gemittelte Winkelhypothese stark von der bisherigen Historie abwich.
Mit den verrauschten Hypothesen ist es auch mitunter zu begründen weshalb die gefilterten Ergebnisse des Referenzsystems einen höheren Fehler vorweisen: Zum einen ist der Kalmanfilter für Kopfdrehungen ungeeignet, weil er in seiner Ursprungsform eine lineare Prozessdynamik voraussetzt und dies bei Kopfdrehungen nicht erwartet werden kann (siehe Diskussion in Abschnitt 3.1). Zum anderen sind die Schätzungen zu verrauscht, um sie durch Filtern den eigentlichen Winkelwerten annähern zu können. In Folge werden die Winkelhypothesen zwar ruhiger und schwanken weniger stark wie in den originalen Schätzungen, passen sich Ausreißern aber

auch weniger stark an und folgen einzelnen Schätzungen mit konsistenten Kopflokalisierungen und optimaler Kameraselektion nur unwesentlich. Das Rauschen wird im Gesamten minimiert, im Mittel passen sich die gefilterten Werte aber den dominanten Fehlschätzungen an.

4. Der Visuelle Aufmerksamkeitsfokus

Die Blickrichtung einer Person, insbesondere das Zielobjekt auf das diese Person schaut, gibt wichtigen Aufschluss über den Kontext, in dem diese Person agiert. In Diskussionen fühlen sich Menschen nicht nur dann angesprochen, wenn sie im Dialog adressiert werden, sondern auch wenn der Sprecher Blickkontakt zu seinem Gegenüber sucht. Während Tätigkeiten deutet die Blickrichtung darauf hin, auf welchen Gegenstand sich der Tätige konzentriert. Ein Vortragender erkennt an den Blicken des Publikums ob diese ihm folgen oder abgelenkt sind.
Der Blick ist ein wichtiges Instrument - nicht nur zur empathischen Nachvollziehbarkeit menschlicher Aktivitäten. Für ein umfassendes Verständnis über eine zu beobachtende Situation ist es deswegen unersetzlich die Blickrichtung als Indikator einzusetzen, um festzustellen wohin Personen ihre Aufmerksamkeit richten.
In der kognitiven Psychologie wird die *Aufmerksamkeit* in einen Bezug zum Arbeitsgedächtnis gesetzt [OB09]. Die Aufmerksamkeit wird dabei als Mechanismus beschrieben, mit der (kognitive) Aktionen ausgewählt und anschließend gesteuert werden [All87]. Damit stellt sie einen Bewusstseinszustand dar, der dafür verantwortlich worauf eine Person ihre Konzentration richtet.
Der Begriff der *visuellen Aufmerksamkeit* bezieht sich dagegen direkt auf die Blickrichtung und das damit verknüpfte Ziel, auf das der Blick gelenkt wird. Mit dem Nachvollziehen des visuellen Aufmerksamkeitsfokus wird der systematische Prozess beschrieben, der ausgehend von der Blickrichtung einer Person auf ihr Interaktionsgegenüber oder ihren Aktionsgegenstand schließt.
Geprägt wurde der Begriff der visuellen Aufmerksamkeit dabei unter anderem von Langton et al., die das Nachvollziehen der Blickrichtung als erste in einen Zusammenhang mit der Kopfdrehung brachten, um ein effektiveres Verständnis der visuellen Aufmerksamkeitszuwendung einer Person zu erhalten[LWB00]. Demnach ist das alleinige Beobachten der Pupillen nicht ausreichend, um eine effektive Aussage über die Aufmerksamkeitszuwendung zu erhalten. Stattdessen wird diese erst durch weitere, sekundäre Indikatoren hergestellt, dementsprechend Beobachter erst anhand der Kopfdrehung und Gesten den sozialen Bezug herstellen und auf einen möglichen Interaktionspartner schließen können.
Im Rahmen dieser Arbeit wird der Problemstellung nachgegangen, die Zuwendung der visuellen Aufmerksamkeit einer Person in aufmerksamen Umgebungen automatisch und kameragestützt zu erkennen. Mit dem Anwendungsfall verbunden ist eine Anbringung der Kameras, die das Geschehen im Raum aus unterschiedlichen Blickwinkeln beobachten. Aufgrund der Kame-

raausrichtungen können dabei weder zuverlässige Vorderkopfansichten erwartet werden, noch detaillierte Aufnahmen der Gesichtsmerkmale, die ein Beobachten der Pupillen jener Person ermöglichen. Eine Bestimmung der visuellen Aufmerksamkeit muss damit statt auf der beobachteten Blickrichtung, auf rein sekundären Merkmale erfolgen. Hierzu wurde im vorigen Kapitel ein System vorgestellt, das es ermöglicht die Kopfdrehung einer Person visuell zu erkennen. Nun soll in diesem Kapitel insbesondere dem Problem nachgegangen werden, von dieser auf eine mögliche Aufmerksamkeitszuwendung zu schließen. Als Ausgabe soll diejenige Person oder dasjenige Objekt erkannt werden, auf das eine beobachtete Person ihren Blick richtet.

4.1. Probabilistisches Schließen der visuellen Aufmerksamkeit in statischen Szenarien

Obwohl mit der Ausrichtung der Pupillen die Blickrichtung bestimmbar ist, haben Langton et al. nachweisen können, dass erst mit einem Berücksichtigen der Kopfdrehung die Zuwendung der Aufmerksamkeit zuverlässig erkannt werden kann [LWB00]. Dafür bezogen sie sich auf Studien, in denen die Stimulation von Nervenzellen im temporalen Kortex des menschlichen Gehirns betrachtet wurden: Darin wurde die Reaktion der Zellen bei Betrachten ausgewählter Fotos von Augenpaaren gemessen und mit Reaktionen verglichen, wenn stattdessen fotografische Kombinationen aus Blickrichtung, Kopf- und Torsoorientierungen vorgelegt wurden. In dieser physiologischen Begründung sehen Langton und seine Kollegen die Kopfdrehung als Bezugssystem zur Ausrichtung der Pupillen sowie die Torsoorientierung als Bezugssystem des Kopfs. Für ein menschliches Bestimmen der Aufmerksamkeit wurde so ein hierarchischer Prozess verstanden, der zunächst anhand der Stellung der Augen im Blickfeld, bei Verdeckung dieser anhand des Blickfelds und wiederum stattdessen anhand der Körperausrichtung kategorisiert wurde. Die Autoren stellten darüber hinaus in Experimenten fest, dass die Kopfdrehung mit der Blickrichtung stärker verbunden ist als in bisherigen Studien angenommen. Ihrer Auffassung nach muss eine systematische Zuwendungsbestimmung so anhand beider Merkmale geschehen. Dass unter bestimmten Bedingungen die Kopfdrehung als Annäherung die eigentliche Blickrichtung sogar gänzlich ersetzen kann, konnten dabei Stiefelhagen et al. empirisch bekräftigen [SFYW98, SFYW99].

4.1.1. Die Kopfdrehung als Annäherung der Blickrichtung

Im Bezug zur Kopforientierung bietet die Ausrichtung der Pupillen einen eindeutigen Richtungsvektor, entlang dessen sich der Aufmerksamkeitsfokus nachvollziehen lässt. Im Kontrast hierzu kann anhand der Kopfdrehung allein nur eine Aussage über die Ausrichtung des Blickfelds geschehen, in dessen Volumen sich mögliche Aufmerksamkeitsziele befinden müssen. Als binokulares Blickfeld wird derjenige Bereich bezeichnet, den ein Mensch bei ruhig gestelltem

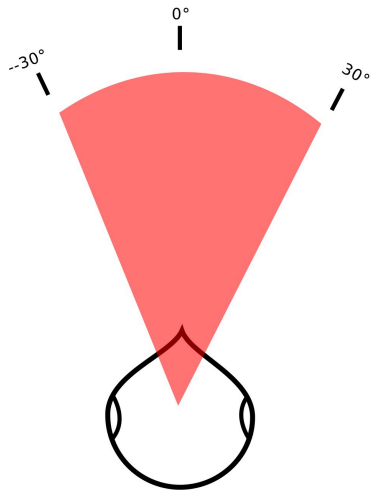


Abb. 4.1.: Darstellung des horizontalen binokularen Blickfelds des Menschen.

Kopf und Körper mit Bewegung beider Augen fixieren kann [Sch73]. Nachfolgend wird entsprechend den in der Literatur gängigen Werten der menschlichen Physiologie und Ergonomie ein ein Fixierfeld von $[-30°, 30°]$ horizontal, beziehungsweise $[-50°, 50°]$ vertikal vorausgesetzt. Mit seiner Größe entspricht das Blickfeld nur einem sehr groben Indikator der eigentlichen Blickrichtung. Das direkte Schließen auf ein Aktionsgegenüber erscheint damit zunächst nur unter solchen Umständen möglich, in denen sich nur ein vorhandenes Ziel im Blickfeld der beobachteten Person aufhält.
Dass jedoch systematisch auch Situationen berücksichtigt werden können, in denen sich mehrere Personen darin aufhalten, wiesen Stiefelhagen et al. empirisch in [SFYW98, SFYW99] nach. In einem Besprechungsszenario untersuchte Stiefelhagen hierzu die beobachtbaren Kopfdrehungen, wenn wiederholt einer der übrigen Besprechungsteilnehmer angeschaut wurde. Dabei zeigte sich in der Praxis, dass Personen ähnliche Kopfdrehungen einsetzen, wenn sie wiederholt dasselbe statische Ziel betrachten. In seinen Beobachtungen konnte Stiefelhagen dieses Muster durch Normalverteilungen modellieren, um die jeweilige Zuwendung einer Person probabilistisch auf eines der damit modellierten Ziele abbilden zu können.

Parametrisierung kognitiver Verhaltensmuster bei Aufmerksamkeitszuwendungen

Untersuchungen an Primaten unterstützen die Annahme, dass bei sich wiederholenden Aufmerksamkeitszuwendungen äquivalente Kopfdrehungen beobachtbar sind [FS08]. Zumindest für statische und vorgegebene Anordnungen der entsprechenden Aufmerksamkeitsziele besteht seither ein indirekter Beweis, dass Blickrichtung und Kopfdrehung demselben kognitiven Prozess unterliegen und die Kopfdrehung damit mehr als nur ein reines Bezugssystem für die Augenbewegung darstellt.

Indem man diesen Sachverhalt ausnutzt, lassen sich zwei Ansätze finden, die von der Kopfdrehung auf Aufmerksamkeitsziele schließen lassen: (1) Lernt man die individuellen Zuwendungsmuster ein, lassen sich für gegebene Aufmerksamkeitsziele Vorhersagen treffen, welche Kopfdrehungen hierbei erwartet werden können. (2) Die zu erwartenden Kopfdrehungen können probabilistisch beschrieben werden, um stochastisch auf das wahrscheinlichste Interaktionsziel schließen zu können.

In Stiefelhagens Untersuchungen wiesen die Kopfdrehungen dabei eine normalverteilte Streuung um denjenigen Referenzwinkel $\tilde{\theta}$ auf, der im Mittel der faktischen, zu erwartenden Kopfdrehung entsprach, wenn ein bestimmtes Ziel wiederholt betrachtet wurde [SYW01a]. Mit einer gegebenen Kovarianzmatrix $\mathbf{C}_{\tilde{\theta}}$ konnte für die zu beobachtenden Kopforientierungen damit der folgende Sachverhalt festgehalten werden:

$$\theta \sim \mathcal{N}(\tilde{\theta}, \mathbf{C}_{\tilde{\theta}}) \tag{4.1}$$

In Abbildung 4.2 wird dieser Zusammenhang schematisch dargestellt.
Damit konnte die Kopfdrehung nicht nur als Ausrichtung des Blickfelds interpretiert, sondern von deren Winkelbeträgen probabilistisch auf das wahrscheinlichste Aufmerksamkeitsziel im Blickfeld geschlossen werden. In Studien konnte hierzu unter anderem von Freedman et al.

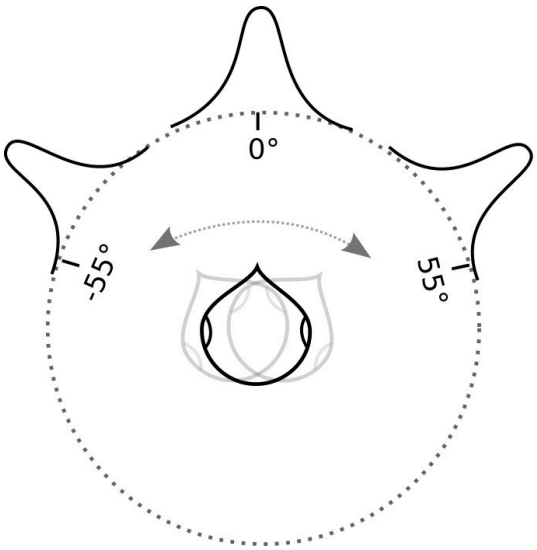


Abb. 4.2.: Beispielhafte Modellierung der zu erwartenden Kopfdrehungen bei drei Aufmerksamkeitszielen. Die Ziele sind um die Person herum, bei $-55°$, $0°$ und $55°$, angeordnet. Bei Zuwendung der Aufmerksamkeit zu einem der Ziele wird der Kopf anteilsmäßig mit gedreht. Links: Bei sich wiederholenden Zuwendungen, erscheinen die Kopfdrehungen normalverteilt um die eigentliche Blickrichtung herum. Rechts: Die Verteilungen der beobachtbaren Kopfdrehungen werden von individuellen Verhaltensmustern beeinflusst. Demnach entsprechend die Mittelwerte nicht den eigentlichen Blickrichtungen sondern hierzu verschoben. Die restliche Sichtstrecke wird innerhalb des Blickfelds mit Augenbewegungen ausgeglichen.

bekräftigt werden, dass die Referenzwinkel $\tilde{\theta}$ dabei linear abhängig zu den eigentlichen Blickwinkeln $\hat{\theta}$ erscheinen, mit denen Objekte relativ zur Oberkörperorientierung einer Person positioniert sind [FS08]. Je weiter ein Ziel damit in die periphere Sicht einer Person wandert, desto stärker wird der Kopf aus der Ruhelage gedreht, um dieses mit den Augen schließlich erfassen zu können. Die Tatsache dass dabei insbesondere $\tilde{\theta} \neq \hat{\theta}$ gilt, ließ sich damit begründen, dass das Ziel faktisch von den Augen erfasst wird aber ein weitwinkliges Drehen in den Randbereich des Blickfelds unangenehm empfunden wird. Infolgedessen wird der Kopf hinreichend weit gedreht, um das Ziel in einen zentraleren Bereich des Blickfelds zu rücken. Für die Referenzwinkel kann diese Abhängigkeit damit wie folgt beschrieben werden:

$$\tilde{\theta} = \kappa \cdot \hat{\theta} \tag{4.2}$$

$\kappa \in [0,1]$ gibt hierbei einen Abbildungsfaktor an, der individuell im angegebenen Definitionsbereich liegen kann, den Untersuchungen nach aber pro Person als konstant angenommen wird.

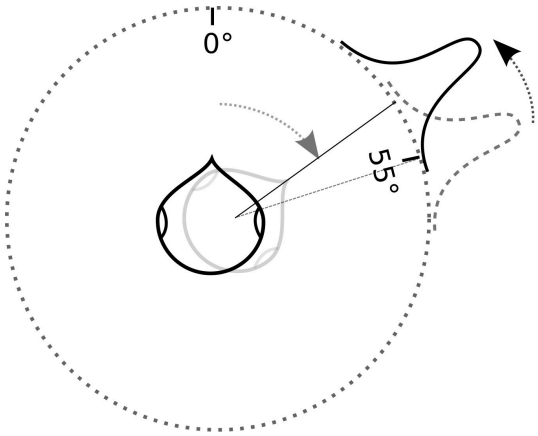


Abb. 4.3.: Kognitive Mittelwertverteilung der Kopfdrehung nach Freedman et al. Um sich einem Ziel bei 55° zuzuwenden, wird der Kopf um $\kappa \cdot 55°$ gedreht, die restliche Strecke wird mit den Augen überbrückt.

Naive Klassifikation der Aufmerksamkeitszuwendung

Bis dato wurde in der Literatur maßgeblich von statischen Anordnungen der Aufmerksamkeitsziele ausgegangen. Sowohl Stiefelhagen als auch die von Freedman et al. veröffentlichten Untersuchungen beließen die Positionen aller Aufmerksamkeitsziele in ihren Untersuchungen unverändert. Wurde eine Person oder ein Objekt wiederholt angeschaut, änderte sich der Referenzwinkel $\tilde{\theta}$ so dazu nicht. Durch die von Stiefelhagen hervorgehobene Normalverteilung der beobachteten Kopfdrehungen lassen sich Zuwendungen zu einem Ziel damit durch eine

naive Bayes'sche Klassifikation erkennen. Mit der Menge $\mathscr{F} = \{F^j\}, j \in \mathbb{N}$ der vorhandenen Aufmerksamkeitsziele bedeutet das insbesondere

$$F^j = \arg\max_{F^j} \{P(F^j|\theta)\} \tag{4.3}$$

mit $P(F^j|\theta) \propto p(\theta|F^j) \cdot P(F^j)$.
Die diskreten A-Priori-Dichten $P(F^j)$ ermöglichen darin im speziellen einen Bias zugunsten einzelner Aufmerksamkeitsziele. Vereinfachend werden diese Wahrscheinlichkeiten jedoch häufig als uniform angenommen, um keines der Ziele in seiner Auswahl zu bevorzugen [BO08a]. Ein erster Versuch diese jedoch explizit zu berücksichtigen findet sich in [SYW01a], worin Stiefelhagen unüberwacht Mischverteilungen einlernt, deren Komponentengewichtungen er als A-Priori-Bezüge versteht. Damit stützt er sich direkt auf die in den Daten vorhandene Zuwendungssystematik der jeweiligen Personen für die die Modelle berechnet werden und übernimmt deren jeweilige Bevorzugung einzelner Ziele. In verwandten Arbeiten zeichnet sich hierzu die Entwicklung ab, dass diese inzwischen anhand des beobachtbaren Kontextgeschehens und jeweiligen Interaktionen der Personen untereinander generativ berechnet werden [BHO09, BO11, HJB+08].

4.2. Schwerpunkte dynamischer Szenenanordnungen

Die im vorherigen Abschnitt umrissene Klassifikation erlaubt probabilistisch auf das am wahrscheinlichsten betrachtete Fokusziel zu schließen, wenn man die Kopfdrehung der Person beobachtet. Hierzu müssen die Blickwinkel zu den jeweiligen Aufmerksamkeitszielen im Blickfeld der beobachteten Person bekannt sein. Für hinreichend diskriminative Wahrscheinlichkeiten wird dabei von der Annahme ausgegangen, dass die Ziele (1) relativ zu jener Person ausreichend disjunkt platziert sind und (2) sich vor allem während der zu evaluierenden Situation nicht bewegen. Dies hängt damit zusammen, dass die eingelernten Referenzwinkel des Beobachtungsmodells in direkter Abhängigkeit zu der Positionierung der Objekte im Blickfeld stehen.
In Folge sind dynamische Szenen, in denen weder die Anzahl der Aufmerksamkeitsziele noch deren (sich änderbare) Position initial bekannt sind, bisher unberücksichtigt geblieben. Tatsächlich aber stellen gerade diese Situationen den größeren Praxisbezug dar, weil sie unvermeidbar im Kontext einer aufmerksamen Umgebung auftreten.
Dieser Abschnitt setzt sich deshalb gezielt mit den Herausforderungen auseinander, die in dynamischen Szenen vorkommen. Weil in der Literatur eine solche Problemstellung bislang unberücksichtigt blieb, sollen dabei im Einzelnen jene Kennzeichen herausgestellt werden, die während der hiesigen Untersuchungen maßgeblichen Einfluss auf die Aufmerksamkeitszuwendung vorwiesen und dem weiteren Systementwurf zugrunde gelegt wurden.

4.2.1. Berücksichtigung unerwarteter Aufmerksamkeitsziele

Bisher in der Forschung berücksichtigte Szenarien setzten vorgegebene Zielpositionen voraus. Alle Personen und Objekte verharren so an fest definierten Orten. Unter dieser Annahme konnten Kopfdrehungen in einen Bezug zum Aufenthaltsort der jeweiligen Aufmerksamkeitsziele gestellt werden. In Arbeiten, in denen probabilistische Modelle benutzt wurden, konnte dieser Bezug dabei durch Trainingsdaten eingelernt werden. Mit dem Einlernen auf dedizierte Zielanordnungen wird das System aber unflexibel gegenüber neuen Situationen. Insbesondere die variable Anzahl vorhandener Aufmerksamkeitsziele stellt eingelernte Modellen vor eine Herausforderung. Um mit neuen Situationen umgehen zu können, müssten so Modelle für alle Ziel- und Positionspermutationen vorliegen, die im Anschluss jeweils angewandt werden können. Im Hinblick auf die Vielfalt möglicher Permutationen scheint das Benutzen statischer Annahmen so nicht empfehlenswert.

4.2.2. Anpassen der Modelle

Der Mittelwert der in Gleichung 4.1 beschriebenen Normalverteilung entspricht der faktischen Kopfdrehung die zu beobachten ist, wenn die Aufmerksamkeit dem entsprechenden Ziel zugewandt wird. Aus kognitiven Gründen entspricht dieser Winkel dabei nicht der tatsächlichen Blickrichtung, sondern fällt in der Regel geringer aus, weil die Zuwendung als Kombination der Kopf- und Augenbewegung aufzufassen ist. Für statische Zielanordnungen konnte nachgewiesen werden, dass dieser Mittelwert implizit aus Trainingsdaten eingelernt oder über ein kognitives Modell bestimmt werden kann, das den Zusammenhang zwischen Kopfdrehung und Blickrichtung zu erfassen versucht.
Unter Berücksichtigung dass im zugrunde gelegten Anwendungsfall Personen nicht konstant an einer Stelle im Raum verharren, sondern sich frei umher bewegen dürfen, bietet die Möglichkeit Referenzwinkelwerte der Kopfdrehungen vorhersagen zu können eine großen Vorteil. Die kombinatorische Vielfalt möglicher Trajektorien vorhandener Personen und Objekte im Raum erlaubt eine nur sehr spärliche Abdeckung durch Trainingsdaten im Vorfeld. Somit ließe sich ein Einlernen auf sehr gezielte und damit beschränkende Situationen umgehen. Mit der Vorhersage der Referenzwinkel erscheint es hierbei intuitiv die vorhandenen Modelle an die Zieltrajektorien ihrer entsprechenden Personen und Objekte adäquat zu adaptieren.
Die in den jeweiligen Studien gemachten Annahmen gleichen dabei jedoch der bisher zugrunde gelegten Unveränderlichbarkeit der Zielpositionen. Die dabei beschriebene, lineare Abbildung des faktischen Blicks auf einen Referenzwinkel der Kopfdrehung berücksichtigte keineswegs die äußeren Umstände unter denen die Aufmerksamkeitszuwendung geschah. Setzt man die Kopfdrehungen nämlich in einen Bezug zu der eigentlichen Interaktion zwischen Personen, dann fällt auf, dass die Referenzwinkel davon abhängig erscheinen, in welchem Diskurs die Zuwendung erfolgt und zwischen welchen Personen hin- und hergewechselt wird. Insbesonde-

re spielen dabei die Ausgangs- und zielgerichtete Orientierung des Kopfs eine tragende Rolle. Bei einem konstanten Wechsel der Aufmerksamkeit zwischen zwei sich nicht bewegenden Personen, erscheinen konstante Referenzmittelwerte vertretbar. Ein linearer Abbildungsfaktor κ bliebe demnach unverändert und würde zu einer korrekten Vorhersage führen. Die Ursprungsorientierung für den Zuwendungswechsel ist in diesem Fall die Kopfdrehung zu jeweils jener Person, auf der die Aufmerksamkeit bislang lag und von der im Anschluss wegrotiert wird. Die Kopfdrehung fluktuiert zwischen den beiden Zielen und fällt dabei stets ähnlich aus weil immer dieselbe Winkeldifferenz zum nächsten Ziel überbrückt werden muss. Mit weiteren Zielen entsteht aber eine größere Permutationsmenge möglicher Zuwendungssequenzen. In Folge treten verschiedene Ausgangs- und zielgerichtete Orientierungen auf, je nachdem welche der Personen im Rahmen des Diskurs fokussiert wird. Weil Menschen in solchen Situationen Annahmen über den weiteren Verlauf der Interaktion machen, treten Zielreferenzwinkel in Abhängigkeit vom Kontext des Geschehens auf. Folglich beeinflusst das die Mittelwertverteilung der Modelle, die unter diesem Gesichtspunkt ebenfalls abhängig von der Interaktionsdynamik und Anordnung der Ziele erscheinen. In Rahmen der Experimente stellte sich die in Gleichung 4.2 aufgeführte Abbildung damit als nicht anwendbar heraus. Der Abbildungsfaktor κ fluktuierte, je nachdem in welchem Kontext die Aufmerksamkeitszuwendungen geschahen. Der in Abbildung 4.4 dargestellte Verlauf unterstreicht diese Erkenntnis dabei auf einer ausgesuchten Teilsequenz der zugrunde gelegten Evaluationsdaten. Prinzipiell stellt sich damit die Frage, wie kognitive Modelle mit normalverteilten Beobachtungen übereinstimmen? Die durch letzterem einbezogene Varianz der Kopfdrehungen wird im Zusammenhang sowohl Rauschen unterliegen als auch Folge eigentlicher Diskurseinflüsse sein. Der Mittelwert kann damit als Referenz hinsichtlich unbeeinflusster Zuwendungsmuster in jeweils immer gleichen Ausgangslagen verstanden werden. Mit einbezogener Dynamik erscheint diese faktische Abbildung jedoch inkonstant und nicht übertragbar. Weil äquivalente Situationen mit exakt denselben Umstände in einer dynamischen und bewegten Umgebung nur sehr spärlich erfasst werden können, muss davon ausgegangen werden, dass die Anwendbarkeit bisheriger kognitiver Vorhersagen hierfür fragwürdig ist. Für eine probabilistische Beschreibung sind aber Referenzwinkelwerte notwendig, die neben entsprechender Varianzen ein diskriminatives Schließen auf Aufmerksamkeitsziele ermöglichen. Die Initialisierung und notwendige Adaption des Beobachtungsmodells stellt somit eine der wesentlichen Herausforderung dar, um mit dynamischen Szenen umgehen zu können.

4.2.3. Interaktionsdynamik und Salienz als Merkmal bevorzugter Zuwendungen

Mit einer festen Anordnung der Ziele werden restriktive Annahmen über die zu erwartenden Handlungen gemacht. Die A-Priori-Wahrscheinlichkeit beziffert dabei eine faktische Zuwendungsbevorzugung einzelner Aufmerksamkeitsziele. Mit unveränderlichen Modellen wird da-

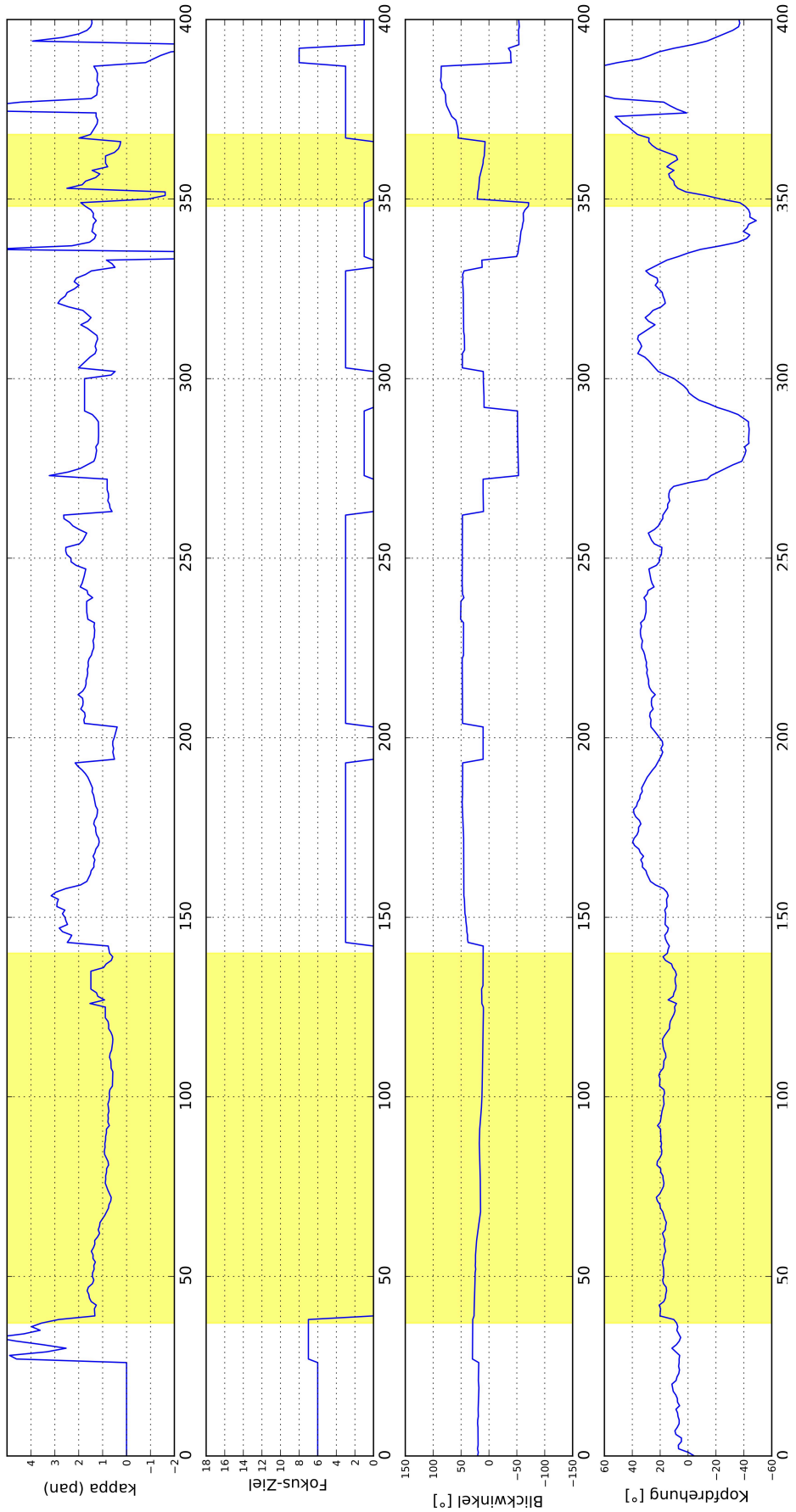


Abb. 4.4.: Beispielsequenz aus einer der Besprechungen. Dargestellt ist die Veränderung des Abbildungsfaktors κ bei horizontalen Kopfdrehungen. Aus Gründen der Vollständigkeit ist der Wechsel der Aufmerksamkeit (zweite Reihe), der eigentliche Blickwinkel zum Zentroiden des entsprechenden Zieles (dritte Reihe), die beobachtete Kopfdrehung (vierte Reihe) sowie die Blickwinkeldifferenz zwischen der gemachten Kopfdrehung und dem Blickwinkel (fünfte Reihe) aufgetragen.

bei gezielt auf eine eingelernte Situation Bezug genommen. In der Forschung wird deswegen zunehmend versucht diesem Aspekt durch ein Beobachten entscheidender Situationsmerkmale gerecht zu werden, so dass die A-Priori-Attraktivität eines Aufmerksamkeitsziels der Rahmenhandlung und dem Kontext heuristisch angepasst werden kann [BHO09, BO11, HJB+08]. Mit dem expliziten Einbeziehen primärer Bevorzugungen konnte bereits eine deutliche Steigerung der Korrektklassifikationsrate nachgewiesen werden.
Die Situationserfassung impliziert allerdings nicht erschöpfend alle Gründe für die Motivation hinter einem Aufmerksamkeitswechsel. Visuelle Reize, insbesondere im peripheren Sehfeld des Menschen, bewegen diesen dazu seinen Blick zu einem Ziel auszurichten [IK01, Sch73]. Diese umfassen Merkmale wie zum Beispiel Farb- und Kontrastunterschiede, Gradienten, Richtung und Beschleunigung beobachteter Bewegungen sowie Stereodisparitäten. Weil diese Faktoren Einflüsse von außen darstellen und die Aufmerksamkeit einer Person wegen ihrer jeweiligen Salienz auf sich ziehen, können situationsabhängige Bezüge als mentale Zustandsannahmen aufgefasst werden, die die Aufmerksamkeitszuwendung von innen heraus steuern. In Kombination kann die Suche nach einem gezielten Objekt im Blickfeld entsprechend beschleunigt werden [NI06, SRF10]. Eine gesamtheitliche Erfassung dieser Faktoren würde so eine generative Beschreibung der einhergehenden Zuwendungsmotivation darstellen [NI05].
Hat man keinen Einblick in die Bewegung der Pupillen, dann entspricht das Nachvollziehen der Aufmerksamkeitszuwendung einer faktischen Schätzung auf welche salienten Merkmale eine Person ihren Blick im Blickfeld lenkt. Die Erscheinung von Salienzen ist dabei jedoch abhängig von der jeweiligen Perspektive der Person. Aus einem anderen Blickwinkel erscheinen Kontraste und Gradienten unterschiedlich. Um solche Merkmale erfassen zu können, wäre infolgedessen ein exaktes Rekonstruieren der Szene aus Sicht der jeweiligen Person notwendig. Mit der Anbringung der Kameras in aufmerksamen Umgebungen kann dies nicht ausreichend geschehen. Zum einen führen hierbei Verdeckungen dazu, dass die ihnen entsprechenden Bereiche nicht hinreichend erfasst werden können. Zum anderen dürfen auch Licht und Schattenwurf in der Szene nicht ignoriert werden, weil insbesondere durch sie Salienz beeinflusst wird. Damit bleibt die Herausforderung perspektivisch invariante Reize in einer Szene zu beobachten und diese zu erfassen.

4.2.4. Sichtbarkeit im perspektivischen Blickfeld

Mit bewegten Zielen sind Situationen unvermeidbar, in denen sich Trajektorien zweier Personen kreuzen oder Objekte hintereinander platziert werden. Aus Sicht einer Person erscheinen diese Objekte damit (teil-)verdeckt. Ein Aufmerksamkeitsziel erfährt dadurch einen dramatischen Einfluss auf seine Form und Erscheinung. Saliente Merkmale, die für den Blick bislang ausschlaggebend waren um ein Objekt wahrnehmen zu können, können somit plötzlich verschwinden.

Die bisher in der Forschung benutzte Repräsentation von Aufmerksamkeitszielen bezieht sich ausschließlich auf den Blickwinkel, wenn ein Ziel betrachtet und fokussiert wird. Insbesondere wird hierbei vom Mittelpunkt des Objekts im Blickfeld der zuwendenden Person ausgegangen und dessen Lage in Form einer einfachen Winkelangabe beziffert.
Eine solche Darstellung ist nicht in der Lage die Verdeckung eines Objekts zu erkennen. Während bei einer vollständigen Verdeckung das Objekt noch gegebenenfalls unberücksichtigt bleiben kann, wirft eine teilweise Sichtbarkeit aber darüber hinaus die Fragestellung auf, wie bisherige Beobachtungsmodelle auf die sich verändernde Form eines Ziels eingehen sollen? Hierfür ist das Einbeziehen der Objektmaße notwendig, was durch die einfache Beschreibung normalverteilter Kopfdrehwinkel in Bezug zum Objektmittelpunkt nicht möglich ist. Damit kann auch keine Aussage gemacht werden, welche salienten Merkmale zur Objektwahrnehmung sichtbar bleiben - ein Aspekt der mit bisherigen Modellen generell unberücksichtigt blieb, weil der Zielmittelpunkt im Blickfeld keine Erfassung der Textur oder Struktur eines Objekts erlaubt.

Mit der Frage nach der Sichtbarkeit eines Ziels stellt sich darüber hinaus auch das Problem der Sichtbarkeitsprüfung. Eine adäquate Repräsentation und Modellierung muss die Möglichkeit bieten auf sichtbar bleibende Segmente, beziehungsweise saliente Wahrnehmungsmerkmale eines Objekts oder einer Person, Bezug nehmen zu können. Bei einer angenommen Berücksichtigung der Objektmaße muss damit auch die Möglichkeit einhergehen, verdeckte Teilbereiche vom übrigen Volumen segmentieren zu können - ein Schritt, der in der Regel nur unter aufwendigen geometrischen Berechnungen möglich ist und ferner die vollständige, dreidimensionale Erfassung und Modellierung aller Objekte und Personen voraussetzt. Doch nur unter diesen Gesichtspunkten kann für ein Objekt entschieden werden, ob es Bestandteil der Aufmerksamkeit einer Person sein kann.

4.3. Systementwurf zur Bestimmung der visuellen Aufmerksamkeit in dynamischen Umgebungen

In diesem Abschnitt wird ein System vorgestellt, das gezielt auf dynamische Szenen ausgerichtet ist. Im ersten Teil wird hierfür auf eine modifizierte Repräsentation der Aufmerksamkeitsziele eingegangen. Diese werden in atomare Einheiten zerlegt, auf die die in Abschnitt 4.1.1 beschriebenen und etablierten Modelle jeweils übertragen werden. Zielobjekte werden so mit ihren Maßen dreidimensional beschrieben, was die Segmentierung verdeckter Bereiche deutlich vereinfacht.
Durch die Unterteilung in atomare Elemente werden saliente Merkmale in der Erscheinung der Aufmerksamkeitsziele implizit berücksichtigt. Daneben werden dedizierte, äußere Reize einbezogen, um ansichtsinvariente Beobachtungen möglicher A-Priori-Zuwendungen zu erhalten. Hierfür wird das von Itti et al. vorgeschlagene Konzept der *Bayes'schen Überraschung* genutzt,

um für beobachtetes Verhalten in einer Szene eine Aussage zu treffen, in wie weit es unerwartet und überraschend für einen Betrachter wirkt.
Das System wird auf einem dedizierten Datensatz evaluiert, der eigens für diese Arbeit aufgezeichnet und annotiert wurde. Weil die Problemstellung dynamischer Zielanordnungen bislang unberücksichtigt blieb, soll mit dem Datensatz eine erste Evaluationsreferenz für nachfolgende und weitere Ansätze geschaffen werden.

4.3.1. Voxelbasierte Repräsentation der Interaktionsziele

Unter dem Begriff *Voxelisierung* versteht man die Segmentierung eines dreidimensionalen Körpers in äquidistant große Elemente gleicher Kantenlänge. Der Körper wird hierbei diskretisiert und so durch eine Menge abzählbarer Elemente, sogenannter *Voxel*, repräsentiert. Jedes dieser Voxel ist dabei eindeutig nummeriert und wird so adressierbar.
Sei im Folgenden mit $c \in \mathbb{R}$ eine beliebige, aber feste Kantenlänge aller Voxel vorgegeben. Die Diskretisierung eines Körpervolumens führt zu der ihm entsprechenden Voxelmenge

$$\mathscr{V} = \{\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3 | \exists k_1, k_2, k_3 < k : x_1 = k_1 c \wedge x_2 = k_2 c \wedge x_3 = k_3 c\} \tag{4.4}$$

Objekte oder sogar Räume können so in eine ihnen entsprechende Voxelmenge überführt werden. Personen und Objekte F^j die hier im Kontext als mögliche Ziele der Aufmerksamkeitszuwendung betrachtet werden, sollen so im Folgenden in Form von Untermengen $F^j := \mathscr{V}^j \subseteq \mathscr{V}$ beschrieben werden. Zur verdeutlichten Objektivierung der Voxel und Abgrenzung einfacher Vektoren, soll im Folgenden statt des Vektors $\mathbf{x}$ ein Voxel durch V bezeichnet werden. Die dem Ziel F^j entsprechende Voxelmenge $\mathscr{V}^j$ wird damit im Folgenden durch ihre Elemente $\mathscr{V}^j = \{V^{j,l}\}, l = [1, 2, ..., N_{\mathscr{V}_j}]$ bezeichnet. Damit führt die Beschreibung eines Ziels im Rahmen dieser Arbeit zu einer quaderförmigen Approximation der Personen und Objekte (zu erkennen in den Darstellungen der Systemvisualisierung 4.6 und 4.7). Bei geringer Kantenlänge c wäre eine detaillierte Berücksichtigung der Objektsilhouette möglich, deren Erfassung soll jedoch nicht Bestandteil dieser Arbeit sein. Im Folgenden wird von der Annahme ausgegangen, dass die annähernde Darstellung für den weiteren Verlauf ausreichend ist und eine hinreichende Unterscheidung der Objekte und Personen erlaubt. Bezüglich einer möglichen Silhouettenerfassung, sei der interessierte Leser an dieser Stelle deswegen auf weiterführende Arbeiten, wie zum Beispiel [SvdCIS09] hingewiesen.

Raycasting zur Sichtbarkeitsprüfung

Mit der atomaren Zusammensetzung aller Aufmerksamkeitsziele kann das Problem der Objektsichtbarkeit darauf reduziert werden, welche der dem Objekt zugehörigen Elemente verdeckt werden. Aus Sicht der beobachteten und zu unterstützenden Person spannt sich deren Blickfeld

damit als ein Sichtkegel in die diskretisierte Voxelmenge des Raumvolumens auf. Eine Prüfung, welche der darin erfassten Voxel der Person sichtbar erscheinen, kann infolgedessen einfach dadurch erreicht werden, dass das perspektivische Blickfeld von Sichtstrahlen abgetastet und jeweils auf Schnitte mit Voxel untersucht wird. Dieses, in der Computergrafik mit *Raycasting* bezeichnete Verfahren, stellt dabei eine intuitive Vorgehensweise dar, die von einer Person betrachtete Szene nachvollziehen und darstellen zu können und bietet so eine inzwischen schnelle Möglichkeit sichtbare Bereiche bestimmen und festlegen zu können [WW92]. Abbildung 4.5 verdeutlicht das Vorgehen dabei noch einmal im Detail.

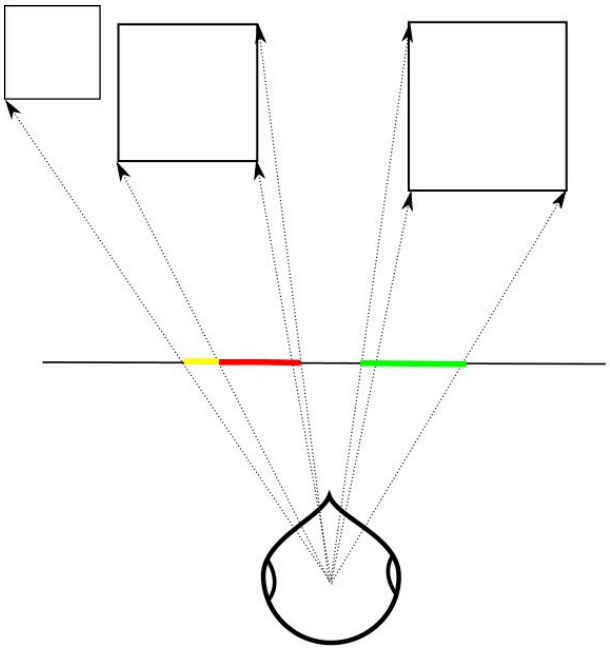


Abb. 4.5.: Schematische Darstellung des Raycastings. Die Szene wird auf eine Bildebene, vor dem Betrachter projiziert. Hierzu wird ausgehend von der Blickposition eine Sichtgerade durch jedes Pixel der Bildebene geschickt. Schneidet diese einen Körper in der Szene, so werden dessen Eigenschaften auf das Pixel übertragen (hier verschiedenfarbig pro Objekt dargestellt). Verdeckungen werden so implizit erkannt und berücksichtigt.

Im Einzelnen wird das Blickfeld hierzu auf eine Bildebene vor dem Betrachter projiziert - ähnlich dem Prinzip einer Lochkamera, in der die Sicht durch die Öffnung des Objektives auf die Bildebene gelangt. Die diskrete Darstellung der Bildebene in Form adressierbarer Pixel gibt zweidimensionale Elemente vor, durch deren Mittelpunkt eine Sichtlinie gelegt werden kann, deren Ursprung in der Position des Betrachters liegt. Die Sichtlinie erstreckt sich in das Volumen des aufgespannten Blickfelds. Schneidet die Gerade dabei ein Objekt, wird eine Weiterverfolgung der Sichtlinie abgebrochen und die Eigenschaften des Schnittpunkts auf das der Sichtlinie jeweilige Pixel der Bildebene übertragen - im Fall hier entspräche das dem Schnitt mit einem Voxel eines Aufmerksamkeitsziels. Weil weitere Voxel entlang der jeweiligen Sichtgerade unberücksichtigt bleiben, stellen die auf die Bildebene projizierten Voxelschnitte im

Anschluss eine Momentaufnahme dar, welche Objektelemente schließlich sichtbar erscheinen und in die weitere Berechnung einbezogen werden müssen. Mit hinreichend kleiner Voxelgröße und Bildebenendiskretisierung kann dabei davon ausgegangen werden, dass die Erfassung der sichtbaren Segmente ausreichend detailliert geschehen kann.

4.3.2. Schließen der visuellen Aufmerksamkeitszuwendung

Mit der atomaren Beschreibung der Aufmerksamkeitsziele bleiben die in der Forschung etablierten Beobachtungsmodelle prinzipiell übertragbar. Die eigentliche Aufmerksamkeitszuwendung bezieht sich infolge nicht mehr auf den faktischen Blickwinkel zum Körpermittelpunkt eines Objekts, sondern auf die Mittelpunkte der sichtbaren Voxel, die diesem zugeordnet wurden. Damit kann der in in Gleichung 4.3 beschriebene, probabilistische Ansatz auf die hier zugrunde gelegte Repräsentation der Aufmerksamkeitsziele übertragen werden. Um dabei darauf einzugehen, dass Zielobjekte nicht mehr fest positioniert sein müssen, soll das Beobachtungsmodell nicht die Kopfdrehung im Weltbezug θ berücksichtigen, sondern die zum Voxelmittelpunkt relative Drehung $\theta_{Vj,l}$. Die Betrachtung eines Ziels wird so als eine Rotation zu den dem Ziel zugehörigen Voxel aufgefasst. Mit den etablierten Normalverteilungen als Beobachtungsmodell, gilt für die zu erwartetenden Kopfdrehungen zu einem Ziel, beziehungsweise einem Zielvoxel, demnach $\theta_{Vj,l} \sim \mathcal{N}(\tilde{\theta}_{Vj,l}, \mathbf{C}_{Vj,l})$. Als Kovarianzmatrix soll im Folgenden vereinfachend eine Diagonalmatrix $\mathbf{C}_{Vj,l} = (\sigma^{j,l})^2\mathbf{I}$ eingesetzt werden.
Damit kann sichergestellt werden, dass Drehwinkel unabhängig von der Position des Betrachters und der Trajektorie des jeweiligen Ziels einbezogen werden:

$$P(F^j|\theta_{Vj,l}) \propto \frac{\sum_{l=1}^{|\mathcal{V}^j|} p(\tilde{\theta}_{Vj,l}|V^{j,l})}{|\mathcal{V}^j|} P(F^j) \tag{4.5}$$

Der bisherige, nur dem gesamten Ziel F^j entsprechende Bezug geschieht nun über die Summe aller dem Objekt zugehörigen Elemente $V^{j,l}$. Mit der Untermenge $\hat{\mathcal{V}}^j$ kann sich dabei auf lediglich diejenigen Voxel beschränkt werden, die aus Sicht der zu bewertenden Person sichtbar bleiben. Die eigentliche Zuwendungsentscheidung aus Gleichung 4.5 lässt sich damit zu einer Summenbetrachtung über die sichtbaren Segmente des (teil-)verdeckten Ziels umschreiben:

$$P(F^j|\theta) \propto \frac{\sum_{l=1}^{|\mathcal{V}^j|} p(\tilde{\theta}_{Vj,l}|\hat{V}^{j,l})}{|\mathcal{V}^j|} P(F^j) \tag{4.6}$$

4.3.3. Berücksichtigung exogener Aufmerksamkeitslenkungen

Wie Posner in [Pos80] beschreibt, können Ursachen für Aufmerksamkeitszuwendungen *endogener* oder *exogener* Natur sein. Während sich endogene Lenkungen auf intentionale Zustands-

wechsel im Bewusstseinsmodell einer Person beziehen, beschreiben exogene Lenkungen Reize von außerhalb, die eine unwillkürliche Zuwendung der Aufmerksamkeit auf sich triggern.

Im Kontext der Blickanalyse und -vorhersage auf Videodaten stellten Itti et al. einen Ansatz vor, probabilistische Zustandsannahmen für Beobachtungen als Ausdruck unerwarteten oder erwarteten Verhaltens zu interpretieren [IB05]. Merkmalsausprägungen, die von gemachten Annahmen abweichen, werden so durch die von den Autoren definierte *Bayes'sche Überraschung* bewertet und überraschende Verhaltensmuster damit quantisierbar.
Die Zustandsyhypothese gibt hierbei eine Aussage darüber, welche Beobachtungen im aktuellen Moment erwartet werden. Unerwartete Ausprägungen führen so zu einer messbaren Abweichung der adaptierten Modellbeschreibung von der zuvor jeweiligen Zustandsannahme. Nach den Verfassern sind diese dabei ein generelles Konzept, das entsprechend auf raumzeitliche Merkmale, Sensormodalitäten und beliebigen Datentypen und -quellen definiert und angewandt werden kann. Überraschende Beobachtungen treten in diesem Zusammenhang jedoch nur dann auf, wenn im Modell implizite Unsicherheiten zugrunde gelegt werden, was insbesondere bei stochastischen Beschreibungen, fehlenden Informationen oder begrenzten Ressourcen der Fall ist.

Nach [Jay03, Sav72, IB09], stellt Bayes Theorem die einzig konsistente Methode dar, um ein Modellieren und Schließen mit Unsicherheiten bewerkstelligen zu können. A-Priori-Dichten entsprechen darin Hypothesenannahmen im zugrunde liegenden Zustandsraum. Mit neuen Beobachtungen führen diese nach Bayes zu A-Posteriori-Wahrscheinlichkeiten, die bei veränderten Ausprägungen darauf schließen lassen, dass die vorigen Annahmefunktionen über mögliche Systemzustände angepasst werden mussten und damit nicht adäquat ausfielen. Mit gegebenen Modellen $M \in \mathscr{M}$ und einer gegebenen Beobachtung $\mathbf{d}$, lassen sich zuvor einhergehende Prämissen über Zustandsausprägungen damit wie folgt in entsprechende A-Posteriori-Wahrscheinlichkeiten $\{P(M|\mathbf{d})\}_{M\in\mathscr{M}}$ umformen:

$$\forall M \in \mathscr{M} : P(M|\mathbf{d}) = \frac{p(\mathbf{d}|M)}{p(\mathbf{d})} P(M) \tag{4.7}$$

Bei unveränderter Form der damit erhaltenen Dichtefunktionen bestanden mit den gemachten Beobachtungen keine überraschenden Erkenntnisse und jeweilige Merkmalsausprägungen traten wie erwartet ein. Unerwartetes Verhalten würde hingegen zu einer messbaren Differenz der Dichtefunktionen führen, wofür Itti et al. die Kullback-Leibler-Divergenz (im Folgenden KL genannt) einsetzen, um mit diesem Maß eine Bewertung der vorhandenen Divergenzen in den Modellen zu erhalten:

$$S(\mathbf{d},\mathscr{M}) := KL(P(M|\mathbf{d}),P(M)) = \int_{\mathscr{M}} P(M|\mathbf{d}) \log \frac{P(M|\mathbf{d})}{P(M)} \mathrm{d}M \tag{4.8}$$

Die von der KL-Divergenz erhaltenen Werte entsprechen jedoch keiner stochastischen Aussage. Hierzu müssen die Divergenzen als solche interpretiert werden. Im Rahmen dieser Arbeit wird deshalb die Exponentialverteilung als Wahrscheinlichkeitsmaß der KL-Divergenz vorgeschlagen, so dass für eine Modellannahme $\mathscr{M}^j$ eines Aufmerksamkeitsziels F^j und eine dem Ziel entsprechende Beobachtung $\mathbf{d}^j$ im wesentlichen der Sachverhalt $P(\mathscr{M}^j|\mathbf{d}^j) \propto 1 - \mathrm{e}^{-\lambda^j S(\mathbf{d}^j, \mathscr{M}^j)}$ angenommen werden kann. Damit der A-Priori-Bezug zum Ziel hierzu in Abhängigkeit von der restlichen Menge vorhandener Aufmerksamkeitsziele gesetzt wird, wird zusätzlich über diese Überraschungsbezüge der übrigen Ziele normalisiert. Eine endgültige A-Priori-Wahrscheinlichkeit ergibt sich damit aus:

$$P(\mathscr{M}^j|\mathbf{d}^j) \propto \frac{1 - \mathrm{e}^{-\lambda^j S(\mathbf{d}, \mathscr{M}^j)}}{\sum_k 1 - \mathrm{e}^{-\lambda^k S(\mathbf{d}^k, \mathscr{M}^k)}} \tag{4.9}$$

Die Zuwendungsattraktivität zu einem Ziel kann so als ein momentaner Ausdruck des aktuellen Geschehens aufgefasst werden und stellt hierzu ferner alle Ziele untereinander in einen Bezug. Die A-Priori-Wahrscheinlichkeit $P(F^j)$ eines Ziels F^j betrachtet zu werden korrelliert also mit all seinen vorliegenden Kontextbeobachtungen $\mathscr{D}^j = \{\mathbf{d}^{j,i}\}, i \in \mathbb{N}$. Es gilt also:

$$P(F^j) := P(F^j|\mathscr{D}^j) \tag{4.10}$$

womit die Berechnung der Zuwendungswahrscheinlichkeiten nach Gleichung 4.6 wie folgt erweitert werden kann:

$$P(F^j|\theta) \propto \frac{\sum_{l=1}^{|\mathscr{V}^j|} p(\theta|\hat{V}^{j,l})}{|\mathscr{V}^j|} P(F^j|\mathscr{D}^j) \tag{4.11}$$

Statistische Ausreißer statt Überraschungsbezüge

Im Gegensatz zu statistischen Ausreißern soll mit dem Überraschungswert erfasst werden, ob Beobachtungen genügend Entropie für Aussagen über ihre jeweiligen Zustandsraumausprägungen beinhalten. Die Autoren verdeutlichen diesen Sachverhalt am Beispiel zweier Hypothesen M und der ihr gegensätzlichen $\bar{M}$, für die jeweils eine A-Priori-Dichte die zu erwartende Beobachtung im nächsten Messschritt beschreibt. Eine neue Beobachtung $\mathbf{d}$, die von beiden Dichten niedrig bewertet wird, würde so für beide Modelle einen statistischen Ausreißer darstellen. Weil $\mathbf{d}$ aber weder für M noch für $\bar{M}$ eine Aussage darüber zulässt, welche der beiden Hypothesen sie eher zuzuordnen ist, trägt die Beobachtung keinerlei verwendbare Information. Aufgrund dessen liefert sie auch kein Ausmaß an Überraschung, weil nicht ausreichend diskriminativ zwischen den beiden Hypothesenklassen M und $\bar{M}$ unterschieden werden kann.
Um für eine gegebene Beobachtung einen Überraschungsbezug gewinnen zu können, muss die A-Priori-Erwartung eine gegensätzliche Annahme zum gemachten Messwert darstellen. Im erläuterten Beispiel käme das einem Bias für Hypothese M gleich, während die Beobachtung eher

der Annahme der Gegenhypothese entsprechen müsste. Nur so würde die darauf folgende A-Posteriori-Beschreibung, dass im nächsten Sequenzschritt eher die Gegenhypothese zu erwarten sein wird, damit unerwartet ausfallen und schließlich zu einem entsprechenden Überraschungswert $S(\mathbf{d}, \{M, \bar{M}\})$ nach Gleichung 4.8 führen.

Bewegung als ansichtsinvariante Salienz

Neben Farbe und Gradienten stellt Bewegung in einer Szene ebenfalls ein salientes Merkmal dar, das die Aufmerksamkeit einer Person und die damit verbundene Blickrichtung auf sich ziehen kann [IK01]. Bezogen auf die Anbringung der Kameras in der hiesigen Problemstellung stellt Bewegung damit ein ansichtsinvariantes Merkmal dar, das nicht von der Perspektive einer bestimmten Person abhängig ist. Die Bewegung von Personen lässt sich dabei mit Hilfe von Personentrackingverfahren nachvollziehen. Stattdessen kann auch das in Kapitel 3 beschriebene Kopftrackingverfahren benutzt werden, in dem die Kopfpositionen der Personen als Positionshypothesen im Raum eingesetzt werden. Als Folge lässt sich die Bewegungsgeschwindigkeit jeder Person im Raum errechnen und modellieren. Bewegungsgeschwindigkeiten, die dabei von gemachten Annahmen abweichen, können so als äußere Reize interpretiert werden, die die Aufmerksamkeiten anwesender Kollegen auf sich ziehen.

Sei im Folgenden mit b_t^j die Bewegungsgeschwindigkeit einer Person F^j zum Zeitpunkt t bezeichnet. Eine A-Priori-Annahme über das Bewegungsverhalten sei ferner durch eine Normalverteilung gegeben, so dass die erwartete Ausprägung der Beobachtung stochastisch modelliert wird und eine Dichtefunktion über den damit verbundenen Zustandsraum liefert. Der Mittelwert der A-Priori-Annahme sei dabei durch $hat b_t^j$, die Varianz durch $\hat{\sigma}_t^j)^2$ beschrieben. Für die erwartete Messung b_t^j gilt demnach der Zusammenhang $b_t^j \sim \mathcal{N}(\hat{b}_t^j, (\hat{\sigma}_t^j)^2)$.
Mit Vorliegen einer neuen Beobachtung wird die A-Priori-Dichte in die jeweilige A-Posteriori-Wahrscheinlichkeitsfunktion überführt und zur Messung der Divergenz nach Gleichung 4.8 genutzt. Aus Gründen der Berechenbarkeit wird die Beobachtung als Normalverteilung repräsentiert: Mit einer normalverteilten Beobachtung tritt der Sonderfall ein, dass diese *konjugiert* zur Funktionsfamilie der A-Priori-Dichtefunktion ist. Konjugierte A-Priori- und Beobachtungsfunktionen resultieren in A-Posteriori-Dichten derselben Funktionsfamilie, welche dadurch in geschlossener Form darstell- und insbesondere berechenbar werden [GCSR03, RS61, BI05]. Im Allgemeinen lässt sich das auf ein einfaches Anpassen der entsprechenden Hyperparameter der Funktionsfamilie reduzieren - für eine Normalverteilung betrifft das Mittelwert und Varianz.

Für Beobachtungen $b_t^j \approx \mathcal{N}(b_t^j, \sigma_b^2 \to 0)$ ergibt sich die A-Posteriori-Funktion damit aus der Aktualisierung des A-Priori-Mittelwerts und der zugehörigen Varianz [GCSR03, RS61]:

$$\begin{aligned} \hat{b}_t^j &= \left(\frac{\hat{b}_{t-1}^j}{(\hat{\sigma}_{t-1}^j)^2} + \frac{b_t^j}{(\sigma_b)^2}\right)\left(\frac{1}{(\hat{\sigma}_{t-1}^j)^2} + \frac{1}{\sigma_b^2}\right)^{-1} \\ (\hat{\sigma}_t^j)^2 &= \left(\frac{1}{(\hat{\sigma}_{t-1}^j)^2} + \frac{1}{\sigma_b^2}\right)^{-1} \end{aligned} \tag{4.12}$$

Die daraus gewonnene A-Posteriori-Modellierung der Zustandsannahme kann schließlich durch Anwenden der Kullback-Leibler-Divergenz einen Bezug auf den Überraschungswert der gemachten Beobachtung liefern:

$$KL(\mathcal{N}(\hat{b}_{t-1}^j, (\hat{\sigma}_{t-1}^j)^2), \mathcal{N}(\hat{b}_t^j, (\hat{\sigma}_t^j)^2)) = \frac{((\hat{b}_{t-1}^j - \hat{b}_t^j)^2 + (\hat{\sigma}_{t-1}^j)^2 - (\hat{\sigma}_t^j)^2}{(2(\hat{\sigma}_t^j)^2)} + \ln\frac{(\hat{\sigma}_t^j)^2}{(\hat{\sigma}_{t-1}^j)^2} \tag{4.13}$$

Sprechermodellierung

Neben Bewegungen stellen Interaktionsmerkmale einen herausragenden Aspekt der Situationsbeobachtung dar [HGP08, HF08, HJB+08, HGPHF08]. Im Einzelnen bietet hierbei die Beteiligung an Diskussionen und Gesprächen einen wesentlichen Faktor, der individuellen Beitrag und Einsatz in interagierenden Gruppen bezeichnet. Hinsichtlich Personen die sich bisher wenig am Diskurs beteiligen kann die Annahme gemacht werden, dass sie sich auch in Folge nur selten äußern werden. Die Häufigkeit ihrer Sprachäußerungen stellt somit eine messbare Beobachtung dar, die als die Anteilnahme der Person in einem Gespräch interpretiert werden kann und die im Folgenden ein weiterer externer Attraktor für Aufmerksamkeitszuwendungen darstellen soll[1].

Bei der Modellierung der Sprachhäufigkeit handelt es sich um das stochastische Modellieren der Auftrittswahrscheinlichkeit eines Ereignisses in einem vordefinierten Zeitfenster. In folgenden soll daher ein Poisson-Prozess zugrunde gelegt werden, dessen Parameter k_t^j die Häufigkeit angibt, mit der eine Person F^j im vorgegebenen Zeitintervall Sprachäußerungen von sich gibt. Hierbei gilt insbesondere

$$f(k_t^j; \hat{\lambda}_t^j) = \frac{(\hat{\lambda}_t^j)^{k_t^j} \mathrm{e}^{-\hat{\lambda}_t^j}}{k_t^j!} \tag{4.14}$$

wobei $\hat{\lambda}_t^j$ die aktuelle Zustands- und A-Priori-Annahme der erwarteten Sprachhäufigkeit bezeichnet.

[1] Die Detektion ob eine bestimmte Person spricht oder nicht, wird in der Literatur gängig mit Hilfe von Nahbesprechungsmikrofonen umgesetzt, auf deren aufgezeichnetem Signal die Energie einem Schwellwert unterzogen wird [BO11]. Weil solche Mikrofone während der Datenaufnahme in dieser Arbeit nicht verfügbar waren, wurden Sprecherhäufigkeiten auf den Aufzeichnungen eines Tischmikrofons manuell annotiert und ersatzweise eingesetzt.

Durch die hierzu konjugierte Gamma-Verteilung lässt sich ein geeignetes Modell finden, um die Annahme bezüglich $\hat{\lambda}_t^j$ a-priori zu modellieren [RS61, GCSR03, BI05]. So kann mit der Gamma-Funktion $\Gamma(\cdot)$, den Hyperparametern α und dem umgekehrten Skalierungsparameter β die Zustandsannahme über die modellierte Sprachhäufigkeit wie folgt beschrieben werden:

$$\gamma(\hat{\lambda}_t^j; \hat{\alpha}_t^j, \hat{\beta}_t^j) = \frac{(\hat{\beta}_t^j)^{\hat{\alpha}_t^j}}{\Gamma(\hat{\alpha}_t^j)} (\hat{\lambda}_t^j)^{\hat{\alpha}_{t-1}^j} \mathrm{e}^{-\hat{\beta}_t^j \hat{\lambda}_t^j}, \hat{\lambda}_t^j > 0 \tag{4.15}$$

Bei neuer Beobachtung λ_t zum Zeitpunkt t führt wiederum durch den Sonderfall konjugierter Beobachtungen, ein einfaches Aktualisieren der Hyperparameter zur Berechnung der A-Posteriori-Dichte. Im Einzelnen gilt hierbei:

$$\hat{\alpha}_t^j = \hat{\alpha}_{t-1}^j + k_t^j \tag{4.16}$$

$$\hat{\beta}_t^j = \hat{\beta}_{t-1}^j + 1 \tag{4.17}$$

Ein Überraschungsbezug erfolgt schließlich wie auch in Gleichung 4.13 durch Anwenden der Kullback-Leibler-Divergenz auf den jeweiligen Annahmemodellen. Mit der Digamma-Funktion $\Psi(\cdot)$ ergibt sich somit:

$$KL(\gamma(\hat{\lambda}_{t-1}^j; \hat{\alpha}_{t-1}^j, \hat{\beta}_{t-1}^j), \gamma(\hat{\lambda}_t^j; \hat{\alpha}_t^j, \hat{\beta}_t^j)) = \tag{4.18}$$

$$\hat{\alpha}_t^j \log(\frac{\hat{\beta}_{t-1}^j}{\hat{\beta}_t^j}) + \log(\frac{\Gamma(\hat{\alpha}_t^j)}{\Gamma(\hat{\alpha}_{t-1}^j)}) + \hat{\beta}_t^j \frac{\hat{\alpha}_{t-1}^j}{\hat{\beta}_{t-1}^j} + (\hat{\alpha}_{t-1}^j - \hat{\alpha}_t^j)\Psi(\hat{\alpha}_{t-1}^j) \tag{4.19}$$

4.3.4. Adaption des Beobachtungsmodells zur Laufzeit

Mit der Voxelisierung der Zielobjekte, wird deren Darstellung im Blickfeld einer Person diskretisiert. Systematisch wird die Erscheinung eines Aufmerksamkeitsziels für einen Beobachter, wie in Abschnitt 4.3.1 beschrieben, in Form einer Sichtebenenprojektion modelliert. Die sichtbaren Voxel werden dabei reflektiert und die Szene aus der Perspektive der beobachtenden Person annähernd rekonstruiert. Jedes Voxelmodell kann daneben als implizite Repräsentation möglicher Salienzen innerhalb der Objekterscheinung im Blickfeld verstanden werden. Kontrastunterschiede, Farben oder Gradienten in der projizierten Objekttextur müssen so nicht explizit nachvollzogen werden, sondern werden durch Voxel an entsprechenden Stelle generalisiert.

Wie in Abschnitt 4.1.1 und 4.2 diskutiert wurde, können persönliche Verhaltensmuster und der Kontext einer Szene als maßgebliche Faktoren angesehen werden, die die Orientierung der Kopfdrehung bei Aufmerksamkeitszuwendungen beeinflussen. Das äußert sich insbesondere in der Mittelwertverteilung des zugrunde gelegten Beobachtungsmodells, denn durch dieses wird

eine Referenzwinkelannahme gemacht, die die faktische Kopfdrehung bei der Zuwendung zu einem Ziel beschreibt.

Durch die Ausgangslage den visuellen Fokus der Personen in einer aufmerksamen Umgebung nachvollziehen zu können, wird implizit die Forderung gestellt, adaptiv auf sich ändernde Situationen reagieren zu können. Ein einmaliges Einlernen der Modelle, wie es zum Beispiel bei Stiefelhagen et al. [SFYW98, SFYW99, SYW01a] der Fall ist, ist somit nicht ausreichend, wenn sie auf verschiedene Handlungen übertragbar sein sollen. Um auf die sich verändernden Zielanordnungen dynamischer Szenen einzugehen, findet sich deshalb ein grundlegender Beitrag dieser Arbeit deshalb darin, Bewegung und Sprachhäufigkeit der Fokusziele als zwei ansichtsinvariante Merkmalsbeobachtungen nachzubilden, die zur Erfassung des Kontextgeschehens dienen sollen. Ihre Einflüsse sollen zu einer Adaption des Beobachtungsmodells benutzt werden. Dazu wird das Modell zur Laufzeit entsprechend der A-Posteriori Wahrscheinlichkeiten der zugehörigen Ziele angepasst. Je stärker die A-Priori-Attraktivität eines Zielobjekts betrachtet zu werden steigt, desto stärker sollen die jeweiligen Normalverteilungen der Voxelmodelle auf die momentan zu beobachtende Kopfdrehung angepasst werden. Durch den Einfluss externer Reize sollen die Modelle so auf sich ändernde Interaktionsdynamiken reagieren und die angenommenen Referenzwinkel darüber hinaus einer sich ändernden Zielanordnung nachführen.
Dafür wurde eine Laufzeitvariante des *Expectation Maximization* Algorithmus implementiert, der die Verschiebung und Skalierung der normalverteilten Beobachtungsdichten in jedem Sequenzschritt aktualisiert. Die Implementierung stützt sich dabei auf Arbeiten von Stauffer et al. [SG00], in denen Mischverteilungen mit Hilfe einer laufzeitfähigen Adaption des Algorithmus an neue Beobachtungen angepasst wurden. Die Autoren beziehen sich in ihren Arbeiten dabei auf Mischverteilungen als Farbmodelle pro Pixel in einem Kamerabild: Für jedes Pixel wird eine Mischverteilung, bestend aus einer vorgegebenen Anzahl an Komponenten, initialisiert und träge auf neue Farbwerte des Pixels hin angepasst. Aktualisierte Kamerabilder können daraufhin pixelweise mit den bestehenden jeweiligen Modellen verglichen und stark veränderte Areale im Bild automatisch segmentiert werden. Durch eine träge Adaptierung der Modelle erreichten die Autoren dabei, dass die Farbmodelle die ursprünglichen Farbwerte ihrer jeweiligen Pixel beibehielten und erst bei einer längerfristigen Veränderung der Farben entsprechend angepasst wurden. In der Bildverarbeitung wird dieses Vorgehen seither populär dazu eingesetzt ein statisches Bild als Hintergrundmodell einzulernen und Bewegung im Vordergrund, durch die spontane Farbveränderung in den jeweiligen Pixeln, segmentieren zu können. Die Adaptierung erfolgt dabei dadurch, dass für ein Pixel die aktuelle Farbe mit den Komponenten der Mischverteilung im Farbraum verglichen wird und jene Komponente angepasst wird, deren Funktionswert für die neue Beobachtung einen vorgegebenen Schwellwert übersteigt. Entsprechend eines ebenfalls vorgegebenen Lernfaktors wird infolgedessen Mittelwert und Varianz der jeweiligen Gausskomponente an die neue Beobachtung angepasst. Auf den hiesigen Anwen-

dungsfall übertragen, erfahren die Voxelmodelle eine Adaption, indem Mittelwert $\tilde{\theta}_{V^{j,l},t-1}$ und Varianz $\sigma^2_{V^{j,l},t-1}$ des vorigen Zeitschritts an neue Kopfdrehungsbeobachtungen θ_t adaptiert werden. Nach [SG00] lässt sich dies im Einzelnen wie folgt umsetzen:

$$\begin{aligned}\tilde{\theta}_{V^{j,l},t} &= (1-\rho_t^j)\tilde{\theta}_{V^{j,l},t-1} + \rho_t^j \theta_{V^{j,l},t} \\ \sigma^2_{V^{j,l},t} &= (1-\rho_t^j)\sigma^2_{V^{j,l},t-1} + \rho_t^j (\theta_{V^{j,l},t} - \tilde{\theta}_{V^{j,l},t})^T (\theta_{V^{j,l},t} - \tilde{\theta}_{V^{j,l},t})\end{aligned} \tag{4.20}$$

wobei für ρ_t^j gilt:

$$\rho_t^j = aP(F^j|\mathscr{D}^{j,t}) \tag{4.21}$$

Die Stärke der Adaption wird damit durch den angesprochenen Lernfaktor a gesteuert, der zusammen mit der Ausprägung des Voxelmodells für die vorliegende Beobachtung das endgültige Ausmaß der Anpassung vorgibt. In den nachfolgenden Evaluationen wurde dieser empirisch auf 0,2 festgelegt.

4.4. Evaluationen

In diesem Abschnitt soll das entworfene System auf seine Korrektklassifikationsrate hin untersucht und evaluiert werden. Wegen der bisher neuartigen Schwerpunktsetzung dieser Arbeit mit dynamischen Szenen umzugehen, war hierzu kein dedizierter Datensatz verfügbar. Infolgedessen wurde ein eigener aufgenommen und annotiert. Dieser Datensatz soll hierzu im folgenden Abschnitt zunächst erläutert und vorgestellt werden.

4.4.1. Datensatz

Arbeiten, die sich bisher damit beschäftigt haben menschliche Aufmerksamkeitszuwendungen von Kopfdrehungen abzuleiten, legten den Schwerpunkt allesamt weniger auf die Dynamik in der Szene, als auf dedizierte Situationen, in denen der systematische Mehrwert einer umfassenderen Erfassung damit hervorgehoben werden kann (für eine detaillierte Diskussion sei der Leser an Abschnitt 2.2 erinnert). Dabei beschränken sie sich auf statische Szenen oder wenigstens fest vorgegebene Zielanordnungen. Damit sind bis dato keine Referenzdatensätze verfügbar, die vollständig annotiert sind und dynamische Zielanordnungen beinhalten.
Im Rahmen dieser Arbeit wurde deshalb ein dedizierter Datensatz aufgenommen und mit Annotationen versehen. Dieser umfasst insgesamt zehn Besprechungen, mit einer jeweils unterschiedlichen Anzahl an Teilnehmern und Sitzplätzen.
Um während der Aufnahmen Aktionen beobachten zu können, die zu Bewegungen der Personen, Verschieben vorhandener Objekte, Unterbrechungen der Gespräche und weiterem, unerwartetem Verhalten führten, folgten die Besprechungen einem vordefinierten Drehbuch. In jeder Aufnahme waren dafür jeweils drei Teilnehmer in die Geschehnisse eingeweiht, die so den

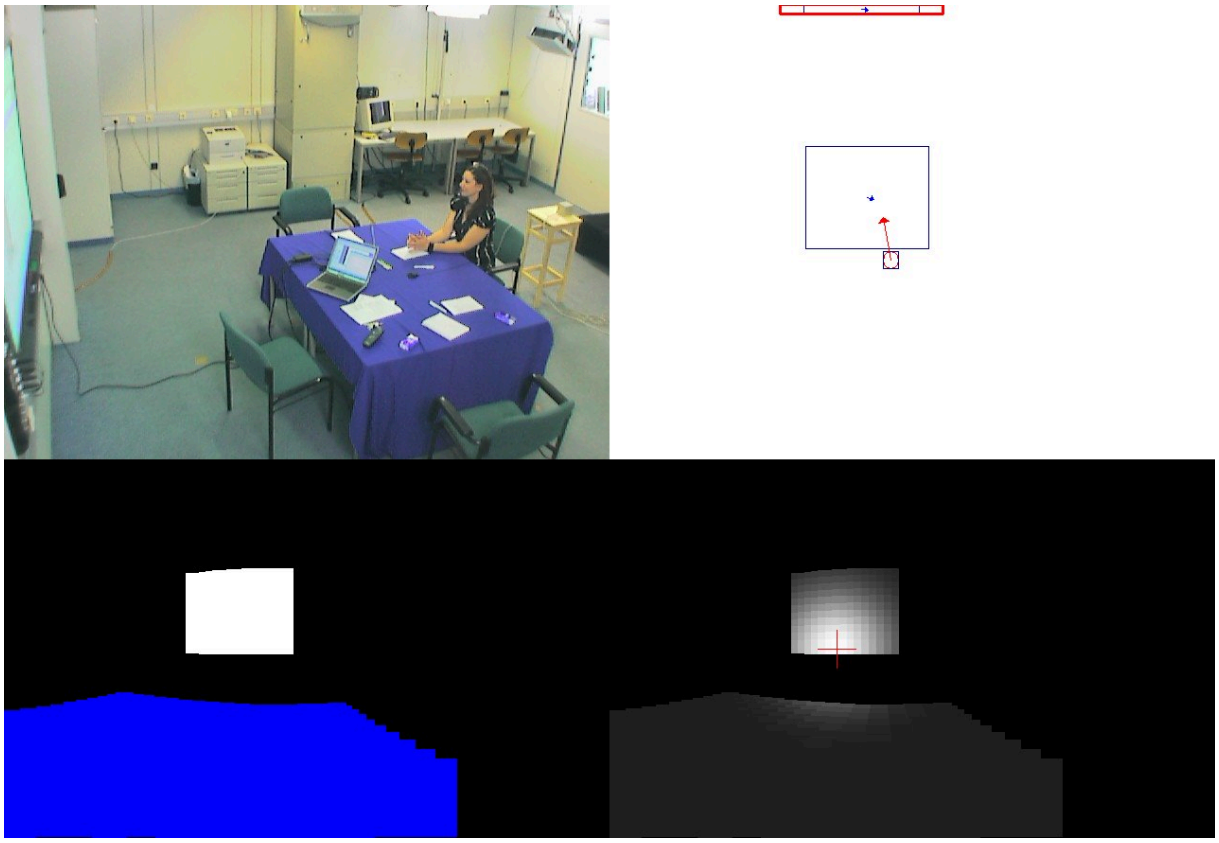


Abb. 4.6.: Visualisierung des Systems zur Aufmerksamkeitsschätzung. Dargestellt ist ein Moment kurz bevor die Besprechung beginnt. Links oben: Die Person mit dem Sensor auf dem Kopf (zur Protokollierung der Kopfrehung) sitzt wegen der Sensorinstallation bereits am Tisch und blickt geradeaus zur Leinwand. Rechts oben: Rekonstruierte Ansicht des Systems aus der Vogelperspektive. Die genutzte Kopfdrehung zum Schätzen des Aufmerksamkeitsziels ist als roter Pfeil hervorgehoben. Das erkannte Ziel ist rot umrandet. Links unten: Rendering der Voxel, aus Sicht der Person deren Kopfdrehung genutzt wird; die möglichen Ziele im Blickfeld sind zur Unterscheidung farblich verschieden kodiert - die Leinwand ist weiß eingefärbt, der Tisch blau. Rechts unten: Rendering der Voxel und ihrer Wahrscheinlichkeiten betrachtet zu werden, aus Sicht der Person deren Kopfdrehung genutzt wird; je heller ein Voxel dargestellt ist, desto wahrscheinlicher ist es angesehen zu werden. Das rote Zielkreuz in der Darstellung symbolisiert die die eigentliche Kopfdrehung.

weiteren Verlauf des Geschehens beeinflussten. Die übrigen Teilnehmer wurden weder über den Ablauf noch über Grund ihres Zusammentreffens informiert. Damit sollte sichergestellt werden, dass diese unbeeinflusst auf die Ereignisse reagieren und eine natürliche Haltung und Teilnahme während der Aufnahme beibehalten konnten.

Die Dauer der Besprechungen wurde auf maximal zehn Minuten beschränkt. Eine detaillierte Übersicht über die jeweilige Aufnahmedauer und Teilnehmeranzahl wird in Tabelle 4.1 aufgelistet. Daneben stellt Tabelle 4.2 einen Überblick dar, welche Personen im Einzelnen an den Besprechungen teilnahmen und welchen Sitzplatz sie während der Aufnahme einnahmen. Daran soll verdeutlicht werden, welche der Teilnehmer in die Ereignisse eingeweiht waren (jeweils

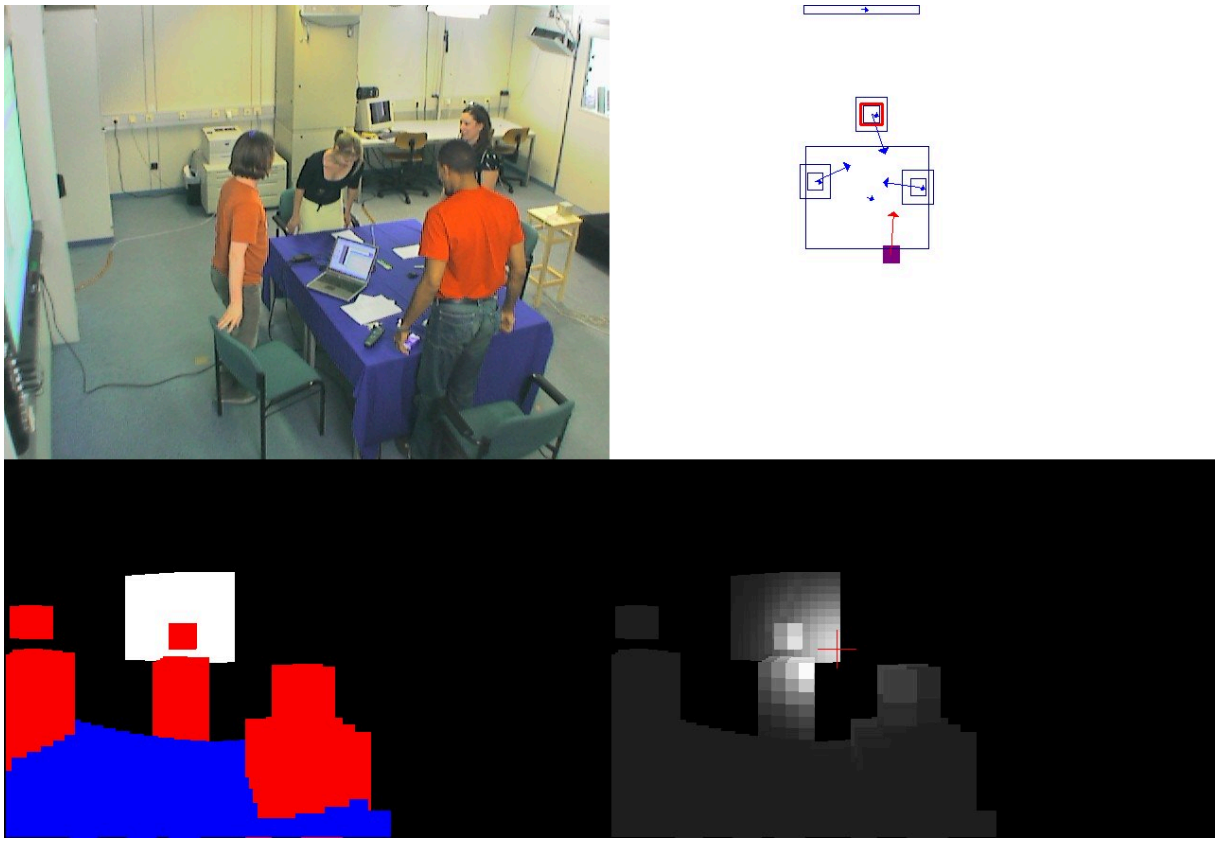


Abb. 4.7.: Weitere beispielhafte Szene und die zugehörige Visualisierung des Systems zur Aufmerksamkeitsschätzung. Hier sind die Besprechungsteilnehmer im Begriff sich nach der gegenseitigen Begrüßung zu setzen. Wie auch in Abbildung 4.6 wird für die Person an der südlichen Tischkante (im linken oberen Bild die rechte weibliche Person) die Aufmerksamkeit nachvollzogen. In der Vogelperspektive rechts oben ist ihre Kopfdrehung durch den roten Pfeil eingezeichnet, während die Kopfdrehungen der übrigen Teilnehmer mit blauen Pfeilen dargestellt sind. Das abgeleitete Fokusziel ist rot umrandet: in diesem Zeitschritt wurde die gegenüberstehende Person als visuelles Aufmerksamkeitsziel erkannt. Dabei hervorzuheben ist, dass die Kopfdrehung deutlich sichtbar zur Leinwand orientiert ist. Durch die Kontextmerkmale Bewegung und Sprachaktivität wird jedoch die Person hervorgehoben und stellt somit das wahrscheinlichere Aufmerksamkeitsziel dar.

gekennzeichnet als Schauspieler durch S_1, S_2 oder S_3) und wie die Sitzplatzverteilung der uneingeweihten Personen ausfiel (jeweils dargestellt durch $P_{1..16}$). Die unterschiedliche Nummerierung der fremden Personen unterstreicht dabei insbesondere, dass keine wiederholt aufgenommen wurde. Um die Anordnung der drei Ereignisauslöser variabel zu halten, wurde deren Sitzplatzverteilung pro Aufnahme permutiert.

Ereignisablauf

Im folgenden soll der Ablauf der vorgegebenen Ereignisse während der Aufnahmen zusammengefasst werden. Bei der Erstellung des Ablaufs wurde darauf geachtet, dass sich die vorge-

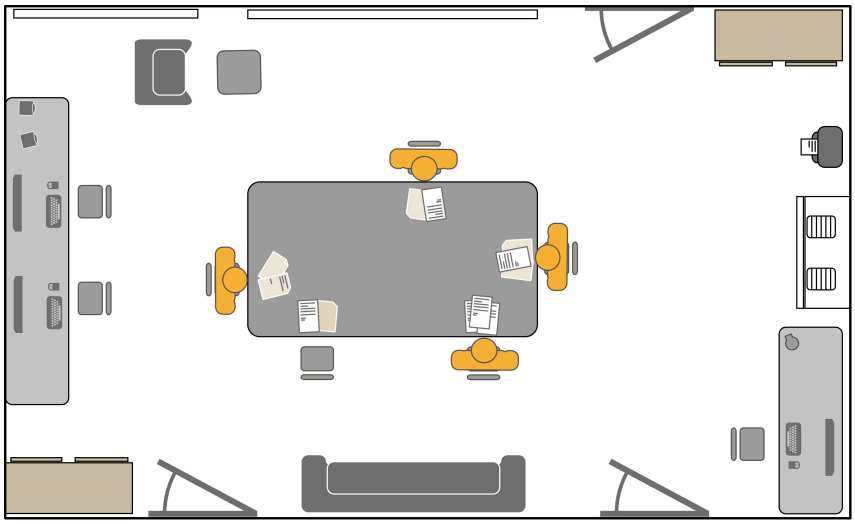


Abb. 4.8.: Schematische Darstellung des Besprechungszimmers und der annotierten Aufmerksamkeitsziele darin. Alle Objekte und das gesamte Mobiliar wurde hierzu vermessen und protokolliert. Die Positionen der Personen ergaben sich aus den Kopfdrehungsschätzungen.

gebenen Geschehnisse sowohl in den Handlungsablauf der jeweiligen Situationen integrierten als auch dedizierte Unterbrechungen des Diskurses und ferner Hintergrundaktivitäten umfassten. Damit sollte ein regulärer Arbeitsalltag nachgestellt, mit unerwarteten Ereignissen aber die Aufmerksamkeit der anwesenden Person gezielt abgelenkt werden. Die folgenden Stichpunkte listen den chronologischen Handlungsrahmen aller Aufnahmen auf.

- Die Teilnehmer betreten den Raum, begrüßen sich stehend am Besprechungstisch und nehmen jeweils auf beliebigen Sitzplätzen um den Tisch herum Platz. Im folgenden ist die Kennzeichnung der Personen so vorgenommen worden, dass die Personen im Uhrzeigersinn um den Tisch herum beziffert werden, beginnend mit *Person 1* an der Nordkante des Tischs. *Person 2* sitzt demnach an der Ostkante, *Person 3* an der Südkante, usw.

- Während alle Diskutanten am Tisch sitzen, betritt eine weitere Person (im Folgenden als *Extra-Person* bezeichnet) den Raum und begrüßt die anwesenden Teilnehmer. Sie sucht einen nahestehenden Stuhl und setzt sich mit an den Besprechungstisch. Weil die bisherigen Teilnehmer nicht damit gerechnet haben, müssen sie ihre Stühle verrücken, um Platz für den zusätzlichen Schauspieler zu machen. Er beteiligt sich kurzzeitig am Gespräch. Anschließend verlässt er den Tisch, lässt den zugestellten Stuhl aber stehen. Er wendet sich einem der umstehenden Arbeitsplätze. Nach kurzer Zeit löst der Schauspieler von seinem Arbeitsplatz aus eine laute und störende Audioausgabe an angeschlossenen Lautsprechern aus. Er unterbricht damit das bisherige Geschehen und lenkt die Aufmerksamkeit auf sich. Nach kurzer Entschuldigung, eilt er zu den Lautsprechern, schaltet diese auf lautlos und verlässt den Raum auf anderem Weg, als er zuvor eingetreten ist.

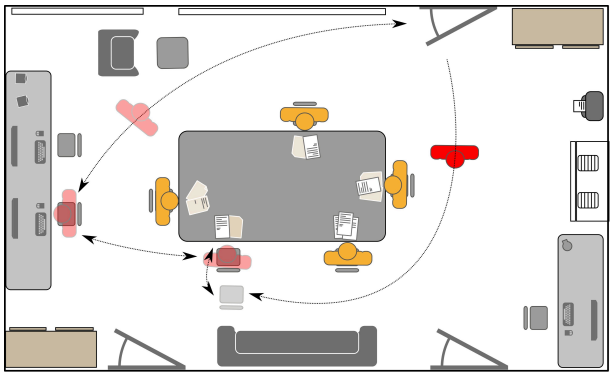


Abb. 4.9.: Beispielhafte Darstellung einer Unterbrechung während einer Besprechung: während alle Diskutanten am Tisch sitzen, betritt eine weitere Person den Raum (rot dargestellt), holt sich einen nahstehenden Stuhl und setzt sich an den Tisch. Nach kurzer Zeit verlässt die Person die Besprechung und widmet sich anderen Aufgaben.

- Einer der anwesenden Schauspieler steht auf, geht zur Projektionsleinwand und hält dort einen Vortrag mit aufgelegten Folien. Das Thema orientiert sich grob an seinem jeweiligen Fachgebiet und gliedert sich flüssig in die vorhergehende Besprechung ein.

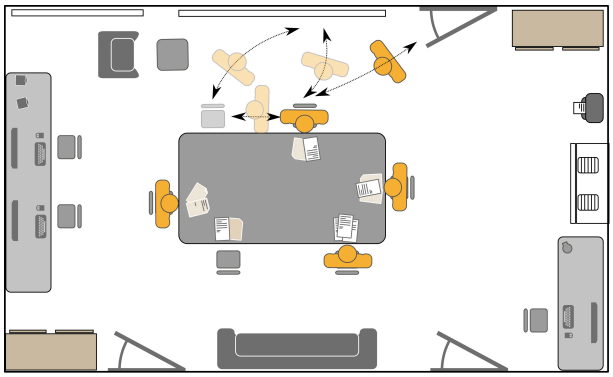


Abb. 4.10.: Beispielhafte Darstellung eines Vortrags während einer Besprechung: ein Teilnehmer steht auf und geht zur Projektionsleinwand. Dort hält er einen Vortrag und setzt sich anschließend wieder an seinen Platz. In manchen Fällen verschiebt er seinen Stuhl darüberhinaus seinen Stuhl und setzt sich an die verschobene Position.

- Während der Besprechung startet ein Schauspieler von außerhalb eine Ausgabe auf den im Raum platzierten Drucker. Er betritt den Raum schweigend und steuert auf den Drucker zu. Dort täuscht er einen Papierstau vor und beginnt lautstark den Drucker zu untersuchen, wobei er mehrfach die Verschlussklappen des Druckers öffnet und wieder schließt um dem vorgetäuschten Papierstau entgegenzuwirken. Er entnimmt seinen Ausdruck und verlässt erneut die Szene.

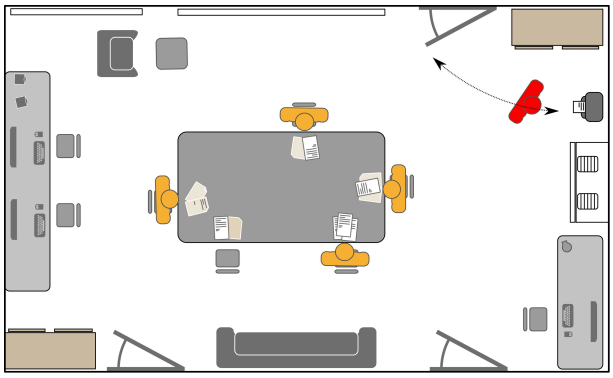


Abb. 4.11.: Beispielhafte Darstellung der Unterbrechung durch einen Papierstau im Drucker: während der Besprechung löst eine Person (rot dargestellt) einen Druckauftrag aus. Sie geht zum Drucker und täuscht dort einen Papierstau vor.

- Die Besprechung wird von einem Telefonläuten unterbrochen. Das Mobiltelefon wurde hierzu vor der Besprechung unbemerkt in einem Schrank platziert und im Anschluss von einem Schauspieler außerhalb des Raums angerufen. Während des Läutens betritt dieser den Raum und spricht die anwesenden Kollegen gezielt darauf, an ob sie das Telefon gesehen haben. Nach kurzem Orten der Lärmquelle geht er zu dem entsprechenden Schrank, öffnet ihn und entnimmt das Telefon. Er entschuldigt sich bei den übrigen Personen und verlässt mit dem Telefon erneut den Raum.

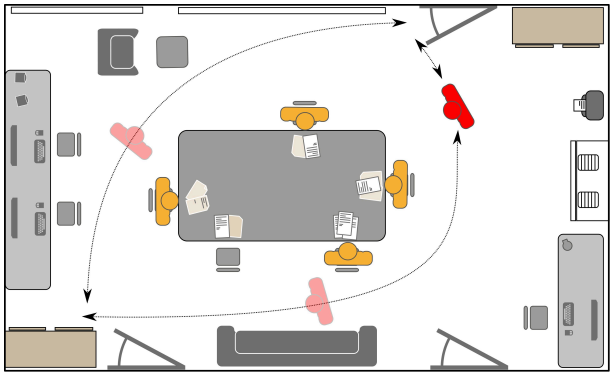


Abb. 4.12.: Beispielhafte Darstellung der Unterbrechung durch ein Telefonläuten während einer Besprechung: während alle Diskutanten am Tisch sitzen, fängt ein Mobiltelefon an zu klingeln, das vor der Besprechung in einem Schrank platziert wurde. Um die Besprechung sicherstellend zu unterbrechen, betritt eine Person den Raum (rot dargestellt) und fragt die Besprechungsteilnehmer ob diese ihr vermisstes Telefon gefunden hätte. Die Person folgt dem Läuten zum Schrank, nimmt das Telefon an sich und verlässt daraufhin wieder den Raum.

- Einer der Schauspieler sucht die Überleitung zu einer entfernt liegenden Stereokamera. Er holt sie, kehrt zurück und zeigt sie allen übrigen Besprechungsteilnehmern. Anschließend legt er die Kamera zum Besichtigen auf den Besprechungstisch.

Besprechungsszenario	Dauer (Minuten)	Frames	Teilnehmer
1	10,2	9147	5
2	8,7	7801	4
3	7,9	7115	5
4	6,3	5650	5
5	6,2	5621	4
6	6,2	5588	5
7	8,0	7232	4
8	7,0	6368	5
9	7,8	7028	4
10	7,6	6870	5

Tab. 4.1.: Dauer und Anzahl Teilnehmer pro aufgezeichnetem Besprechungsszenario.

- Die Besprechung wird beendet. Die Teilnehmer stehen auf, verabschieden sich stehend voneinander indem sie sich die Hände schütteln und verlassen einzeln und nacheinander die Szene.

Neben der in jeder Aufnahme unterschiedlichen Anzahl an Teilnehmern ändert sich diese auch während der Aufnahme selbst durch den wiederholt unterbrechende Schauspieler. Durch dessen spontanes Hinzusetzen an den Besprechungstisch wird infolgedessen erzwungen, dass sich die bisherigen Sitzplatzpositionen ändern und nicht statisch erscheinen. Ferner ändert sich die Anzahl der Aufmerksamkeitsziele um den Besprechungstisch, was die notwendige Adaption des Beobachtungsmodells verstärkt unterstreichen soll. Das Miteinbeziehen diversen Mobiliars und unterschiedlicher Objekte die nicht Bestandteil der eigentlichen Besprechung sind, wird das System mit der Herausforderung konfrontiert, bisher unberücksichtigte Objekte erkennen und miteinbeziehen zu müssen. Durch die verschiedenen Trajektorien der Personen, den Vortrag vor einer Projektionsleinwand und die unterschiedlichen Objektmaße soll darüber hinaus erneut das Problem der Sichtbarkeit hervorgehoben werden und explizit im Datensatz einbezogen werden.

Sensoraufbau während der Aufzeichnung

Die Besprechungen wurden mit vier Kameras des Typs Sony DFW-VL500 aufgezeichnet, die jeweils unterhalb der Raumdecke, in den oberen Ecken des Raums angebracht wurden. Eine weitere Kamera vom Typ Scorpion SCOR-03NSC in der Raumdeckenmitte erfasste mit einem 180°-Fischaugenobjektiv die gesamte Szene von oben. Alle Kameras lieferten jeweils digitalisierte Videoströme, die mit 15 Hz zu 640×480 Pixel aufgezeichnet wurden. Die Videos wurden als JPEG-Bildfolge gespeichert.
Ferner wurden die Besprechungen auditiv mit Mikrofonen aufgezeichnet. Hierzu wurde ein

Besprechungsszenario	Nord	West	Süd	Ost	Unterbrechung	Vortragender
1	S_1	S_2	P_1	P_2	S_3	S_1
2	S_1	S_2	P_3	-	S_3	S_1
3	S_1	S_2	P_4	P_5	S_3	S_1
4	S_1	S_3	P_6	P_7	S_2	S_1
5	S_1	S_3	P_8	-	S_2	S_1
6	S_3	P_9	P_{10}	S_1	S_2	S_1
7	-	S_2	P_{11}	S_1	S_3	S_1
8	P_{12}	S_3	P_{13}	S_1	S_2	S_3
9	-	S_2	P_{14}	S_1	S_3	S_2
10	S_1	P_{15}	P_{16}	S_3	S_2	S_1

Tab. 4.2.: Platzbelegung und Rollenverteilung pro aufgezeichnetem Besprechungsszenario. Schauspieler sind mit *S*, reguläre Teilnehmer mit *P* bezeichnet. Treten Personen in mehreren Besprechungen auf, teilen sie auch dieselbe Kennzeichnung.

omnidirektionales auf der Tischmitte platziert. Vier Mikrofonarrays in T-Form an jeder der vier Raumwänden, wurden darüber hinaus dazu benutzt eine Grundlage für weiterführende Evaluationen zu schaffen, in denen die gemessenen Schalllaufzeiten zu einer automatischen Sprecher- und Tonquellenlokalisierung genutzt werden kann [SBE+06].

Annotationen

Während der Aufnahmen wurde einer der nicht schauspielenden Teilnehmer der in Kapitel 3 angegebene Magnetsensor auf dem Kopf befestigt um dessen Kopforientierung in Echtzeit zu protokollieren. Aus der mangelnden Verfügbarkeit weiterer Sensoren, konnte das pro Aufnahme nur für eine Person genutzt werden. Hierzu wurde immer diejenige Person ausgesucht, die den südlichen Sitzplatz am Besprechungstisch einnahm und frontal zur Projektionsleinwand sehen konnte.
Der Datensatz wurde nach der Aufzeichnung manuell annotiert. Hierzu wurden zunächst die Objektmaße und -positionen aller Objekte im Raum erfasst. Anschließend wurde framebasiert von einem einzelnen Annotator die Kopfposition und -maße aller Teilnehmer in den Kameraansichten festgehalten.
Der Audiostrom des Tischmikrofons wurde dazu eingesetzt, manuelle Protokolle darüber anzufertigen zu welchem Zeitpunkt welche der Teilnehmer gesprochen haben. Wie bereits in Abschnitt 4.3.3 erklärt wurde, geschah dies in Anlehnung daran, dass sich in verwandten Ansätzen der Ansatz etablieren konnte, anhand der Energie eines durch ein Nahbesprechungsmikrofon aufgezeichneten Signals, auf Sprachaktivität zu schließen. Im Rahmen der Aufnahmen war eine solche Sensorik nicht verfügbar, weswegen die Annotationen im weiteren Verlauf als Ersatz dienen sollen.

Drei weitere, unterschiedliche Annotatoren, entschieden anschließend framebasiert für alle Teilnehmer, welche Person oder welches Objekt respektive angeschaut wurde. Hierzu wurden den Annotatoren dieselben Kameraaufnahmen vorgelegt, die im Datensatz aufgezeichnet und in den Evaluationen des Systems zugrunde gelegt wurden. Damit sollen die menschlichen Entscheidungen in den nachfolgenden Evaluationen einen direkten Kontrast zu den Systemhypothesen bieten.

Analyse der Annotationen

Um die menschliche Erkennungsleistung mit den Systemhypothesen vergleichen zu können, wurden die Fokuszielentscheidungen der Annotatoren einander gegenübergestellt und verglichen. Zwei in der Statistik bewährte Gütekriterien für Übereinstimmungen wurden dabei mit dem Gesamtverhältnis der Übereinstimmungen p_0 und Cohens Kappa κ ausgesucht [KP03].
Mit dem Gesamtverhältnis p_0 wird direkt das Verhältnis der Übereinstimmungen zu der gesamten Anzahl vorhandener Annotationen beziffert:

$$p_0 = \frac{\text{Anzahl Übereinstimmungen}}{\text{Anzahl Annotationen}}, \qquad p_0 \in [0,1] \tag{4.22}$$

Cohens Kappa berücksichtigt hingegen die Menge übereinstimmender Annotationen, die als solche unerwartet eintrafen und nur durch Zufall zustande kamen. Hierzu wird die Anzahl der erwarteten zufälligen Übereinstimmungen p_e von der Anzahl tatsächlicher Übereinstimunngen p_0 abgezogen und durch die erwartete Anzahl der Fälle geteilt, in denen Übereinstimmungen nicht durch Zufall geschahen:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \qquad \kappa \in [0,1] \tag{4.23}$$

Stimmen die beiden Annotatoren dabei in jeder ihrer Entscheidungen überein, nehmen die Gütemaße jeweils den Wert 1 an. Je häufiger divergente Entscheidungen vorliegen, desto niedriger sinken sie in ihren Werten. Für eine Aussage über die Qualität der beiden Annotationen, stellt insbesondere $\kappa > 0,4$ ein in der Literatur gängiger, unterer Schwellwert für eine noch hinnehmbare, hinreichend übereinstimmende Anzahl an Übereinstimmungen dar.
Die Ausprägungen der beiden Gütemaße auf den Datensatzannotationen werden in Tabelle 4.3 aufgelistet. Darin werden für alle Aufmerksamkeitsziele die Übereinstimmungen der Annotatoren untereinander bewertet und im Einzelnen aufgeführt. *A1-2* bezieht sich dabei die Übereinstimmungen der Annotatoren A1 und A2. Entsprechend beschreibt *A2-3* die Reliabilität der Annotatoren A2 und A3. Wie man in der Tabelle deutlich sehen kann, stellen die Personen neben der Projektionsleinwand und den jeweiligen Besprechungsunterlagen auf dem Tisch die dominante Menge ausgesuchter Aufmerksamkeitsziele dar: Sie sind es, die von den Annotatoren am häufigsten übereinstimmend selektiert wurden. Die übrigen Objekte im Raum werden

	Fokussiert [%]			κ			p_0		
	A1	A2	A3	A1-2	A1-3	A2-3	A1-2	A1-3	A2-3
Person 1	13,3	14,8	16,3	0,7	0,7	0,6	0,6	0,6	0,5
Person 2	13,5	13,3	13,7	0,7	0,7	0,7	0,6	0,6	0,6
Person 3	11,0	11,0	9,9	0,7	0,7	0,7	0,6	0,6	0,6
Person 4	10,0	10,2	10,7	0,7	0,7	0,7	0,6	0,6	0,6
Unterlagen 1	4,1	4,1	3,4	0,7	0,7	0,7	0,6	0,5	0,5
Unterlagen 2	2,8	2,5	2,2	0,8	0,7	0,7	0,6	0,6	0,6
Unterlagen 3	2,6	2,4	2,3	0,8	0,8	0,8	0,7	0,7	0,7
Unterlagen 4	2,8	2,8	2,3	0,7	0,7	0,7	0,6	0,5	0,5
Person Extra	9,1	9,0	8,0	0,8	0,8	0,7	0,7	0,6	0,5
Stuhl Extra	0,5	0,4	0,6	0,5	0,5	0,4	0,3	0,3	0,2
Unterlagen Extra	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Tisch	0,5	1,2	2,8	0,2	0,1	0,2	0,1	0,1	0,1
Whiteboard	0,0	0,0	0,1	0,1	0,0	0,1	0,1	0,0	0,1
Leinwand	15,5	13,4	11,6	0,8	0,7	0,7	0,7	0,6	0,6
Drucker	1,0	0,9	1,1	0,9	0,8	0,8	0,8	0,7	0,7
Klimaanlage	0,1	0,0	0,1	0,1	0,1	0,1	0,0	0,1	0,0
Schrank Nord	0,2	0,1	0,1	0,2	0,2	0,2	0,1	0,1	0,1
Schrank Süd	0,6	0,9	0,7	0,6	0,7	0,6	0,4	0,5	0,5
Eingang	0,9	1,0	1,1	0,7	0,6	0,5	0,5	0,4	0,4
Tür Labor	0,1	0,1	0,1	0,3	0,2	0,2	0,2	0,1	0,1
Tür Büro	0,1	0,1	0,1	0,1	0,2	0,1	0,1	0,1	0,0
Lautsprecher	0,5	0,5	0,5	0,7	0,6	0,6	0,6	0,5	0,4
Beistelltisch	0,1	0,1	0,1	0,3	0,3	0,3	0,2	0,2	0,2
Desktop 1	0,2	0,0	0,1	0,1	0,3	0,1	0,0	0,2	0,1
Desktop 2	3,9	3,9	3,9	0,9	0,9	0,9	0,9	0,9	0,9
Desktop 3	0,5	0,2	0,4	0,4	0,5	0,5	0,2	0,3	0,3
Kamera	4,2	5,0	5,1	0,6	0,7	0,7	0,5	0,5	0,5
Handy	0,3	0,4	0,5	0,6	0,7	0,5	0,4	0,5	0,4
Helmi	0,2	0,2	0,1	0,2	0,1	0,2	0,1	0,1	0,1
Sofa	0,2	0,1	0,2	0,2	0,2	0,2	0,1	0,1	0,1
Sessel	0,0	0,0	0,0	0,1	0,1	0,1	0,0	0,1	0,1
Lampe 1	0,0	0,0	0,0	0,1	0,1	0,5	0,1	0,1	0,3
Lampe 2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Boden	1,1	1,0	1,7	0,5	0,5	0,5	0,4	0,3	0,3
Hocker	0,1	0,1	0,1	0,3	0,3	0,3	0,2	0,2	0,2
Decke	0,0	0,1	0,1	0,0	0,0	0,6	0,0	0,0	0,4
Stuhl 1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Stuhl 2	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Stuhl 3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Stuhl 4	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Gesamt:	100,0	100,0	100,0	0,7	0,7	0,7	0,8	0,7	0,7

Tab. 4.3.: Übereinstimmung der manuellen Annotationen.

seltener betrachtet. Das kommt daher, dass die meisten während der Aufnahmen nicht Inhalt der Besprechungen waren und damit nur als Requisite im Hintergrund auftraten. Als Fokusziel wurden sie von den Annotatoren nur dann ausgewählt, wenn (1) der Aufmerksamkeitsfokus des annotierten Teilnehmers im Begriff war zu wechseln und die Kopfdrehung sich von einem Ziel zum nächsten wegwandte oder (2) die Sicht auf den Kopf der annotierten Person nicht ausreichend vorhanden war und die Aufmerksamkeit damit uneindeutig interpretiert wurde. Letzteres bekräftigen auch die jeweiligen Übereinstimmungen: Mit Kappa-Werten $> 0,6$, gibt es bei den Hauptzielen nur selten Unstimmigkeiten zwischen den Annotatoren. Bedeutsamer werden sie allerdings bei fast allen Requisiten: An den niedrigen Kappa-Werten von zum Beispiel $0,1$ bei der Klimaanlage oder $0,2$ beim Sessel lässt sich ablesen, dass diese Ziele zum einen nur sehr selten von den Annotatoren ausgesucht wurden (die Klimaanlage wurde von Annotator 1 nur in $0,1\%$ aller Fälle als Aufmerksamkeitsziel ausgewählt). Zum anderen konnte erkannt werden, dass ein anderer Annotator nur selten mit der Wahl übereinstimmt. Damit wird insbesondere deutlich, dass das Ziel der Aufmerksamkeitszuwendung selbst für menschliche Betrachter stark vom Kontext der Situation abhängt. Mit der Qualität der vorgelegten Videoaufzeichnungen entschieden die unterschiedlichen Beobachter ohne weitere Informationen damit deutlich divergenter bezüglich derjenigen Ziele, die nicht offensichtlicher Teil des Geschehens darstellten.

Allerdings bedeutet ein Kappa-Wert < 1 ebenfalls, dass Divergenzen in den Entscheidungen vorhanden sind und die beiden im Vergleich vorliegenden Annotationen nicht immer übereinstimmen. Auch die Hauptziele weisen solche Unstimmigkeiten auf. Mit Kappa-Werten um $\sim 0,7$ hinsichtlich der Besprechungsteilnehmer stimmen die Annotatoren zwar zu einem Großteil überein, es stellt sich allerdings die Frage in welchen Situationen selbst auf diesen prinzipiellen Aufmerksamkeitszielen Unstimmigkeiten zu beobachten sind.
Tabelle 4.4 führt hierzu eine Konfusionsmatrix auf, die die Unterschiede zwischen Annotator 1 und Annotator 2 für eine der annotierten Besprechungen verdeutlichen soll. Weil mit Angabe aller Ziele die Matrix den Rahmen einer Seite sprengen würde, soll an dieser Stelle nur ein kleiner Ausschnitt gegeben werden, der sich auf die Hauptziele beschränken soll. Fett hervorgehoben ist darin insbesondere die Diagonale, die übereinstimmende Annotationen bedeutet.
Wir dargestellt ist wurde in $10,4\%$ aller Videoframes der Besprechung von beiden Annotatoren Person 1 als Aufmerksamkeitsziel ausgewählt. In ebenfalls $10,4\%$ aller Frames stimmten beide überein, dass Person 2 von einem der Teilnehmer betrachtet wurde. In $0,9\%$ der Frames, entschied sich Annotator 1 jedoch dafür, dass statt Person 1 die Leinwand fokussiert würde, der zweite Annotator entschied sogar in 2% aller Frames zugunsten der Leinwand. Dass die Leinwand überhaupt als Aufmerksamkeitsziel in Frage kommt deutet deutlich darauf hin, dass es sich um Situationen handeln muss, in denen eine Person vor der Leinwand stand oder saß, beziehungsweise einen Vortrag vor der Leinwand hielt. Es muss sich damit um Momentaufnah-

men gehandelt haben, in denen eine rein visuelle Unterscheidung wohin die zu annotierende Person blickte nicht ausreichte, weil beide Ziele sinnvoll erschienen.
Ebenso verhält es sich in Momenten, in denen die entfernte Kamera während des Vortrags geholt und vom Vortragenden den übrigen Personen erklärt und gezeigt wurde. Auch hier lässt sich eine Verwechslung in den Annotationen feststellen: In 3% aller Frames der Besprechung stimmten die beiden Annotatoren damit überein, dass die Kamera von einer Person betrachtet wurde. Hierzu stellen $1,9\%$ der Frames in denen sich Annotator 2 für eines der anderen aufgeführten Aufmerksamkeitsziele entschied, eine vergleichsweise hohe Unstimmigkeit dar.
Den Annotatoren lagen dieselben Videobilder vor, wie sie im Datensatz aufgezeichnet wurden. Damit standen die Annotatoren vor derselben Herausforderung keine detaillierte Sicht des jeweiligen Augenpaars einer Person vorgelegt zu bekommen. Stattdessen waren sie gezwungen maßgeblich anhand sekundärer Indikatoren auf die Aufmerksamkeitszuwendungen schließen zu müssen. Die Beispiele zeigen dabei dass sich das visuelle Erfassen einer Szene aus äquivalenten Blickwinkeln allein, auch für den Menschen als solches, schwierig erweist, um eine Aussage darüber treffen zu können wohin eine Person ihre Aufmerksamkeit richtet. Bezüge zu Objekten werden erst im Kontext der Situation hergestellt. Die Entscheidung wohin jemand blickt orientiert sich daran. Die Beispiele deuten deshalb darauf hin, dass die Größe eines Aufmerksamkeitsziels in direkter Abhängigkeit zur qualitativen Auflösung der Erkennung der Blickfeldorientierung steht sowie zu nicht-visuellen Situationsmerkmalen, die den Annotatoren per se nicht zur Verfügung standen.

Evaluationsgrundlage: komplette Zielmenge versus reduzierte Zielmenge

Wie in den Annotationen erkennbar wird, können jene Fokusziele hervorgehoben werden, die Bestandteil der Besprechungen waren und häufig von den Teilnehmern betrachtet wurden. Im Gegensatz dazu stehen Mobiliar und andere Objekte, die im Hintergrund Teil der Raumausstattung waren und häufig ignoriert wurden oder sogar in manchen Fällen nicht eindeutig betrachtet wurden, so dass selbst die Annotatoren jene Ziele nicht einstimming als Aufmerksamkeitsziel hervorheben konnten.
Verwandte Arbeiten die sich mit dieser Fragestellung auseinandersetzen, reduzieren die Zielmenge in der Regel auf die prinzipiellen Personen und Objekte die Bestandteil der Aufnahmen sind und ignorieren Hintergrundziele. Im Rahmen dieser Arbeit sollen deswegen vergleichsweise beide Fälle evaluiert werden: das System wird sowohl auf der gesamten Zielmenge evaluiert als auch einer solchen reduzierten Zielmenge gegenübergestellt. Die gesamte Zielmenge umfasst dabei all jene Personen und Objekte, die auch in Tabelle 4.3 aufgelistet sind. Das umfasst 38-40 Ziele, abhängig vom zugrundeliegenden Video, weil jeweils eine unterschiedliche Anzahl an Personen in den Aufnahmen vorkommen. Die reduzierte Zielmenge umfasst lediglich alle Teilnehmer, den Besprechungstisch und die Projektionsleinwand, also 5-7 Ziele, auch hier

wieder abhängig vom jeweiligen Video wegen der unterschiedlichen Anzahl an vorkommenden Personen.
Die Annotationen beziehen sich auf alle Objekte und Personen im Raum. Es liegen leider keine weiteren Annotationen vor, in denen lediglich zwischen den dominanten Zielen unterschieden wurde. Aus Grundlage für die reduzierte Zielmenge wird das System deswegen nur auf jenen Frames evaluiert, in denen die Annotatoren jeweils eines dieser Ziele als Fokusziel angaben.

		Annotator 2														
		Person 1	Person 2	Person 3	Person 4	Unterlagen 1	Unterlagen 2	Unterlagen 3	Unterlagen 4	Person Extra	Stuhl Extra	Unterlagen Extra	Tisch	Leinwand	Kamera	Handy
Annotator 1	Person 1	**10,4**	0,6	0,1	0,2	0,1	0,0	0,1	0,1	0,1	0,0	0,0	0,1	0,9	0,6	0,0
	Person 2	0,6	**10,4**	0,7	0,1	0,2	0,1	0,0	0,3	0,3	0,0	0,0	0,1	0,2	0,4	0,0
	Person 3	0,1	0,5	**8,1**	0,7	0,4	0,1	0,0	0,1	0,3	0,0	0,0	0,1	0,0	0,3	0,0
	Person 4	0,3	0,1	0,8	**7,7**	0,1	0,1	0,0	0,1	0,3	0,0	0,0	0,1	0,1	0,3	0,0
	Unterlagen 1	0,2	0,1	0,2	0,1	**3,0**	0,1	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,0
	Unterlagen 2	0,0	0,1	0,1	0,1	0,0	**2,0**	0,1	0,0	0,0	0,0	0,0	0,1	0,0	0,1	0,0
	Unterlagen 3	0,1	0,0	0,1	0,0	0,0	0,0	**2,0**	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0
	Unterlagen 4	0,1	0,2	0,2	0,1	0,1	0,0	0,0	**2,0**	0,0	0,0	0,0	0,1	0,0	0,0	0,0
	Person Extra	0,1	0,4	0,3	0,3	0,0	0,0	0,0	0,0	**7,3**	0,0	0,0	0,0	0,1	0,0	0,2
	Stuhl Extra	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,1	**0,2**	0,0	0,0	0,0	0,0	0,0
	Unterlagen Extra	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**	0,0	0,0	0,0	0,0
	Tisch	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,1**	0,0	0,0	0,0
	Leinwand	2,0	0,4	0,0	0,4	0,0	0,0	0,1	0,0	0,1	0,0	0,0	0,1	**11,9**	0,2	0,0
	Kamera	0,6	0,1	0,2	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**3,0**	0,0
	Handy	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,2**

Tab. 4.4.: Ausschnitt der Konfusionsmatrix zwischen Annotatoren 1 und 2, die beschreibt wie häufig in [%] einzelne Ziele übereinstimmend oder inkonform annotiert wurden.

4.4.2. Annahmenevaluation der Kopfdrehung als faktische Blickrichtung

Die Evaluation des beschriebenen Systemansatzes soll im Hinblick auf die Referenzannahme geschehen, dass die Kopfdrehung der eigentlichen Blickrichtung gleichgesetzt werden kann. Darin soll jenes Ziel als Aufmerksamkeitszuwendung klassifiziert werden, das der geometrischen Sichtgerade einer Person euklidisch am nächsten erscheint. Die Sichtgerade hat dabei als Ursprung den Kopfmittelpunkt und wird entsprechend des Kopfdrehwinkels rotiert.
Für alle Aufmerksamkeitsziele wird daraufhin die jeweilige Distanz zur Sichtgerade berechnet. Schneidet die Sichtgerade das mögliche Fokusziel, wird der Schnitt als faktischer Blickpunkt festgelegt. Falls kein Objekt von der Gerade geschnitten wird, wird jenes als Aufmerksamkeitsziel interpretiert, zu dessen Kanten die Sichtgerade am dichtesten verläuft.
Im Fall verdeckter Ziele wird die Gerade mit allen Objekten geschnitten und jenes als Fokusziel hervorgehoben, das dem Kopf und der Sichtgerade am nächsten liegt.

Die Resultate sind in den Tabellen 4.5 und 4.6 aufgelistet. Im Detail stellt Tabelle 4.5 hierfür die Ergebnisse dar, wenn anhand der geschätzten Kopfdrehung zwischen den hauptsächlichen Besprechungszielen unterschieden wird. Diese umfassen dabei die Teilnehmer, die Projektionsleinwand sowie den Besprechungstisch. Darüber hinaus wird in den Ergebnissen der Vergleich gegenübergestellt, wenn nicht geschätzte Kopfdrehungen sondern protokollierte als Ausgangsbasis dienten. Infolgedessen ist die qualitative Auflösung der Blickfeldorientierung höher - weil aber nur Protokolle für einen der Besprechungsteilnehmer pro Video vorliegen, beschränken sich die Evaluationen auf eine einzelne Person. Tabelle 4.6 gibt die Korrektklassifikationsrate an, wenn alle Aufmerksamkeitsziele im Raum einbezogen wurden. Respektive wird hierbei auch wieder zwischen den Ergebnissen mit geschätzten Kopfdrehungen und protokollierten unterschieden. Die Menge der Ziele umfasste dabei alle Personen, alle Objekte und alles Mobiliar im Raum (40 Ziele im gesamten). Die dargestellte Konfusionsmatrix in Tabelle 4.7 verdeutlicht die Unterschiede zwischen den Hypothesen auf protokollierten Kopfdrehungen und den manu-

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	63,2	58,7	43,5	53,6	53,9
	konf. Annot.	70,9	65,2	46,9	58,5	59,8
Winkelprotokolle	alle Frames	-	-	45,0	-	-
	konf. Annot.	-	-	50,9	-	-

Tab. 4.5.: Korrektklassifikationsrate, wenn die Kopfdrehung geometrisch als Blick interpretiert wird. Menge einbezogener Ziele: Personen, Tisch, Leinwand (insg. 5-7 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.1 (Seite: 182).

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	33,0	38,3	30,6	25,7	32,1
	konf. Annot.	37,2	43,0	31,9	28,1	35,3
Winkelprotokolle	alle Frames	-	-	31,5	-	-
	konf. Annot.	-	-	36,2	-	-

Tab. 4.6.: Korrektklassifikationsrate, wenn die Kopfdrehung geometrisch als Blick interpretiert wird. Menge einbezogener Ziele: alle Personen, Objekte und Mobiliar (insg. 38-40 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.2 (Seite: 183).

ellen Annotationen. Die Konfusionsmatrix in Tabelle 4.8 stellt hingegen die protokollierten mit den geschätzten Drehwinkeln derselben Person in einen Vergleich.

In den Ergebnissen wird deutlich, dass eine Interpretation der Kopfdrehung als Blickrichtung zur Folge hat, dass immer das frontal erscheinende Objekt als Aufmerksamkeitsziel selektiert wird, ungeachtet dessen wie sich die Situation entwickelt. Für den Fall einer Präsentation vor der Leinwand hat das zur Folge, dass immer nur dann die Leinwand als Fokusziel klassifiziert wird, wenn der Vortragende nicht die Sicht versperrt. In allen übrigen Momenten, in denen der Sprecher zum Beispiel zur Seite ausweicht, um auf selektive Bereiche auf der Leinwand hinzudeuten, beziehungsweise sich generell während des Vortrags vor der Leinwand bewegt, entfernt sich dieser von der Sichtgeraden und gerät damit aus dem Fokus der Aufmerksamkeit. Im Gegensatz hierzu entsprechen die Annotationen nicht der interpretierten Blickrichtung, sondern lassen den Fokus weiter auf dem Sprecher haften. In den aufgeführten Konfusionsmatrizen wird das dadurch verdeutlicht, dass in $10,7\%$ aller Frames die Leinwand als Fokusziel hypothetisiert wird, die Annotation aber zum Beispiel für Person 1 als Sprecher votiert. Umgekehrt wurde in $1,6\%$ aller Frames Person 1 als Sprecher hypothetisiert, die Annotationen ordneten den Fokus jedoch der Leinwand zu. Mit nur $8,7\%$ bzw. $13,1\%$ der Frames in denen Person 1 bzw. die Leinwand korrekt hypothetisiert wurde, hat das zur Folge, dass in mehr als der Hälfte aller Frames falsch zwischen diesen beiden Zielen unterschieden wurde. In genanntem Beispiel wurde deutlich gemacht, dass die zu annotierende Person ihre Aufmerksamkeit auf dem Sprecher belässt, auch wenn dieser sich von der Leinwand fortbewegt. Daneben konnte beobachtet werden, dass der Blick auf den Folien belassen wurde, auch wenn der Sprecher die Sicht zur Leinwand kreuzte und deutlich in das Blickfeld der jeweiligen Person geriet. Die Annotationen machen in solchen Momenten deutlich, dass nicht allein aufgrund der Kopfdrehung zwischen den beiden Zielobjekten unterschieden wird, sondern manchmal das Geschehen den ausschlaggebenden Faktor darstellt: mit Gesten und Sprache zieht der Vortragende die Aufmerksamkeit hierbei auf sich, auch wenn die als Blick interpretierte Kopfdrehung eindeutig auf die Leinwand schließen lässt. Der Vortragende bezieht sich unter Umständen nicht direkt auf Informationen

		Annotation A1						
		Person 1	Person 2	Person 3	Person 4	Person Extra	Tisch	Leinwand
Hypothesen (Winkelschätzung)	Person 1	**8,7**	0,0	1,5	2,4	1,4	0,1	1,6
	Person 2	0,6	**0,1**	1,9	1,2	2,6	0,0	0,0
	Person 3	0,0	0,0	**0,0**	0,0	0,0	0,0	0,0
	Person 4	0,9	0,0	0,4	**1,6**	2,5	0,1	0,3
	Person Extra	0,4	0,0	0,2	0,4	**3,3**	0,0	0,0
	Tisch	2,2	0,0	1,8	2,8	0,6	**0,0**	5,1
	Leinwand	10,7	0,2	1,3	4,2	3,5	0,1	**13,1**

Tab. 4.7.: Konfusionsmatrix der Hypothesen für Person 3 in Besprechungsszenario 1, bezogen auf Annotationen A1. Die Matrix beschreibt wie häufig in [%] einzelne Ziele übereinstimmend oder inkonform annotiert wurden.

auf der Leinwand, sondern lenkt von diesen ab. Im Gegensatz treten Momentaufnahmen auf, in denen der Kopf eindeutig dem Sprecher zugewandt erscheint, anhand der Gestik des Vortragenden aber deutlich wird, dass dieser sich auf selektive Bereiche auf der Leinwand bezieht und die Aufmerksamkeit des Publikums damit von sich fernhält. Die Erfassung der Gesten in solchen Moment führt in den Annotationen dazu, dass die Kopfdrehung von den Annotatoren nur unwesentlich zur Zielentscheidung beiträgt - ein Aspekt der sowohl bei der rein geometrischen Klassifikation als auch bei der im System beschriebenen Vorgehensweise noch gänzlich unberücksichtigt bleibt. Gestik stellt damit ein visuelles Kontextmerkmal dar, das zwar den Annotatoren zugänglich war aber erst durch eine explizite Modellierung im System von diesem berücksichtigt werden kann. In solchen Situationen erscheint die Kopfdrehung nur als Abbild der eigentlichen Aufmerksamkeit, tatsächlich orientiert sich die Zuwendung aber an anderen Faktoren. Schnelle Fluktuationen der Augen erlauben dabei den flinken Blickwechsel auf externe Reize, die innerhalb des Fixierfelds deutlich wahrgenommen werden können ohne den Kopf der Zuwendung beisteuern zu lassen. Im Gegenzug deutet die träge Orientierung des Kopfs an, womit sich eine Person eigentlich beschäftigt, auch wenn der Fokus dabei nicht frontal vor ihr liegen muss.

Wie auch schon in Abschnitt 4.2 aufgeführt wurde, zeigt sich diese Divergenz zwischen der Kopfdrehung und Blickrichtung in all jenen Situationen, in denen kontinuierlich zwischen zwei Zielen gewechselt wird. Das ist zum Beispiel immer dann der Fall, wenn in einer Diskussion die Aufmerksamkeit einer Person zum aktiven Sprecher folgt. Die Kopfdrehung scheint dabei immer dort zum liegen zu kommen, wo der minimale Drehwinkel eine schnelle Reaktion auf den weiteren Verlauf der Handlung erlaubt. In der Regel entspricht das bei einem Wechsel zwischen zwei Personen einer Orientierung, die mittig zwischen diesen erscheint. Die direkte

Interpretation der Kopfdrehung als Blickrichtung entscheidet in solchen Fällen für ein eventuelles Objekt im Hintergrund, das die Sichtgerade direkt schneidet. Existiert ein solches nicht wird stattdessen zu der nächstliegendsten Person tendiert.
Die geometrische Abbildung der Kopfdrehung auf Fokusziele kann allein aufgrund des gänzlich unberücksichtigten Kontexts kein gesamtheitliches Schließen der Zuwendungsorientierung ermöglichen. Bereits in der Literatur konnte nachgewiesen werden, dass die Referenzwinkel der Kopfdrehung bei einer Aufmerksamkeitszuwendung zu einem Zielobjekt nicht den faktischen Blickwinkeln entsprechen müssen. Dementsprechend werden bereits individuelle Faktoren, die hier eine ausschlaggebende Rolle spielen wie der Kopf einem Ziel zugewandt wird, gänzlich ignoriert. Zusammenfassend kann damit gesagt werden, dass der Blick ein Abbild dessen gibt, was der Betrachter überwacht und sucht, die Kopfdrehung aber eher der trägen Ausrichtung der eigentlichen Aufmerksamkeit entspricht. Bei hinreichend disjunkten Zielobjekten, lässt die Kopfdrehung zwar einen direkten, geometrischen Bezug zur Zuwendungsorientierung zu, in Situationen wo jedoch keine ausreichende Disjunktivität der Ziele gegeben ist, sind weitere Merkmalsbeobachtungen unabdingbar die den Kontext des Geschehens vervollständigen und so wertvolle Hinweise über die eigentliche Motivation einer Aufmerksamkeitszuwendung geben.

Hypothesen: geom. Ansatz (Winkelprotokolle)	Hypothesen: geom. Ansatz (Winkelschätzung)														
		Person 1	Person 2	Person 3	Person 4	Person Extra	Unterlagen 1	Unterlagen 2	Unterlagen 3	Unterlagen 4	Unterlagen Extra	Stuhl 1	Stuhl 2	Stuhl 3	Stuhl 4	Leinwand	Kamera
Person 1		**10,2**	0,1	0,0	0,2	0,1	1,4	0,0	0,0	0,3	0,0	1,1	0,0	0,0	0,0	0,6	0,0
Person 2		0,0	**0,8**	0,0	0,1	0,0	0,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Person 3		0,0	0,0	**0,0**	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Person 4		0,0	0,0	0,0	**1,4**	0,0	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Person Extra		0,1	0,0	0,0	0,0	**0,2**	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Unterlagen 1		1,1	0,1	0,0	0,0	0,0	**5,6**	0,0	0,0	3,4	0,0	0,4	0,0	0,0	0,0	1,2	0,0
Unterlagen 2		0,0	0,0	0,0	0,0	0,0	0,0	**0,0**	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Unterlagen 3		0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,4**	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Unterlagen 4		0,5	0,0	0,0	0,6	0,0	0,0	0,0	0,1	**3,2**	0,0	0,0	0,0	0,0	0,0	0,2	0,0
Unterlagen Extra		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**	0,0	0,0	0,0	0,0	0,0	0,0
Stuhl 1		1,5	0,0	0,0	0,0	0,0	2,0	0,0	0,0	0,1	0,0	**2,0**	0,0	0,0	0,0	0,1	0,0
Stuhl 2		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**	0,0	0,0	0,0	0,0
Stuhl 3		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**	0,0	0,0	0,0
Stuhl 4		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	**0,0**	0,0	0,0
Leinwand		0,1	0,1	0,0	0,2	0,0	0,6	0,0	0,0	1,2	0,0	0,0	0,0	0,0	0,0	**1,6**	0,0
Kamera		0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**

Tab. 4.8.: Konfusionsmatrix der Hypothesen für Person 3 in Besprechungsszenario 1.

4.4.3. Evaluation des entworfenen Systemansatzes

Im folgenden soll der eigentliche Systemansatz - der in Abschnitt 4.3 vorgestellt wurde - evaluiert werden. Das Verfahren nimmt Rücksicht auf Kontextbezüge und adaptiert das Beobachtungsmodell, um mit der sich ändernden Zielanordnung und dem unterschiedlichen Interaktionsgeschehen in der Szene umgehen zu können.
Um insbesondere den Einfluss und entstehenden Mehrwert der berücksichtigten Kontextmerkmale hervorzuheben, sind die Evaluationen in zwei Abschnitte aufgeteilt. Im ersten werden die Ergebnisse des gesamten Systems aufgeführt. Die Kontextbeobachtungen sind in den A-Priori-Wahrscheinlichkeiten einbezogen und beeinflussen damit die Adaption. Im zweiten Abschnitt sind die Kontextbeobachtungen sowohl komplett ignoriert oder nur teilweise einbezogen worden, um ihren jeweiligen Einfluss genauer zu untersuchen.

Korrektklassifikationsrate der Systemhypothesen mit Kontextbezug

Die Ergebnisse inklusive Kontextbezug stehen in direktem Kontrast zum Referenzsystem in Abschnitt 4.4.2, das die Kopfdrehung geometrisch als Blickrichtung interpretierte. Das Beobachtungsmodell bezieht sich hierbei zunächst auf Verhaltensmuster und Rauschen der gemessenen Kopfdrehungen, das Einbeziehen der Kontextmerkmale soll darüber hinaus aber gezielt während der Adaption einen Bezug zum Geschehen herstellen. Damit soll versucht werden, unerwartete Reize und Attraktionen im Kontext der Situation und des Geschehens erfassen zu können (Bewegung und Sprachaktivität) und die Systemannahme über das Zuwendungsverhalten einer Person entsprechend zu nivellieren. Mit der priorisierten Adaption, die durch den Kontextbezug geschieht, verliert das Beobachtungsmodell nicht allesamt seinen Ursprung. Eine Verschiebung der Mittelwerte geschieht nur in Abhängigkeit zu überraschendem Verhalten. So soll bei einer Anpassung der Modelle auf die aktuelle Kopfdrehung ein eindeutiger Bezug auf das zu berücksichtigende Aufmerksamkeitsziel sichergestellt werden. Weil dafür die Bewegung und Sprachaktivität der Personen zu Rate gezogen wird, darf es nicht überraschen,

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	67,0	63,5	54,8	58,7	60,5
	konf. Annot.	74,8	71,8	60,0	62,6	67,1
Winkelprotokolle	alle Frames	-	-	59,9	-	-
	konf. Annot.	-	-	66,9	-	-

Tab. 4.9.: Korrektklassifikationsrate des Systems mit Kontextbezug. Menge einbezogener Ziele: alle Personen, Tisch, Leinwand (insg. 5-7 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.3 (Seite: 184).

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	48,0	45,1	39,1	41,4	43,1
	konf. Annot.	52,9	51,6	41,8	45,0	47,6
Winkelprotokolle	alle Frames	-	-	38,5	-	
	konf. Annot.	-	-	42,8	-	

Tab. 4.10.: Korrektklassifikationsrate des Systems mit Kontextbezug. Menge einbezogener Ziele: alle Personen, Objekte, Mobiliar (insg. 38-40 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.4 (Seite: 185).

wenn kein eindeutiger Bezug zu Requisiten oder Hintergrundobjekten hergestellt werden kann. Infolgedessen stellen die Ergebnisse bei der Reduktion auf die prinzipiellen Aufmerksamkeitsziele, wie sie in Tabelle 4.9 aufgelistet werden, auch die entsprechende Leistungssteigerung, im Gegensatz zu allen vorhandenen Zielobjekten, dar. Letztgenannte Evaluation auf der gesamten Zielmenge wird hierbei in Tabelle 4.10 aufgeführt. Wie insbesondere durch die Konfusionsmatrix in Tabelle 4.11 gezeigt wird, hat der Kontextbezug eine deutliche Steigerung bei der Unterscheidung zur Folge. Im Vergleich zur geometrischen Interpretation wird die Leinwand hier deutlich seltener klassifiziert. Infolgedessen treten falsche Annahmen weit weniger auf als zuvor und die Hypothesen entsprechen den mit tatsächlichen Annotation erkennbar überein. Das zeigt, dass im Vergleich zu erstgenannter Evaluation, in Fällen in denen Personen dicht bei der Leinwand standen, solche auch als Ziel zugeordnet werden konnten und nicht stattdessen die offensichtliche Leinwand als Hypothese erschien - obwohl diese dichter zum Kopfdrehungsvektor

		Annotationen A1						
		Person 1	Person 2	Person 3	Person 4	Person Extra	Tisch	Leinwand
Hypothesen (Winkelschätzung)	Person 1	**17,8**	0,1	2,5	5,0	4,5	0,1	7,4
	Person 2	0,3	**0,1**	1,0	0,5	2,0	0,0	0,0
	Person 3	0,0	0,0	**0,0**	0,0	0,0	0,0	0,4
	Person 4	0,8	0,0	0,2	**0,9**	1,6	0,0	0,4
	Person Extra	0,2	0,0	0,0	0,1	**0,8**	0,0	0,1
	Tisch	0,1	0,0	0,2	0,0	0,0	**0,0**	1,1
	Leinwand	0,2	0,0	0,1	0,6	0,1	0,1	**5,8**

Tab. 4.11.: Ausschnitt der Konfusionsmatrix für Besprechungsszenario 1, Person 3 für Hypothesen mit Kontextbezug. Die Matrix beschreibt wie häufig in [%] die Zielhypothesen mit den Annotationen übereinstimmen und wie häufig Verwechslungen mit anderen Zielen auftraten.

		Annotation A1						
		Person 1	Person 2	Person 3	Person 4	Person Extra	Tisch	Leinwand
Hypothesen (Winkelprotokolle)	Person 1	**19,1**	0,1	2,4	6,0	5,4	0,0	9,1
	Person 2	2,4	**0,1**	3,2	3,5	2,7	0,1	0,1
	Person 3	0,0	0,0	**0,0**	0,0	0,0	0,0	0,8
	Person 4	0,5	0,0	0,4	**1,6**	1,5	0,2	0,1
	Person Extra	0,2	0,0	0,5	0,6	**4,1**	0,0	0,1
	Tisch	0,6	0,0	0,4	0,3	0,1	**0,0**	3,6
	Leinwand	0,7	0,0	0,2	0,7	0,2	0,1	**6,4**

Tab. 4.12.: Ausschnitt der Konfusionsmatrix für Besprechungsszenario 1, Person 3, in der die Systemhypothesen mit Kontextbezug bei protokollierten Kopfdrehungen im Vergleich zu den Zielannotationen aufgetragen sind. Die Matrix beschreibt wie häufig in [%] die Zielhypothesen mit den Annotationen übereinstimmen und wie häufig Verwechslungen mit anderen Zielen auftraten.

liegen musste. Die Unterscheidung anhand des Kontextbezugs entsprach damit den menschlichen Annotationen. Das Blickfeld muss demnach zur Leinwand ausgerichtet gewesen sein und die Personen darin agiert haben. Unter diesem Aspekt wird deutlich, dass der Mensch hier ähnlich seine Entscheidung anhand äußerer Reize für die beobachtete Person bestimmt und nicht anhand der prinzipiellen Ausrichtung des Blickfelds. Bewegung und Verdeckung kann damit

		Hypothesen (Winkelprotokolle)						
		Person 1	Person 2	Person 3	Person 4	Person Extra	Tisch	Leinwand
Hypothesen (Winkelschätzung)	Person 1	**48,8**	4,1	0,0	0,7	0,2	1,6	3,3
	Person 2	0,8	**10,2**	0,0	0,0	4,1	0,0	0,0
	Person 3	0,0	0,0	**0,3**	0,0	0,0	0,1	0,0
	Person 4	1,0	0,0	0,0	**4,2**	0,1	0,4	0,2
	Person Extra	0,2	0,2	0,0	0,0	**1,3**	0,0	0,0
	Tisch	0,1	0,6	0,4	0,0	0,0	**4,1**	0,0
	Leinwand	5,0	0,4	0,0	0,0	0,0	0,4	**7,1**

Tab. 4.13.: Ausschnitt der Konfusionsmatrix für Besprechungsszenario 1, Person 3, in der die Systemhypothesen mit Kontextbezug bei protokollierten Kopfdrehungen im Vergleich zu geschätzten Kopfdrehungen aufgetragen sind. Die Matrix beschreibt wie häufig in [%] die Zielhypothesen übereinstimmen und wie häufig Verwechslungen mit anderen Zielen auftraten.

nachgewiesen als A-Priori-Motivation verstanden werden, die für außenstehende Betrachter die Aufmerksamkeit einer Person auf sich zieht.
Im Vergleich zur geometrischen Interpretation wird auch deutlich häufiger Person 1 als eigentliches Ziel erkannt - wie die Annotationen zeigen das auch zurecht. Es zeigt sich damit, dass Person 1 nicht immer der Blickrichtung entsprach, wohl aber als Ziel von den Annotatoren aufgefasst wurde. Mit der gesunkenen Häufigkeit mit der die Leinwand statt Person 1 erkannt wurde (0,2% hier gegen 10,7% geometrisch) kann man erkennen, dass hierbei Situationen gemeint sein müssen, in denen Person 1 vor der Leinwand einen Vortrag hält. Infolgedessen muss Person 1 vor der Leinwand gestanden, den Fokus aber fluktuierend von sich auf die Folien gelenkt haben. Dass hierbei korrekt der Vortragende klassifiziert werden konnte zeigt, dass seine Bewegung die Aufmerksamkeit von Person 3 auf sich lenken musste - zumindest aus Sicht der Annotatoren.

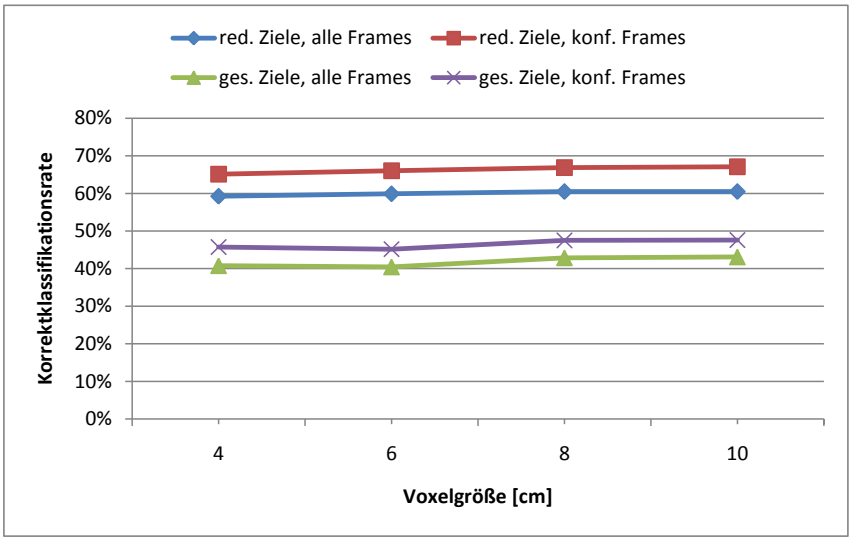


Abb. 4.13.: Korrektklassifikationsrate der Aufmerksamkeitsschätzung bei unterschiedlicher Voxelgröße.

In Abbildung 4.13 ist zusätzlich der Einfluss der Voxelkantenlänge auf die Korrektklassifikationsrate des Gesamtsystems dargestellt. Evaluiert wurden die Kantenlängen 4 cm, 6 cm, 8 cm und 10 cm. Größere Kantenlängen konnten nicht evaluiert werden, weil das kleinste Objekt in der Zielmenge sonst nicht mehr durch ein einzelnes Voxel hätte dargestellt werden können. Mit kleiner werdender Kantenlänge steigt auch der Rechenaufwand pro Zeitschritt entsprechend an. Kleinere Voxelkantenlängen als 4 cm wurden deshalb aufgrund des notwendigen Rechenaufwands außer Acht gelassen.
Auffällig ist, dass die Voxelkantenlänge nur marginal Einfluß auf die Klassifikationsrate des Systems nimmt. Mit geringerer Größe sinkt die Korrektklassifikationsrate ein wenig, die höchste Genauigkeit konnte bei 10cm Kantenlänge beobachtet werden. Mit der geringer werdenden Kantenlänge steigt die Anzahl der Voxel pro Zielobjekt. Damit verbunden steigt auch die Anzahl der Normalverteilungen im Beobachtungsmodell, welche pro Voxel zum Einsatz kommen.

Mit der Adaption der Modelle zu den stetig zu beobachtenten Kopfdrehungen, verschieben sich die Normalverteilungen auf Voxelebene und konvergieren zu den Kopfdrehungen - mehr oder weniger stark, ganz abhängig von der A-Posteriori-Wahrscheinlichkeit des entsprechenden Fokusziels, momentan betrachtet zu werden. In der initialen Repräsentation sind die Glockenfunktionen hierbei noch in den Voxelmittelpunkten verankert und bilden damit eine gleichmäßige Abdeckung über die kompletten Objektmaße. Nach kurzer Zeit sind die Modelle jedoch verschoben zu den eigentlichen Voxeln - weil sie sich der Kopfdrehung angepasst haben - und bilden eine Zielrepräsentation, die der Silhouettenform des Objekts nicht mehr äquivalent sein muss. Insbesondere bei falsch klassifizierten Aufmerksamkeitszielen kann beobachtet werden, dass die Modellrepräsentationen von der eigentlichen Objektform weg divergiert. Mit steigender Anzahl an Voxelmodellen steigt auch die Gefahr einer solchen Fehlrepräsentation und eines solchen Verhaltens zur Laufzeit - ein interessanter Grund wieso die Korrektklassifikationsrate bei feingranularer Voxelauflösung eher sinkt statt zur Genauigkeit beiträgt. Damit wird ersichtlich, dass die einbezogenen Kontextbeobachtungen nicht erschöpfend sind und eine fehlerhafte Adaption sich auf die interne Repräsenation der Ziele überträgt und damit akkumulierend fortgeführt wird.

Untersuchung des Einfluß des Kontextbezugs

Ohne Kontextbezug werden die Zuwendungshypothesen allein aufgrund des adaptiven Beobachtungsmodells getroffen. Dieses passt sich automatisch den Kopfdrehungen zur Laufzeit an, unbeeinflusst von weiteren Situationsmerkmalen. In jedem Sequenzschritt einer zu evaluierenden Besprechung wird es damit den Kopforientierungen angepasst. Damit divergieren die repräsentativen Modelle von den eigentlichen Aufmerksamkeitszielen weg, wenn die Kopfdrehung ausreichend lange vom jeweiligen Zielobjekt weggerichtet erscheint. Weil die Modelle für alle Ziele gleichzeitig angepasst werden, verschieben sich die Referenzwinkel aller Objekte und Personen und insbesondere aller Voxel im Einzelnen. Die Adaption erfolgt dabei nur noch in Anbhängigkeit dazu, wie gut die Modelle der aktuellen Kopfdrehung entsprechen. Für dicht platzierte Ziele, beziehungsweise Objektmodelle mit breiter Standardabweichung oder bereits weit verschobenem Referenzmittelwert, erfolgt die Adaption hierbei äquivalent. Ein Bias auf dedizierte Ziele fehlt beim Aktualisieren somit komplett und einer Divergenz wird auch entsprechend nicht vorgebeugt. Infolgedessen ist bereits nach kurzer Zeit zu beobachten, dass die Modelle nicht mehr einer Reflektion der Zielanordnung gleichen, sondern allesamt ausschließlich der Kopfdrehung nachziehen und der eigentliche Bezug zum Zielobjekt verloren geht.
Die Ergebnisse spiegeln die Divergenzen wider. In Tabelle 4.14 sind die Resultate wieder getrennt für die prinzipiellen Aufmerksamkeitsziele der Besprechungen, in Tabelle 4.15 die entsprechenden bei einer Berücksichtigung aller Ziele. Ersichtlich wird darin besonders, dass der Wertebereich der Korrektklassifikationsrate in beiden Fällen ähnlich ausfällt. Das überrascht zu-

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	20,5	26,1	28,7	31,0	26,8
	konf. Annot.	19,1	25,7	25,3	29,1	24,9
Winkelprotokolle	alle Frames	-	-	30,7	-	-
	konf. Annot.	-	-	27,9	-	-

Tab. 4.14.: Korrektklassifikationsrate des Systems ohne Kontextbezug. Menge einbezogener Ziele: alle Personen, Tisch, Leinwand (insg. 5-7 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.5 (Seite: 186).

nächst, weil die größere Anzahl vorhandener Ziele eigentlich durch ihre geringe Disjunktivität in der Zielanordnung zu deutlich häufigeren Mehrdeutigkeiten führen müsste. Berücksichtigt man allerdings die vorab beschriebene Divergenz der Modelle, kann nachvollzogen werden, dass dies in beiden Fällen zu einer äquivalenten Adaption der Modelle führt, womit sich diese infolgedessen bei der Klassifikation gleichen. Verdeckte Hintergrundobjekte werden durch die Sichtbarkeitsprüfung ignoriert und die Unterscheidung der Ziele im Vordergrund ähnelt vom Prinzip einer Reduktion auf die eigentlich beteiligten Personen und Objekte. Insbesondere bedeutet das, dass sich die beiden verschiedenen Zielmengen in solchen Momenten die Wage halten. In der in Tabelle 4.16 dargestellten Konfusionsmatrix wird insbesondere deutlich, dass im Vergleich zur rein geometrischen Interpretation (siehe Tabelle 4.7) viel deutlicher den übrigen Personen die Aufmerksamkeit zugeordnet wird. Person 1 wird so weit häufiger mit den anderen Teilnehmer verwechselt, während die geometrische Zuordnung diese erkennbar voneinander trennt. Offensichtlich stehen diese dabei ausreichend disjunkt voneinander, dass eine direkte Verwechslung von der Kopfdrehung allein nicht möglich erscheint. Das unterstreicht hier klar erkennbar die verschobenen Modellrepräsentationen, die offenbar zueinander konvergieren und damit keine disjunkte Diskriminierung mehr erlauben.

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	12,3	31,5	34,6	29,0	28,0
	konf. Annot.	11,9	28,8	33,3	27,7	26,3
Winkelprotokolle	alle Frames	-	-	35,1	-	
	konf. Annot.	-	-	33,4	-	

Tab. 4.15.: Korrektklassifikationsrate des Systems ohne Kontextbezug. Menge einbezogener Ziele: alle Personen, Objekte und Mobiliar (insg. 38-40 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.6 (Seite: 187).

		Annotation A1						
		Person 1	Person 2	Person 3	Person 4	Person Extra	Tisch	Leinwand
Hypothesen (Winkelschätzung)	Person 1	**19,7**	0,1	5,1	8,1	9,1	0,2	17,9
	Person 2	1,4	**0,0**	1,1	2,5	0,7	0,1	0,1
	Person 3	0,3	0,0	**0,1**	0,1	0,0	0,0	0,0
	Person 4	0,8	0,0	0,3	**1,4**	0,0	0,0	1,8
	Person Extra	1,3	0,2	0,5	0,4	**4,2**	0,0	0,5
	Tisch	0,0	0,0	0,0	0,0	0,0	**0,0**	0,0
	Leinwand	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**

Tab. 4.16.: Ausschnitt aus der Konfusionsmatrix in Besprechungsszenario 1, Person 3 für Hypothesen ohne Kontextbezug. Die Matrix beschreibt wie häufig in [%] die Zielhypothesen mit den Annotationen übereinstimmen und welche Verwechslungen mit anderen Zielen auftraten.

Im Gegensatz zum geometrischen Schließen wird ferner auch die Leinwand nicht mehr als Ziel klassifiziert. Während dort noch in 30% aller Frames ein Blick dorthin interpretiert wird, erscheint die Leinwand hier kein mal als Zielausgabe. Wie man insbesondere in Tabelle 4.7 sehen kann, wird dort die Leinwand häufig mit Personen verwechselt. Daran ist zu sehen, dass in solchen Momenten die entsprechenden Ziele nah positioniert waren, weil sie entsprechend dicht an der Sichtgerade zur Leinwand erscheinen mussten, um mit ihr verwechselt zu werden. Dieselben Trajektorien lagen somit auch für die hier gezeigte Konfusionsmatrix vor. Dass jedoch im Vergleich dazu stets gegen die Leinwand entschieden wurde, zeigt deutlich, dass deren Modell prägnant von ihrer Lage fort divergiert ist. Mit den Ergebnissen in Tabellen 4.17 und 4.18 ist der explizite Nutzen von Bewegung als alleinigem Kontextbeobachtung erkennbar. Für die reduzierte Zielmenge ist dabei eine Klassifikationsratensteigerung von $26,8\%$ auf $60,5\%$ beobachtbar, bei der gesamten Zielmenge von 28% auf 41%. Es wird damit deutlich erkennbar, wie

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	67,0	63,4	54,6	58,6	60,5
	konf. Annot.	74,7	71,5	60,1	62,8	67,0
Winkelprotokolle	alle Frames	-	-	60,3	-	-
	konf. Annot.	-	-	67,7	-	-

Tab. 4.17.: Korrektklassifikationsrate des Systems mit Bewegung als einzigem Kontextbezug. Menge einbezogener Ziele: alle Personen, Tisch und Leinwand (insg. 5-7 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.7 (Seite: 188).

		KKR pro Teilnehmer				KKR ∅
Kopfdrehung	vFoA Annotationen	1	2	3	4	
Winkelschätzung	alle Frames	45,6	43,2	37,6	38,4	41,0
	konf. Annot.	50,2	49,3	40,4	41,7	45,3
Winkelprotokolle	alle Frames	-	-	37,0	-	-
	konf. Annot.	-	-	41,3	-	-

Tab. 4.18.: Korrektklassifikationsrate des Systems mit Bewegung als einzigem Kontextbezug. Menge einbezogener Ziele: alle Personen, Objekte und Mobiliar (insg. 38-40 Ziele, mit untersch. Anzahl Personen pro Video). (Detaillierte Ergebnisse einzelner Besprechungen und bezüglich aller Annotatoren, siehe Tabelle D.8 (Seite: 189).

stark der Einfluss der Bewegung die Aufmerksamkeit und die Adaptierung der Modelle auf die bewegten Personen im Raum konzentriert. Insbesondere die Adaptierung profitiert davon, weil mit der Bewegung die prägnanten Fokusziele hervorgehoben und die Modelle entsprechend auf sie statt auf nah gelegene Hintergrundobjekte konvergieren. Vor dem Hintergrund der Ergebnisse mit Sprachaktivität als weitere Kontextbeobachtung (den Ergebnissen des Gesamtsystems), lässt sich ablesen, dass die Hinzunahme dieser weiteren Modalität mit $60,5\%$ auf der reduzierten Zielmenge keinen Einfluß zeigt, bei der Gesamtzielmenge aber ein kleiner Beitrag deutlich wird: hier steigt die Korrektklassifikationsrate von $41,0\%$ mit Bewegung allein auf $43,1\%$, wenn Bewegung und Sprachaktivität einbezogen werden. Das hängt damit zusammen, dass in der reduzierten Zielmenge die Fokusziele zu disjunkt positioniert sind, als dass diese Beobachtung zu einer verbesserten Unterscheidung anhand der Kopfdrehung beitragen könnte. Mit der Gesamtzielmenge treten aber plötzlich neben den dominanten Zielen auch Hintergrundobjekte auf, die dicht gelegen mit dem eigentlichen Aufmerksamkeitsfokus verwechselt werden können. Hier hilft die Sprachaktivität, um bei der Adaptierung nicht mehr auf diese zusätzlichen Objekte einzugehen, sondern iterativ die Personen hervorzuheben. Darüberhinaus ist mit der Sprachaktivität ein Merkmal modelliert worden, das nur sehr spärlich zu beobachten ist - im Gegensatz zur Bewegung der Teilnehmer, welche ständig auftritt und damit regelmäßig Überraschungsbezüge in den A-Priori-Wahrscheinlichkeiten beisteuert.

5. Anwendung: Visuelle Perzeption für die Mensch-Maschine-Interaktion

Über das bereits genannte CHIL-Projekt [WS09] hinaus flossen Bestandteile dieser Arbeit in das Fraunhofer-interne Forschungsprojekt "*Visuelle Perzeption für die Mensch-Maschine-Interaktion - Interaktion des Menschen in und mit aufmerksamen Räumen*" am Fraunhofer Institut für Optronik, Systemtechnik und Bildauswertung (IOSB) [SG10]. Ziel der Forschungsgruppe ist unter anderem die Konzeptionierung, Entwicklung und Evaluierung intuitiver und menschlicher Eingabemodalitäten in aufmerksamen Umgebungen. Das Erkennen der Kopfdrehung und des visuellen Aufmerksamkeitsfokus wird darin eingesetzt, um einerseits notwendige Information für weitere perzeptive Verarbeitungsschritte zu erhalten und andererseits Benutzer- und Situationsmodelle sukzessive zu vervollständigen.
Im diesem Kapitel soll das Konzept des entworfenen *Smart-Rooms* erläutert werden. Daneben soll die Einbettung der in dieser Arbeit entwickelten Komponenten den weiteren zum Einsatz kommenden Perzeptionsmodulen gegenübergestellt werden. Hierzu werden die jeweiligen Komponenten zur Erfassung eines vollständigen Benutzermodells vorgestellt und deren Zusammenwirken für eine intuitive Mensch-Maschine-Kommunikation in aufmerksamen Umgebungen beschrieben.

5.1. Der Smart-Room

Die Motivation des Smart-Rooms ist es autonom durch visuelle (und gegebenenfalls akustische) Perzeption einen Eindruck des Geschehens in ihm zu erhalten und auf Personen sowie deren Absichten und Handlungen zu reagieren. Damit soll eine Form der Mensch-Maschine-Interaktion bereitgestellt werden, die auf periphere Eingabemodalitäten verzichtet und sich stattdessen auf menschliche Formen der Interaktion und Kommunikation konzentriert. Statt einzelne Arbeitsplatzrechner bedienen zu müssen, soll die kooperative Arbeit in einer Gruppe in den Vordergrund gestellt werden oder einzelnen Personen die Möglichkeit geboten werden sich auf ihre eigentliche Aufgabe konzentrieren zu können. Dabei soll insbesondere der kognitive Mehraufwand vermieden werden, der sich durch die Bedienung peripherer Eingabegeräte und vorgeschriebener Anwendungsprozesse der zu bedienenden Software unweigerlich aufzwingt. Perzeptive Bildverarbeitungskomponenten werden eingesetzt, um den Handlungsrahmen aller Personen nachvollziehen zu können. Insbesondere interessieren dabei die Fragen wo sich Personen aufhalten, wer diese Personen sind, mit wem oder was sie sich beschäftigen, worauf sie

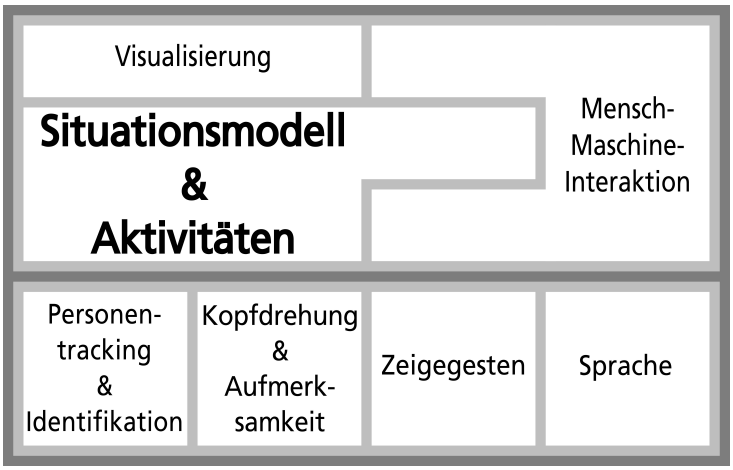


Abb. 5.1.: Schichtendarstellung der Informationsverarbeitung im Smart-Room. Quelle: [IS10].

ihre jeweilige Aufmerksamkeit richten und wohin sie zeigen und sich orientieren. Ein auf dieser perzeptiven Ebene aufsetzendes Situationsmodell greift die Datenströme aller jeweiligen Komponenten ab und fusioniert sie in einem kohärenten Benutzermodell. Zu jedem Zeitpunkt soll damit eine eindeutige Beschreibung des Aktionskontexts aller anwesenden Personen gegeben werden können, was zum einen eine automatische Protokollierung des Geschehens aber auch die Beantwortung kommunikativer Problemstellungen liefern können soll. Indem sich der Mensch in einem solch geschlossenen und einheitlichen System aufhält, besteht die Möglichkeit mit vorhandener Aktorik auf gezielte Situationen proaktiv einzugehen und Information dort einzublenden wo die jeweilige Aufmerksamkeit hingerichtet wird. Durch die unterstützte Erkennung des Handlungskontexts miteinander kooperierender Kollegen, kann deren Gruppentätigkeit als solche erkannt werden und von den übrigen Geschehnissen im Raum anderer Mitarbeiter abgegrenzt werden. Zur Bedienung systematischer Prozesse kann eine solche Umgebung entweder proaktiv oder durch eine dem Menschen vertraute Sprache- und Gesteninteraktion auf die Bedürfnisse der Benutzer reagieren. Damit löst sich das Bedienkonzept von einer starren Anwendung mit vordefinierten Abläufen auf einzelnen Arbeitsplatzrechnern und lässt den Benutzer in einer immersiven Umgebung agieren, in der ein vornehmlich unsichtbares Computersystem omnipräsent beobachtet und unterstützt und durch unterschiedliche Aktorik mit dem Benutzer auf seine Ansprüche bezogen kommuniziert.

5.1.1. Sensorausstattung

Der Vorteil einer geschlossenen Umgebung - wie sie ein sogenannter Smart-Room darstellt - ist die konstruktive Anbringung mehrfacher Sensorik um das Innenvolumen aus unterschiedlichen Blickwinkeln beobachten zu können. Aufgrund der gewünschten Bewegungsfreiheit, können vereinzelte Ansichten mit Verdeckungen konfrontiert werden, die eine konsistente Beobachtung und Modellierung erschweren. Durch den Einsatz vermehrter Sensorik aus komplemen-

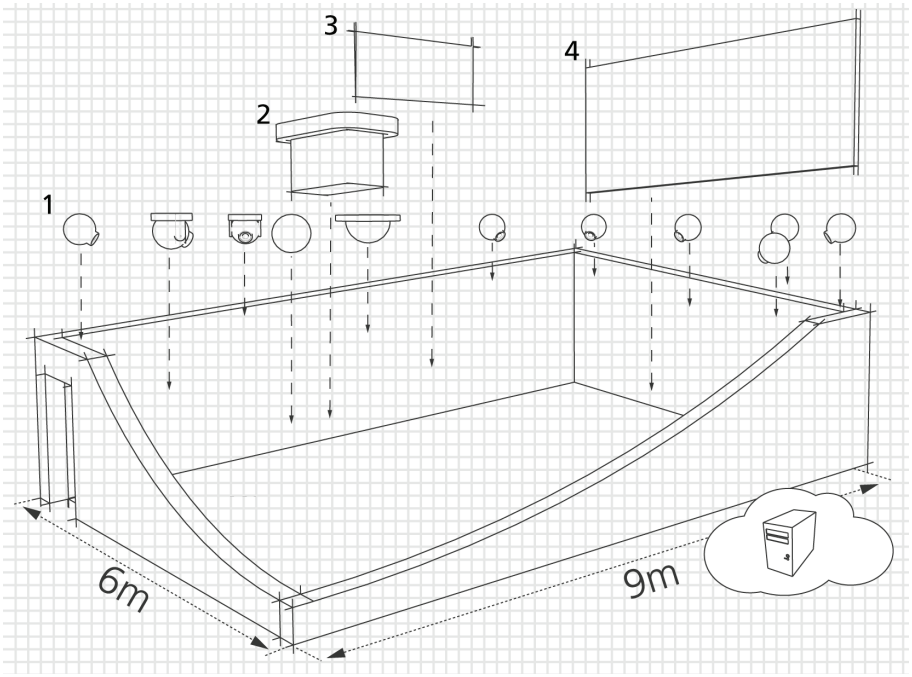


Abb. 5.2.: Sensor- und Aktorausstattung des Smart-Rooms am Fraunhofer IOSB: (1) Kameras mit Fischaugenobjektive in der Decke, Kameras mit fester Brennweite unterhalb der Decke sowie Seitenkameras mit verstellbarer Brennweite beobachten den Raum aus unterschiedlichen Blickwinkeln. Als Aktorik wird bislang ein Tischdisplay (2), ein LC-Display (3) und eine großflächige Videowand (4) eingesetzt.

tären Blickwinkeln, kann in solchen Fällen der Verdeckung vorgebeugt und eine umfassende Erfassung der Szene gewährleistet werden. Der Raum ist hierzu mit einer Vielzahl an Kameras ausgestattet. Deckenkameras bieten eine vollständig verdeckungsfreie Ansicht, die allerdings durch den Einsatz von Fischaugenobjektiven stark radial verzerrt erscheint und darüberhinaus mit dem Blick von oben keine Frontal- oder Profilaufnahmen der Personen bieten. Als Gegenstück hierzu sind Kameras unterhalb der Raumdecke angebracht, die von oben herab die Szene beobachten und damit einsichtigere Aufnahmen garantieren aber vereinzelt Verdeckungen ausgesetzt sein können. Für eine konsistente Erfassung des gesamten Innenvolumens sind die Kameras hierzu mit einer festen Brennweite ausgestattet, die jedoch dazu führt, dass die große Distanz ihrer Positionen zum Geschehen nur zu verrauschten Beobachtungen mit niedriger Auflösung führt. Diesem Problem wurde in dieser Arbeit entgegengetreten, in dem durch die gleichzeitige Nutzung mehrerer Ansichten eine fusionierte und stabilisierte Schätzung der Kopfdrehung erreicht werden sollte. Für eine detaillierte Merkmalsbeobachtung eignen sich solche Ansichten jedoch nicht, weswegen für vereinzelte Aufgaben, wie zum Beispiel einer vollautomatischen Gesichtsidentifikation, Aktivkameras an den Seitenwänden angebracht sind, die auf Augenhöhe eine dreh- und schwenkbare Aufnahme mit verstellbarer Brennweite ermöglichen.

5.1.2. Aktorausstattung

Im Raum werden verschiedenartige Display eingesetzt um Information verteilt darzustellen und aufgrund der unterschiedlichen Displaygrößen verschiedenartige Interaktionstechniken zu evaluieren. Am dominantesten wirkt dabei eine großflächige Videowand, die insbesondere den Nutzen von Zeigegestenerkennung unterstreicht, weil bei einer reinen Touchscreenunterstützung mit den jeweiligen Bildmaßen Randbereiche für den Menschen unerreichbar bleiben.
Neben der Videowand wird ein Tischdisplay - ein sog. *digitaler Lagetisch* [BMT08] - eingesetzt, der Fingergesten durch eine angebrachte Infrarotüberwachung erfasst und damit Information auf dem Tischdisplay gezielt und feingranularer editierbar werden lässt als an der Videowand per Zeigegesten.
Weitere LC-Displays im Raum sollen für eine konsistente aber bildschirmübergreifende Darstellung genutzt werden, so dass manipulativ Oberflächenelemente auf andere Bildschirme verschoben werden und Informationen den Benutzern folgen können.

5.2. Perzeptive Verarbeitungskette

Im folgenden soll die Wahrnehmung des Raums anhand der eingesetzten perzeptiven Verarbeitungskette beschrieben werden. Jede Komponente greift hierzu autonom die für sie notwendigen Kameraströme im Netzwerk ab. Die durch die Komponente berechneten Hypothesen werden im Anschluss zurück in das Netzwerk eingespeist, wo weitere Perzeptionsmodule oder das Situationsmodell diese zur Vervollständigung und Aktualisierung des Benutzermodells abonnieren können. Zur flexiblen Verteilung der Datenströme wurde eine projekteigene Peer-to-Peer-Middleware implementiert, die flexibel auf Abonnements reagieren kann und die Netzwerklast zu minimieren versucht. Über sie werden sowohl Sensorströme, als auch Perzeptionshypothesen und Interaktionsereignisse übertragen.

5.2.1. Personentracking

Um Menschen bei ihrer Arbeit unterstützen zu können, ist es notwendig zu erkennen ob sie überhaupt im Raum anwesend sind: Sobald eine Person den Raum betritt, wird sie erfasst und im nachfolgenden Geschehen verfolgt und von anderen Menschen im Raum unterschieden [vdCVS10]. Für eine individuelle und personalisierte Interaktion, muss darüberhinaus aber auch die Identität zugeordnet werden können.

Detektion von Personenmerkmalen

Mit einer Detektion von Personenmerkmalen wird Aufschluß darüber gewonnen, wo sich Menschen im Raum aufhalten. Die Lokalisationsbestimmung gibt jedoch allein keinen Aufschluß über die zurückliegende Trajektorie oder die Identität. Ebensowenig werden Personen zu diesem

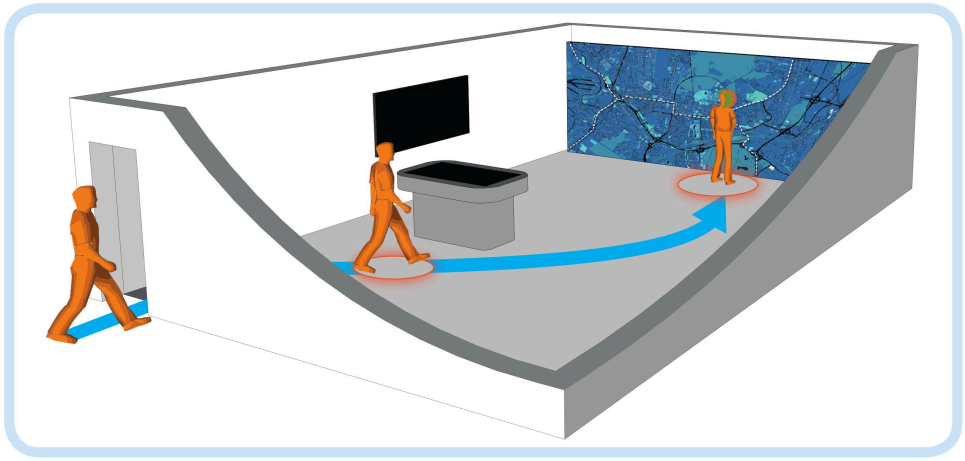


Abb. 5.3.: Schematische Darstellung des Personentrackings: sobald eine Person den Raum betritt, wird sie erfasst und identifiziert. Ihre nachfolgende Trajektorie wird konsequent modelliert und von den übrigen Benutzern im Raum differenziert. Durch das Wissen wo sich eine Person aufhält, kann Information gezielt in ihrer Nähe eingeblendet werden; zum Beispiel kann bei einem Nähern an die großflächige Videowand personalisierte Information dargestellt werden, die die Person bei ihrer aktuellen Aufgabe unterstützt.

Zeitpunkt bereits unterschieden. Die von den Detektionen erhaltenen Positionen und Körpermaße erkannter Benutzer kann allerdings im weitergehenden Prozess dazu eingesetzt werden den Suchraum weiterer Verarbeitungskomponenten hinreichend einschränken zu können.
Im Rahmen des hier beschriebenen Anwendungsfalls werden zur Detektion der Merkmale die Kameras in den oberen Ecken des Raums eingesetzt: auf Oberkörper eingelernte Detektoren werden dazu benutzt, um Personen vor dem Hintergrund zu erkennen. Die Hypothesen werden im Anschluß probabilistisch in die Trajektorienverfolgung einbezogen.

Gesichtsidentifikation

Die Identifikation der Benutzer dient dazu eine personenbezogene Modellierung und Interaktion bewerkstelligen zu können. Erst durch das Wissen wo sich eine bestimmte Person aufhält, kann auf sie individualisiert eingegangen werden oder Wissen über unterschiedliche Aktivitäten der anwesenden Benutzer akquiriert werden.
Die Gesichtsidentifikation erfolgt vollautomatisch durch Einbezug der Aktivkameras an den Raumwänden: Anhand der erkannten Kopfdrehung der Benutzer kann eine optimale Kameraansicht für frontale Aufnahmen ausgewählt werden [vdCVS10]. Verdeckungen können automatisch berücksichtigt werden, da zu jedem Zeitpunkt die Position aller Personen im Raum bekannt ist und damit auf verdeckte und benachteiligte Ansichten geschlossen werden kann.
Die Aktivkamera fokussiert auf den Vorderkopfbereich einer Person und dektektiert die Gesichtsregion, die im Anschluß zur ansichtsbasierten Identifikation herangezogen wird. Ein wie-

derholtes, iteratives Identifizieren wird in der Trajektorienverfolgung dazu eingesetzt, um die Personenmodelle zu überprüfen und gegebenenfalls korrigieren zu können.

Trajektorienverfolgung

Um ab dem Zeitpunkt des Raumbetretens eine Person in ihrer Bewegung nachverfolgen zu können, werden Detektionen und automatische Identifizierungen in ein Trackingrahmenwerk vereint, dass die Frage nach der Position jeder anwesenden Person zu jedem Zeitpunkt beantwortet. Sobald eine Person den Raum betritt, wird sie detektiert und als *unbekannt* markiert. Im Laufe nachfolgender Beobachtungen und Frontalaufnahmen, wird ihr Modell mit einer Identität ausgestattet und personenbezogene Dienste werden der Person angeheftet, indem sie der Trajektorie an nahgelegene Displays folgen können. Ebenso werden individualisierte Interaktionskonfigurationen, Präferenzen in der grafischen Benutzerobfläche und persönliche Verhaltensmuster ableit- und wiederladbar. Darüberhinaus ist mit der Identifizierung eine automatische Autorisierung möglich, die Information vor unautorisierten Benutzern verbirgt und nicht zugänglich macht.

5.2.2. Kopfdrehung und Aufmerksamkeitszuwendung

Wie bereits in Kapitel 1 beschrieben wurde, führt die Sensoranbringung und Bewegungsfreiheit der Benutzer dazu, dass die Blickrichtung nur durch sekundäre Indikatoren angenähert werden kann. Wie im Rahmen dieser Arbeit gezeigt werden konnte, kann mit der Kopfdrehung in einem solchen Rahmen eine hinreichende Messung des Blickfelds einer Person erfolgen, das darüberhinausgehende Aussagen über deren eigentliche Aufmerksamkeitszuwendung erlaubt. Mit dem Wissen wohin eine Person ihre Aufmerksamkeit richtet, erhält man Informationen über den Zuwendungsfokus und der jeweiligen Aktivität und Tätigkeit. Diese Beobachtung kann schließlich dazu eingesetzt werden, um eine Hypothese über unterstützende Informationsdarstellung zu ermöglichen oder kollaborative Aktivitäten zwischen mehreren Personen zu erkennen.

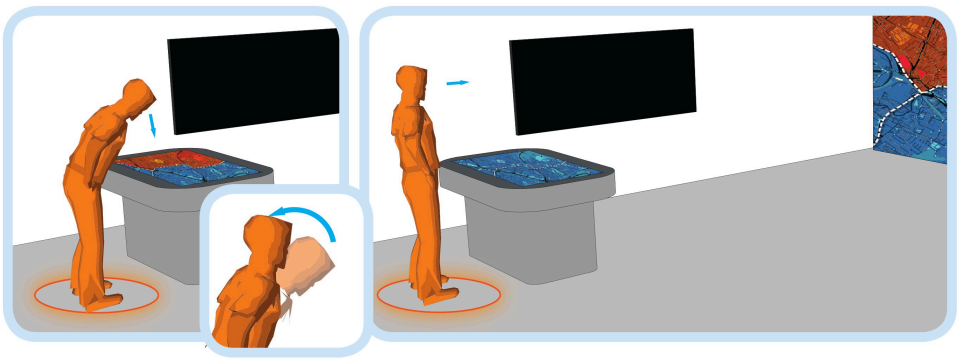


Abb. 5.4.: Schematische Darstellung der Kopfdrehungserkennung in einer aufmerksamen Umgebung.

Als weiteren Einsatz zeigte sich die Kopfdrehung dazu in der Lage die Gesichtsidentifikation während des Personentrackings dahingehend zu unterstützen, dass Frontalansichten zu einzelnen Kameras erkannt und für eine detaillierte Gesichtsbeobachtung ausgenutzt werden können [vdCVS10].

5.2.3. Erkennung von Körperpose und Zeigegesten

Die Detektion einer Person gibt deren Position im Raum wider. Mit der Erkennung der Körperpose können aber erst manipulative Handlungen erkannt und zur Steuerung der Raumanwendung umgesetzt werden. Daneben gibt das Wissen über das artikulare Körpermodell eines Menschen weitergehende Information über dessen Handlung und Kontext.
Zeigegesten können damit als einfacher Ausdruck ausgestreckter Arme interpretiert werden und bieten eine dem Menschen vertraute Möglichkeit sich gezielt auf Bereiche im Raum beziehen, dargestellte Information selektieren und schließlich auch manipulieren zu können. Darüber hinaus gibt die Körperorientierung und -haltung - neben der Kopfdrehung - weitere Indikatoren prinzipieller Aufmerksamkeitszuwendungen. Hierzu wird die Szene aus den verschiedenen Kameraansichten unterhalb der Raumdecke dreidimensional rekonstruiert. Damit kann das Volumen menschlicher Körper extrahiert werden. Im Anschluß wird die Silhouette auf ausgestreckte Gliedmaßen untersucht [SvdCIS09]. Personen wird so die Möglichkeit geschaffen, um zum Beispiel mittels Zeigegesten selektive Bereiche auf der Videowand auszuwählen oder operationelle Werkzeuge anzuwenden - ohne die notwendigen Beihilfe weiterer peripherer Werkzeuge. Tastaturen und Mäuse werden durch die dem Mensch vertraute Art auf Dinge zu zeigen ersetzt. Prinzipiell ist durch die visuelle Erfassung des gesamten Raumvolumens diese Form der Interaktion an jeder Position im Raum möglich. So kann nicht nur vor der Videowand manipulativ eingegriffen werden sondern auch in der gesamten Umgebung auf einzelne Bereiche aufmerksam gemacht werden. In Kombination mit Sprachkommandos werden Zeigegesten damit ein wichtiges Instrument dem Raumsystem Kommandos zu übergeben und die Interaktion der Benutzer untereinander analysieren zu können.

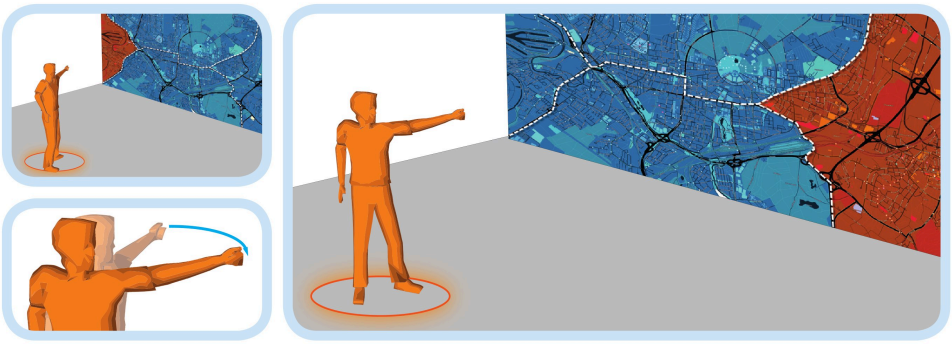


Abb. 5.5.: Schematische Darstellung der Zeigegestenerkennung in einer aufmerksamen Umgebung.

Mit der Möglichkeit dem Raum zeigen zu können worauf man sich bezieht, ohne hierzu weitere Geräte bedienen zu müssen, kann Mensch-Maschine-Interaktion intuitiv und natürlich ausgerichtet und Bedienkomfort und -einfachheit in den Vordergrund gestellt werden.

5.2.4. Situationsmodell

Mit einem auf der perzeptiven Ebene aufsetzenden Situationsmodell werden die Hypothesen der verarbeitenden Komponenten ausgelesen und in einem einheitlichen Benutzermodell zusammengefügt. Damit werden die gesammelten Informationen über Position, Körpermaße, Identität, Trajektorie, Kopfdrehung und Aufmerksamkeitszuwendung, Körpermodell und Gesteninteraktion vereinheitlicht und für alle Personen in der aufmerksamen Umgebung vorgehalten. Über die Historie werden Bewegungen nachvollziehbar und kollaboritave Zusammenarbeit mehrerer Personen nun ableitbar.
Das Situationsmodell setzt Prädikatenlogik ein, um Zusammenhänge zwischen personenbezogenen Beobachtungen und vordefinierten Situationen herzustellen [IS10]. So können Regionen im Raum voneinander abgegrenzt werden, in denen verschiedenartige Handlungen zu beobachten sind. Ebenso zeigt sich das Situationsmodell dafür verantwortlich multimodale Interaktionsereignisse zu korrelieren und darauf entsprechend zu reagieren. Ein solches Beispiel stellen Sprachkommandos dar, die in Kombination mit Zeigegesten zu einer operativen Interaktion an der großflächigen Videowand führen sollen. Die Inkompatibilität der Komponentenausgaben zeigt sich bereits daran, dass Zeigegesten nicht als binäre Aktionen aufgefasst werden können, sondern der Arm zunächst gehoben und anschließend gesenkt wird und die Geste als eine nicht diskret abgrenzbare Gliedmaßenbewegung aufgefasst werden kann. Vielmehr sollte sie erst in einem Kontext der Interaktion als eigentliche Bedienung interpretiert werden. Die darüberhinaus notwendige Zuordnung zu einem zeitversetzten Sprachkommando und die Latenz der Spracherkennung machen bereits deutlich, dass die Modalitätenfusion ein nicht trivialer Prozess ist, der die Notwendigkeit eines konsistenten und über die Historie vollständigen Benutzermodells hervorhebt.

5.3. Konzeptionierung eines Smart-Control-Rooms

Der Smart-Room am Fraunhofer IOSB wird zu Entwicklungszwecken als Vision eines zukunftsorientierten Krisenlagezentrums eingesetzt. Mit den Möglichkeiten einer großflächigen Videowand und weiteren Displays im Raum soll der Vorteil verteilter Informationsdarstellung und intuitiver Interaktionsmodalitäten gezeigt werden. Hierzu wurde eine Anwendung implementiert, die bildschirmübergreifend das Manipulieren von Geodateninformation an einer Lagekarte ermöglicht: Die Videowand zeigt dabei eine große Übersichtskarte eines regionalen Ausschnitts an, in dem Einsätze und Notfälle aufgeführt werden. Der digitale Lagetisch lässt

eine eigene Darstellung des Kartenausschnitt durch Gesten auf dem Tischdisplay verschieben, vergrößern, verkleinern und drehen. Mithilfe weiterer Tablet-PCs kann zusätzliche Information in den Kartenausschnitt eingefügt und manipuliert werden. Die Information steht schließlich im gesamten Raum zur Verfügung.
Die Motivation hinter der Anwendung ist es zu evaluieren, wie die rechnerlose Bedienung einer bildschirmübergreifenden Anwendung umgesetzt werden kann und wie die Infrastruktur dafür ausgelegt werden muss. Die Vernetzung aller Komponenten und die Entwicklung einer bildschirm-, plattform- und betriebssystemübergreifenden Anwendung und Nachrichtendistribution stehen dabei ebenso im Vordergrund, wie die Entwicklung der notwendigen Perzeptionskomponenten, die in Echtzeit entsprechende Hypothesen ausgeben müssen, um eine flüssige Benutzbarkeit sicherzustellen. Neben der eigenentwickelten Middleware kommt so die perzeptive Verarbeitungskette in einem praktischen Einsatz zum Tragen und durch die neuartige Bedienbarkeit stellen sich Fragen bezüglich der Benutzerführung und des Oberflächendesigns, die in Benutzerstudien untersucht und beantwortet werden.
Das Konzept wurde auf der CeBIT im Jahr 2011 der Öffentlichkeit vorgeführt. Der nachfolgende Ereignisablauf beschreibt dabei die vorgeführte Ausstellungsdemonstration. Die eingebundenen Fotos entstammen ebenfalls dem dortigen Auftritt.

5.3.1. Ereignisablauf

Sobald Mitarbeiter den Raum betreten werden diese detektiert und vom System identifiziert. Jeder der Mitarbeiter vor der Videowand erhält einen individuellen Arbeitsbereich, in dem operationelle Werkzeuge abgebildet sind, die er durch eine Zeigegeste auswählen und anwenden kann. Je nach Werkzeug können ihm Informationen über die Position vorhandener Einsatzkräfte, Live-Übertragungen zu Videoströmen abonnierter Überwachungskameras oder dreidimensionale Animationen der ausgewählten Umgebung angezeigt werden. Entfernt sich der Benutzer von der Videowand, verschwindet dort sein Arbeitsbereich. Sobald er wieder vor die Wand tritt oder sich am Tischdisplay aufhält, wird das Arbeitsmenü jeweils dort eingeblendet. Durch das Anhaften des Arbeitsbereichs an die personenbezogene Trajektorie, können darüberhinaus Daten aus einem Kartenausschnitt auf ein anderes Display mitgenommen werden: so können Meldungen oder Informationsdarstellungen die sonst nur auf der Videowand angezeigt würden, zur weiteren Verarbeitung zum Lagetisch mitgenommen werden - oder umgekehrt. Im Fall eines Ereignisses, wird die Information dort eingeblendet wohin die Person ihre Aufmerksamkeit richtet. Arbeitet ein Kollege am Lagetisch und konzentriert er sich das Tischdisplay, wird die Alarmmeldung hierauf dargestellt. Hebt er den Kopf und richtet seinen Blick zur Videowand, verschwindet die Meldung vom Tischdisplay und erscheint auf der Videowand. Umgekehrt wird die Meldung wieder auf dem Tisch angezeigt, sobald er seinen Kopf senkt und den Blick wieder nach unten richtet.

Abb. 5.6.: Momentaufnahmen einer Interaktion an der Videowand im Smart-Control-Room. Eine Person steht vor der Videowand und interagiert mit den operationellen Werkzeugen per Zeigegeste. Im linken Bild wurde ein eine dreidimensionale Darstellung eines Einsatzortes auf der Stadtkarte ausgewählt. Rechts ist das Verankern von Information im persönlichen Arbeitsbereich abgebildet: Mit einer Wischgeste wurde der Arbeitsbereich zuvor rotiert und statt der operationellen Werkzeuge wird der untere Ringbereich, der für persönlichen Speicherplatz benutzt werden kann, sichtbar. Auf der Karte dargestellte Elemente können mittels Zeigegeste ausgewählt und im persönlichen Speicherbereich abgelegt werden. Der Arbeitsbereich und die abgelegten Daten folgen der Person auf ihrer Trajektorie durch den Raum und werden an nächstgelegenen Displays automatisch eingeblendet.

Abb. 5.7.: Verdeutlichung des angehefteten Arbeitsbereichs: im linken Bild ist der Mitarbeiter noch an der Videowand beschäftigt. Weil das System seine Trajektorie verfolgt, wird sein Arbeitsbereich entsprechend an der Videowand angezeigt. Im rechten Bild ist der Mitarbeiter an das Tischdisplay zurückgekehrt (rechts im Bild teilweise sichtbar) - sein Arbeitsbereich wird nun automatisch in der rechten unteren Ecke des Tischdisplays eingeblendet.

Die Unterstützung weiterer Endgeräte, wie zum Beispiel Smartphones, erlaubt den Benutzern Information auf der Videowand abzulegen oder Kartendarstellungen vom Lagetisch mitzunehmen. Um Informationen in das Raumsystem einspeisen zu können, wird auf den Gyrosensor des Smartphones zurückgegriffen, mit dem ein Cursor an der Videowand gesteuert werden kann. An einer beliebigen Position kann schließlich die mitgebrachte Information - wie zum Beispiel

Fotos - kopiert und im System verankert werden. Zum Abgreifen von Kartenmaterial auf das Smartphone, kann dieses darüber hinaus auf den Lagetisch gelegt werden. Die Lage auf dem Tisch wird visuell erkannt. Durch eine Fingergeste kann ein Kartenausschnitt schließlich auf das Telefon übertragen werden.

Abb. 5.8.: Links: Auf einem Smartphone mitgebrachte Daten (drei Bilder im Display des Telefons) können auf die Videowand übertragen werden, indem die Bewegung des Telefons über dessen eingebauten Gyrosensor erkannt und dazu genutzt wird einen Cursor auf der Videowand zu steuern. Mit einfachem Fingerdruck auf die Bilder im Telefondisplay werden diese in das System kopiert und an der ausgewählten Position auf der Videowand verankert. Rechts: Legt man das Telefon auf den Lagetisch, können Karteninformationen auf das Telefon kopiert werden. Eine Fingergeste auf den zu kopierenden Kartenausschnitt auf dem Tisch und parallel auf das Telefon genügen, um die Daten auf das mobile Endgerät zu kopieren.

6. Zusammenfassung

In der vorliegenden Arbeit wurde ein System entworfen, das das Ziel der visuellen Aufmerksamkeit von Menschen erkennt. Mit dem Wissen auf welche Person oder welchen Gegenstand sich jemand visuell bezieht, soll systematisch ein umfassenderes Kontextverständis einer beobachteten Szene erhalten werden. Als Anwendungsfall wurde hierzu eine aufmerksame Umgebung gewählt, in der unrestriktive Besprechungen im Arbeitsalltag beobachtet wurden. Notwendige Sensorik wurde hierfür unaufdringlich im Hintergrund platziert, was die Qualität der Videoaufzeichnungen deutlich beeinträchtigte. Aufgrund der Distanz der angebrachten Kameras zum jeweiligen Geschehen konnte keine ausreichend aufgelöste und kontinuierliche Sicht auf die Augen der anwesenden Personen erwartet werden. Infolgedessen wurde mit dem Erkennen des Blickfelds einer Person ein sekundärer Indikator eingesetzt, der bereits in verwandten Arbeiten einen prinzipiellen Bezug zur beobachteten Aufmerksamkeitszuwendung erlaubte [SFYW99, SYW01b, GC94, LWB00, BO06]. Hierzu wurde ein System entwickelt, dass die Kopfdrehungen der Personen mit mehreren Kameraansichten gleichzeitig visuell erkennt.
Die Beiträge dieser Arbeit lassen sich dabei insbesondere durch die folgenden Punkte zusammenfassen:

- **Berücksichtigung dynamischer Szenenkompositionen**
  Um von der Kopfdrehung auf ein mögliches Aufmerksamkeitsziel zu schließen, ist es notwendig zu wissen, wo sich das Ziel relativ zum Beobachter befindet. Bisherige Verfahren, wie zum Beispiel [SFYW99, SYW01a, BO07a, BO06, OB07, OTYH05, OYTM06], gehen hierbei von einer statischen Zielanordnung aus: in den Aufnahmen sind die Positionen der Ziele bekannt und bleiben unverändert. Im Gegensatz hierzu setzt sich diese Arbeit erstmalig mit nicht vordefinierten Zielanordnungen auseinander: Weder die Anzahl der Personen und Objekte in der Umgebung noch deren Positionen sind während der Besprechungen vorgegeben oder bekannt. Infolgedessen waren bereits bestehende Ansätze nicht anwendbar, was zu einer neuartigen Szenenrepräsentation führte und die Adaptierung der eingesetzten Modelle zur Laufzeit voraussetzte.

- **Berücksichtigung offener Mengen als mögliche Aufmerksamkeitsziele**
  Das Nachvollziehen der visuellen Aufmerksamkeit wurde in verwandten Arbeiten stets auf die Menge der prinzipiellen Aufmerksamkeitsziele, wie den anwesenden Personen, Besprechungstisch und Leinwand während einer Besprechung beschränkt. In dieser Arbeit wurde erstmals die gesamte Umgebung vermessen und jedes Möbelstück und Objekt

im Raum ebenfalls als mögliches Aufmerksamkeitsziel annotiert. Das entworfene System wurde sowohl auf der Menge prinzipieller Ziele evaluiert als auch der offenen Menge gegenübergestellt.

- **Feingranulare Schätzung der Kopfdrehung in einer Mehrkameraumgebung mit Beobachtungen niedriger Auflösung**
  Um eine detaillierte Unterscheidung der Aufmerksamkeitszuwendung zu einzelnen Zielen zu ermöglichen, ist eine hinreichend feine Messung des Blickfelds einer Person notwendig. Um die Bewegungsfreiheit der Personen dabei nicht wesentlich einzuschränken, wurden hierzu mehrere Kameras unterhalb der Raumdecke, in den oberen Ecken des Raums installiert. Die Kameras behielten eine Festbrennweite, damit jederzeit aus jeder Ansicht eine vollständige Übersicht des Raums aufgezeichnet wurde. Damit musste in dieser Arbeit mit zwei Schwierigkeiten umgegangen werden: (1) Durch die uneingeschränkte Bewegungsfreiheit der Personen war in keiner Ansicht eine konsquente Beobachtung des Vorderkopfs garantiert. Vielmehr muss in den übrigen Kameraansichten mit Profil- oder Hinterkopfansichten gerechnet und umgegangen werden. (2) Die distante Sensoranbringung mit fester Brennweite erlaubte bei einer Bildauflösung von nur 640×480 Pixel nur verhältnismäßig kleine Kopfbeobachtungen, die dadurch mit einer nur niedrigen Auflösung vorlagen und wenig Details in der Gesichtstextur offenbarten. Die Problemstellung niedrig-aufgelöster Kopfmotive wurde in der Forschung bislang nur in wenigen Arbeiten angegangen [KBS00, NF96, RR98, WT00, ZPC02]. Aufgrund darin nur vorhandener Einzelkameraansichten lagen Beschränkungen in den Ansätzen insbesondere in der Granularität der Winkeldiskretierung bei der Schätzung (meist 45°) oder in einem beschränkten Winkelwertebereich, der maximal bis ins Profil gedrehte Köpfe voraussetzte.
  Im Rahmen dieser Arbeit wurde erstmalig eine feingranulare Winkelschätzung über den gesamten Winkelwertebereich von 360° horizontal und 180° vertikal voraus- und umgesetzt. Um hierbei mit der Auflösungsproblematik umzugehen, wurden die Schätzungen aller vorhandener Kameraansichten fusioniert und in eine gemeinsame Hypothese überführt. Um darüberhinaus eine Schätzung in Echtzeit sicherstellen zu können, wurden einzelne Komponenten des Systems parallelisiert und damit beschleunigt.

- **Datensätze als Referenz öffentlicher Evaluationen**
  Die Problemstellung der Kopfdrehungsschätzung unter den genannten Bedingungen war bislang in der Forschung unberücksichtigt gewesen. Öffentliche Datensätze für Evaluationen waren infolge dessen nicht vorhanden. Aus diesem Grund wurde ein eigener, dedizierter aufgezeichnet und vollständig annotiert.
  Der Datensatz wurde im Zuge der CLEAR-Evaluationen [SBB+07] veröffentlicht und diente seither als Referenz für weitere Arbeiten, die sich mit dieser Problemstellung aus-

einander gesetzt haben [ZHLH06, CFCP06, PZ07, LB07, BO07b, CFCP07, YZF+08]. Der bereitgestellte Datensatz somit kann als erste Referenz in diesem Forschungsgebiet angesehen werden.

6.1. Diskussion

Zum Schätzen der Kopfdrehung wurde ein Partikelfilter implementiert, dessen Zustandshypothesen die dreidimensionale Position und Größe sowie den horizontalen und vertikalen Drehwinkel eines beobachteten Kopfs umfassen. Die Hypothesen wurden einzelansichtsbezogen durch Gradientenhistogramme und Künstliche Neuronale Netze bewertet und die jeweiligen Bewertungen in einer anschließenden Fusion auf Entscheidungsebene zusammengeführt. Damit konnte erreicht werden, dass das Verfahren unabhängig von der Anzahl vorhandener Kameras angewandt werden kann: Ein Hinzufügen weiterer Ansichten hat so kein notwendiges Neutrainieren der Komponenten zur Folge. In den Evaluationen wurde dieser Vorteil durch die Erkenntnis unterstrichen, dass eine Klassifikation des Drehwinkels auf Einzelkameraansichten deutlich unpräziser ausfiel als unter gleichzeitiger Zuhilfenahme mehrerer komplementärer Aufnahmen. Als Grund kann hierfür zum einen die geringe Motivauflösung abgebildeter Köpfe aufgrund der hohen Distanz zu den Kameras herausgestellt werden, zum anderen die unrestriktiven Aufnahmebedingungen die dazu führen, dass neben frontalen auch Profil- und Hinterkopfansichten der jeweiligen Personen beobachtet werden. Mit der hohen Varianz, die unterschiedlich gedrehte Köpfe darüberhinaus aufgrund verschiedener Frisuren und Gesichtsbekleidungen aufweisen, boten einzelne Kameraansichten nicht genügend Information um eine hinreichend feine Drehwinkelhypothese zu ermöglichen. Nachdem alle jeweiligen Teilkomponenten auf den Einfluss ihrer Parameterbelegung untersucht wurden, wurde das Gesamtsystem evaluiert. Hierbei konnte ein mittlerer Fehler bei der horizontalen Drehwinkelschätzung von $5,3°$ und $8,9°$ bei der vertikalen Schätzung festgehalten werden.
Mit der Entscheidung die Kameraansichten getrennt voneinander zur Schätzung einzusetzen und erst auf Entscheidungsebene zur Fusion einzubringen, kommt jedoch die Notwendigkeit der Parallelisierung zu Tage, weil die Merkmalsberechnung pro Kamerasicht einen deutlichen Zuwachs an Laufzeit voraussetzt. Neuronale Netze spielen hier ihren entscheidenden Vorteil aus, sich aufgrund ihrer Topologie ideal für eine parallele Abarbeitung zu eignen. Ferner stellen sie hinsichtlich der Kopfrotationen und verschiedenen Gesichtstexturen bei der niedrigen Auflösung ihre Generalisierungsfähigkeiten unter Beweis. Durch ihre Sensibilität für eine konsistente Lokalisierung der Bildregion, die schließlich als Eingangssignal zur Klassifikation angelegt wird, erzeugten sie jedoch insbesondere im Rahmen des Partikelfilters ein hohes Maß an Rauschen, weil die vielen verschiedenen Annahmen der Stützstellensamples nicht nur den Drehwinkel unterschiedlich hypothetisierten sondern auch die Position und Größe des zu beobachtenden Kopfs. Mit Gradientenhistogrammen wurde hierzu eine geeignete Repräsentation

einer Kopfsilhouette gefunden, die sich im genannten Rahmen zur Lokalisierungsbewertung zuverlässig unter Beweis stellt. Der notwendige Aufwand der gesamten Merkmalsberechnung stellt jedoch insbesondere für die Echtzeitanforderung den Systementwurf insoweit in Frage, als dass die gleichzeitige Benutzung mehrerer Kameraansichten nur unter dem erhöhten implementierungstechnischen Aufwand einer parallelen Berechnung möglich wird. Gerade die Stärke des Ansatzes, weitere Ansichten einfach einbinden zu können, ohne diese neu eintrainieren zu müssen, macht die ursprüngliche Begründung damit obsolet, weil durch den zusätzlichen Rechenaufwand weitere Ansichten nur mit Kompromissen bezüglich der Echtzeitfähigkeit nutzbar werden. Eine Reduktion des Rechenaufwands kann indes nur erreicht werden, wenn entweder die Merkmalsauswahl geändert oder die Kameraanzahl reduziert werden kann. Beides setzt jedoch voraus, dass Vorderkopfansichten trotz der unscharfen Darstellung hinreichend präzise erfasst und in eine Drehwinkelschätzung überführt werden. Eine geeignetere Merkmalsselektion kann dabei die Rechendauer während der Klassifikation minimieren, wohingegen eine nicht mehr notwendige Verarbeitung verschiedener Blickwinkel auf denselben Kopf die Notwendigkeit reduziert, gleichzeitig auf mehrere Ansichten zurückgreifen zu müssen.

6.1.1. Von der Kopfdrehung zur visuellen Aufmerksamkeit

Mit der feingranularen Schätzung der Kopfdrehung unter den gegebenen Beobachtungsbedingungen konnte eine hinreichende Grundlage geschaffen werden, um im Anschluss auf die Zuwendung der visuellen Aufmerksamkeit einer Person schließen zu können. In dieser Arbeit wurden dafür erstmalig dynamische Szenenanordnungen berücksichtigt. Darin war weder die Anzahl vorhandener Objekte und Personen vorgegeben noch deren Positionierung und Trajektorie. Im Vergleich zu bisherigen Ansätzen in diesem Forschungsfeld stellten sich damit neue Herausforderungen, die in der vorliegenden Arbeit ausgiebig untersucht und beim Systementwurf und dessen Implementierung berücksichtigt wurden. Eine der Schwierigkeiten war hierbei die fehlende Möglichkeit alle Permutationen eventueller Zielanordnungen und -bewegungen im Vorfeld durch Trainingsdaten abdecken zu können. Dadurch wurde eine Adaptierung der Modelle zur Laufzeit notwendig.
Im Gegensatz zu bisherigen Problemstellungen verlangte die Unterstützung bewegter Ziele auch eine neue Repräsentation im Beobachtungsmodell: mit statischen Zielen in einfachen Szenen genügte es die zu erwartenden Kopfdrehungen zu einer Person jeweils durch eine Normalverteilung zu modellieren [SFYW99]. Weil bewegte Ziele nun aber auch zu Verdeckungen anderer Ziele führen, reicht eine einzelne Normalverteilung nicht aus, ohne implizit Rücksicht auf die Größe der Objekte zu nehmen. Die Frage welche Bereiche eines Objekts sichtbar sind muss beantwortet werden können, was im Rahmen dieser Arbeit zum dem Vorschlag führte, alle Objekte und Personen zu voxelisieren und die Szene im Blickfeld der jeweiligen Person, für die die Aufmerksamkeit nachvollzogen werden soll, dreidimensional zu rekonstruieren. Mit der

Voxelisierung können die bisher etablierten Normalverteilungen auf die Voxelebene übertragen werden, während Aufmerksamkeitsziele durch ihre kompakte Beschreibung nun durch ein geometrische Raycastingverfahren auf Sichtbarkeit und Teilverdeckungen geprüft werden können. Dedizierte Kontextbeobachtungen wie die Bewegungsgeschwindigkeit oder Sprachaktivität umstehender Personen wurden ferner als externe Attraktoren verstanden, die eine mögliche Zuwendung der Aufmerksamkeit verursachen können. In das implementierte Rahmenwerk wurde deshalb das von Itti et al. entworfene Konzept der Bayes'schen Überraschung [IB05] einbezogen, das es erlaubt, unerwartete Merkmalsbeobachtungen probabilistisch zu bewerten und diese so als A-Priori-Reize in das Schließen der Aufmerksamkeitszuwendung einzubeziehen.

Das entworfene System kann in Echtzeit für beliebige Personen aufgrund der gemachten Beobachtungen und erkannten Kopfdrehung eine Aussage darüber treffen, wohin diese ihre visuelle Aufmerksamkeit richten. Neben der umstehenden Personen ist bei hinreichender Erfassung aller Gegenstände damit eine prinzipiell offene Menge möglicher Ziele anwendbar. Dies wurde in den Evaluationen berücksichtigt und untersucht. Durch das Fehlen eines öffentlichen Datensatzes für das dedizierte Szenario, wurde im Rahmen der Arbeit ein eigener aufgezeichnet und annotiert. Der Schwerpunkt wurde bei der Datenaufnahme dabei gezielt auf das Einbeziehen unerwarteten Verhaltens und dynamischen Zielanordnungen gelegt. Hierzu nahmen an den Besprechungen unter anderem schauspielende Personen teil, die den Ablauf der Besprechungen einem vorgegebenen Drehbuch folgend steuerten.
Um das Verfahren zu evaluieren, wurden die Systemhypothesen dabei mit menschlichen Annotationen verglichen. Diese geschahen dabei unter denselben Beobachtungsbedingungen, denen während der Evaluationen auch das System ausgesetzt war. Im Einzelnen wurden hierzu die Protokolle drei unterschiedlicher Annotatoren untersucht, die sich manuell und bildbasiert auf den Kamerabeobachtungen für das jeweilige Aufmerksamkeitsziel der aufgezeichneten Personen entscheiden sollten. Die Annotationen wurden hierzu mit in der Statistik etablierten Gütekriterien auf Übereinstimmung geprüft und explizit herausfordernde Mehrdeutigkeiten in den beobachteten Geschehnissen im Einzelnen erkannt. Sowohl die übereinstimmenden als auch die jeweiligen Einzelannotationen wurden daraufhin als Grundlage zur Analyse der automatischen Systemhypothesen angewandt. Dabei konnte eine Korrektklassifikationsrate von $43,8\%$ beobachtet werden, wenn als Menge möglicher Aufmerksamkeitsziele alle in einer Besprechung teilnehmenden Personen, erfassten Gegenstände und das gesamte Mobiliar zugrunde gelegt wurden. Motiviert durch die ausschließliche Evaluationsgrundlage verwandter Arbeiten in diesem Forschungsgebiet, wurde den Ergebnissen hinsichtlich dieser ca. 40 möglichen Objekte und Personen die Evaluation mit einer reduzierten Zielmenge (populärsten Zielen) gegenübergestellt. Mit den teilnehmenden Personen, dem Besprechungstisch sowie der Projektionsleinwand wurde das System dabei auf exklusiv all diejenigen Objekte beschränkt, die während den aufge-

nommenen Besprechungen faktisch Bestandteil der zu beobachtenden Handlungen waren. Im Zuge dessen konnte ein Zuwachs der Korrektklassifikationsrate auf $60,5\%$ beobachtet werden. Betrachtet man die Ergebnisse, stellt sich die Frage wo Ansätze zur Verbesserung des Verfahrens liegen und wie die Genauigkeit weiter erhöht werden kann. In den Experimenten konnte beobachtet werden, dass ein Großteil der Fehlklassifikationen schlicht aus Verwechslungen bei dicht beieinander gelegenen Zielen bestand. So führte der vor der Projektionsleinwand stehende Sprecher während seines Vortrags stets zu Verwechslungen mit der Leinwand selbst, weil bislang keine eindeutige Unterscheidung der beiden Aufmerksamkeitsziele erreicht werden konnte. Weitere Kontextmerkmale wie zum Beispiel Zeigegestenerkennung wären hier von weiterer Hilfe, um Bezüge des Vortragenden auf die Leinwand erkennen zu können und die dadurch beeinflusste Aufmerksamkeit des Publikums wahrzunehmen. Daneben würde das Einbinden von automatischer Spracherkennung den Inhalt des Vortrags und parallel stattfindender Diskussionen zugänglich machen, was bei einem offensichtlichen Gespräch mit dem Publikum die Leinwand heuristisch in den Hintergrund stellen könnte.
Indem Bewegung und Sprachaktivität als Kontextmerkmale einbezogen wurden, wurde generell auch lediglich Rücksicht auf die prinzipiellen Fokusziele während der Besprechungen Wert gelegt: Personen bewegen sich und sprechen, der Tisch wäre zumindest vertikal von den übrigen Zielen unterscheidbar und die Leinwand immer dann eindeutig zu erkennen, wenn sich keine Person vor ihr befindet. Den jedoch ebenfalls annotierten Möbelstücken und Objekten im Raum, die nicht dominanter Bestandteil der Besprechungen waren, konnte mit den beiden Beobachtungen kaum entgegengetreten werden. Ihre Attraktivität von einer Person betrachtet zu werden ergibt sich nicht aus eigener Verhaltensauffälligkeit, sondern allein aus dem Kontext des Handelns um sie herum, zum Beispiel wenn sich eine Person auf das Objekt bezieht, es in seiner Sprachäußerung erwähnt oder es Bestandteil einer Aktivität wird. Neben einer offensichtlichen Verfeinerung der Kopfdrehungserkennung, um fehlerreduzierte Blickfeldmessungen zu erhalten und hiervon ausgehende Verwechslungen ausschließen zu können, erscheint die Hinzunahme weiterer Modalitäten und Kontextbeobachtungen deswegen dringend sinnvoll, um sowohl die Zielmenge vollständiger modellieren als auch Kontextbezüge umfassender erfassen und daraus entstehende Verwechslungen reduzieren zu können.

6.2. Ausblick

Diese Arbeit konzentrierte sich auf die in der Forschung bislang unberücksichtigte Problemstellung die visuelle Aufmerksamkeit von Personen in dynamischen Szenen probabilistisch nachvollziehen zu können. Durch den geschilderten Anwendungsfall einer aufmerksamen Umgebung konnte dabei ein deutlicher, praxisnaher Bezug dargestellt werden. Die Fragestellung konnte so in der Forschung bereits gesteigertes Interesse hervorrufen und führte in nachfolgenden Publikationen verwandter Arbeiten inzwischen zu einer Berücksichtigung [BHO09].

Mit dem darüber hinaus veröffentlichten Datensatz für eine Kopfdrehungserkennung in niedrig aufgelösten Mehrkameraumgebungen konnte ferner bereits eine Referenz für vergleichbare Verfahrensevaluationen etabliert werden [SBB+07].

Die Fragestellung ist mit dieser Arbeit jedoch nicht erschöpfend beantwortet. Im Laufe der Jahre führten kognitive Studien dazu, das Aufmerksamkeitsverhalten des Menschen immer besser verstehen zu lernen. Von den Studien ausgegangene Modelle konnten auf gezielte Anwendungsfälle übertragen werden. Die bereits empirisch nachgewiesene Möglichkeit anhand sekundärer Indikatoren systematisch auf die visuelle Aufmerksamkeit einer Person schließen zu können [SFYW99], erfuhr so einen deutlichen Schub an Interesse. Auf gezielten Anwendungen konnte dabei eine stetig wachsende Steigerung der Korrektklassifikationsrate beobachtet werden.
Der Anwendungsfall einer aufmerksamen Umgebung stellt diese Systeme vor neue Herausforderungen. Dynamische Szenenanordnungen und unrestriktives Verhalten der Personen stellen die Übertragbarkeit bisheriger Modelle dabei noch in Frage. In dieser Arbeit wurden dabei offensichtliche Problemstellungen zusammengefasst und im Einzelnen untersucht. Durch den hohen Praxisbezug der Fragestellung auch in uneingeschränkten Szenen feststellen zu können wohin eine Person ihre visuelle Aufmerksamkeit richtet, konnte mit dieser Arbeit hoffentlich eine Entwicklung angestoßen werden, die sich diesem Anwendungsfall weiter annimmt. Mit dem inzwischen deutlich erkennbaren Fortschritt in der Mensch-Maschine-Kommunikation erscheint es unausweichlich ein umfassenderes Kontextverständnis menschlicher Handlungen zu erhalten, um Schnittstellen intuitiver gestalten zu können. Demzufolge soll mit den entworfenen Komponenten auch ein Ausblick gegeben werden, wie eine Weiterentwicklung stattfinden kann.

Die hauptsächliche Schwierigkeit in diesem Anwendungsfall ist die hohe Varianz der Kopferscheinungen und die niedrige Auflösung der jeweiligen Ausschnitte. Durch stetig verbesserte Sensorik kann hier sicherlich ein deutlicher Leistungsschub bei besserer Motiverfassung erwartet werden. Die geforderte Bewegungsfreiheit der Personen in der aufmerksamen Umgebung setzt jedoch auch weiterhin einen Einsatz mehrerer Kameras voraus. so dass in jedem Zeitschritt zumindest eine Vorderkopfansicht - oder zumindest Profilansicht - verfügbar ist. Im optimalen Fall kann mit einer erhöhten Auflösung oder adäquateren Merkmalsbeschreibung die Berechnung dabei auf eine Ansicht beschränkt werden, womit die Echtzeitfähigkeit drastisch erhöht werden könnte und eine Parallelisierung nicht mehr notwendig erschiene. Besteht eine solche Möglichkeit allerdings nicht bleibt die Herausforderung, dass Hinterkopfansichten ein robustes System voraussetzen, das mit der hohen Varianz der Motiverscheinungen umgehen kann. Im Rahmen dieser Arbeit wurden alle Ansichten gleichzeitig eingesetzt um mit dem Zuwachs an Information durch komplementäre Ansichten eine robuste Schätzung zu erhalten. Durch die parallele Abarbeitung auf einer dedizierten Grafikkartenhardware konnte der dafür notwendi-

ge Rechenaufwand in Echtzeit abgearbeitet werden - ein Trend, der sicherlich in Zukunft auch weiterhin verfolgt werden wird. Mit dem Ansatz ist jedoch prinzipiell nur eine Person modellierbar: Zur Unterstützung weiterer Personen müssen die Ressourcen entsprechend erweitert werden - ein unverhältnismäßiger Aufwand, der den momentanen Ansatz für einen praktischen Nutzen in einer tatsächlichen aufmerksamen Umgebung in Frage stellt.

Entsprechend interessant stellt sich die Frage, welche Einsichten in das kognitive Verhalten des Menschen in naher Zukunft weitere Hinweise auf seine Aufmerksamkeitszuwendungen geben werden - insbesondere unter Berücksichtigung der hier zugrunde gelegten Szenendynamik. Die Entwicklung kontextbezogene Merkmale umfassender zu beobachten und einzubeziehen ist bereits seit geraumer Zeit erkennbar, um menschliches Verhalten systematisch nachvollziehen zu können. Das wird sich auch auf den hiesigen Anwendungsfall einer aufmerksamen Umgebung übertragen. Mit dem Beobachten der Bewegungsgeschwindigkeit und Sprachaktivität von Personen wurde in der vorliegenden Arbeit auch bereits ein erster Versuch unternommen dies anzuwenden. In Experimenten wurde dabei deutlich, dass dies natürlich nur personenbezogen geschehen und bei sachbezogenen Aufmerksamkeitszuwendungen keinen Zugewinn an Information bedeuten kann. So blieben in Situationen Mehrdeutigkeiten bestehen, in denen auf Gegenstände gezeigt oder diese in den jeweiligen Aktivitäten eingesetzt oder benutzt wurden. Dies zu erkennen erfordert mehrere Dinge: Zum einen müssen nicht nur Personenbewegungen sondern auch Objekte vollständig erfasst und automatisch verfolgt werden - nur dann können Gegenstände die hochgehoben und anderen gezeigt auch automatisch als Aufmerksamkeitsziele erkannt werden. Zum anderen müssen weitere perzeptive Komponenten eingesetzt werden, um ein vollständigeres Modell der Personen vorzufinden: Der in Kapitel 5 geschilderte Anwendungsfall zeigte dabei bereits die Möglichkeiten einer rein operationellen Zeigegeste zur Steuerung einer Anwendung. Ein Einbeziehen von erkannten Gesten und Körperposen zum Erkennen der Regionen auf die während Besprechungen explizit hingewiesen wird und artikularer Aktivitäten die eindeutige Hinweise auf das Aufmerksamkeitsziel bieten könnten mindestens Mehrdeutigkeiten auflösen mit denen bislang nicht umgegangen werden konnte. Der gleichzeitige Einsatz eines Situationsmodells könnte ferner heuristische oder eingelernte A-Priori-Bezüge zu einzelnen Zielen beeinflussen, zum Beispiel dass in Vorträgen die Leinwand wahrscheinlicher als Aufmerksamkeisziel berücksichtigt werden muss als in Gruppenbesprechungen an einem Tisch.
Die Erweiterung der perzeptiven Verarbeitungskette um weitere Komponenten, deren Hypothesen zu einem umfassenderen Verständnis einer Person und derer Handlungen in einem Kontext führen können, stellt allerdings hohe Anforderungen an eine echtzeitfähige Implementierung. Wie auch für die Kopfdrehung wird dieser Punkt insbesondere deutlich, wenn mehrere Personen in einer solchen aufmerksamen Umgebung gleichzeitig erfasst und modelliert werden sollen. Die dazu jedoch notwendige Infrastruktur um aqäquate Rechenleistung zu verteilen, wirft auch

Fragen nach der Vernetzung, des Datenaustauschs und weiterführenden Fusionsmechanismen - gerade auch im Hinblick auf multimodale Sensordaten - auf. Mit der Zunahme an Rechenleistung die momentan allein durch die parallelisierbaren Fähigkeiten auf modernen Grafikkarten entstehen, darf gespannt in die Zukunft geblickt werden, um die weitere Entwicklung in diesem Gebiet zu verfolgen.

A. Perspektivische Projektion und Szenenrekonstruktion

A.1. Das Lochkameramodell

Kameras stehen an unterster Stelle der Verarbeitungskette und stellen damit diejenige Quelle dar, die ein System mit Eingabemerkmalen versorgt. Im Rahmen dieser Arbeit sind Kameras an unterschiedlichen Stellen im Raum angebracht, um diesen aus verschiedenen Blickwinkeln komplett zu erfassen - auch redundant. Während eine einzelne Kamera dabei eine dreidimensionale Szene auf ein zweidimensionales Bild abbildet, erlaubt der Einsatz unterschiedlicher Ansichten das Rekonstruieren der dreidimensionalen Situation aufgrund verschiedener Blickwinkel darauf. Als wichtigstes Glied in der Kette soll daher im Folgenden die Funktionsweise dieser Sensorik erläutert und darauf folgend auf den Vorteil mehrkamerabehafteter Umgebungen eingegangen werden. Die Ausführungen orientieren sich dabei im wesentlichen an [HZ03] und [CF09].

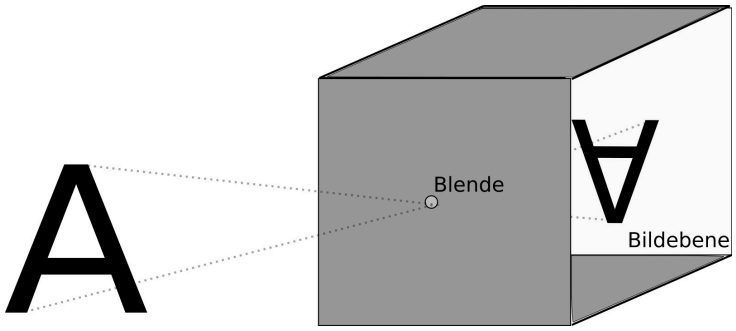


Abb. A.1.: Schematisches Prinzip des Lochkameramodells.

Kameras dienen dazu, eine dreidimensionale Szene persistent auf einer zweidimensionalen Bildebene abzubilden. Das einfachste Verfahren anhand dessen bis heute die Grundzüge bildgebender Sensorik beschrieben wird, basiert dabei auf der *Lochkamera*. Wie in Abbildung A.1 dargestellt wird, besteht diese aus einer verdunkelten Kammer, in die durch eine kleine Öffnung, der sogenannten Blende, an der Vorderseite Lichtstrahlen eindringen. Diese wurden dabei von Objekten vor der Kamera reflektiert, dringen durch das Loch in die Kammer ein und treffen schließlich auf der gegenüberliegenden Seite, auf der Bildebene, auf. Je kleiner dabei die Lochöffnung ist, desto schärfer wird das abgebildete Ergebnis. Das hängt damit zusammen, dass jeder Punkt eines beobachteten Objekts als Punktlichtquelle aufgefasst werden kann. Abbildung A.2 zeigt hierzu wie so ein Lichtkegel, vom Objektpunkt ausgehend, durch die Kammeröffnung eindringt und die Bildebene schneidet. Das Ergebnis sind kleine Kegelschnitte pro

Objektpunkt, sogenannte *Zerstreuungskreise*, die sich mehr oder weniger stark überlappen, je nachdem wie groß die Kammeröffnung davor ist und wie weit die Objektpunkte von der Öffnung entfernt sind. Je kleiner der Lochdurchmesser und je weiter der Objektpunkt von der Öffnung entfernt ist, desto kleiner sind die Zerstreuungskreise. Es ist dabei leicht nachvollziehbar, dass eine nahe und detaillierte Aufnahme eines Objekts eine kleine Lochöffnung benötigt, eine weit entfernte Landschaftsaufnahme dagegen auch mit großen Blenden scharf erscheint. Dabei lässt sich feststellen, dass das Abbild umso schärfer erscheint, je kleiner die Blende eingestellt wird. Desto lichtschwacher wird allerdings auch die Abbildung. Um sich dem zu behelfen, werden optische Sammellinsen in die Blende eingesetzt. Diese bündeln parallele, eintreffende Lichtstrahlen und fokussieren sie in einer für die jeweilige Linse charakteristischen Entfernung, der sogenannten *Brennweite*. In der Literatur wird diese häufig mit f bezeichnet und beschreibt jene Entfernung hinter der Linse, in der die Abbildung scharf erscheint.

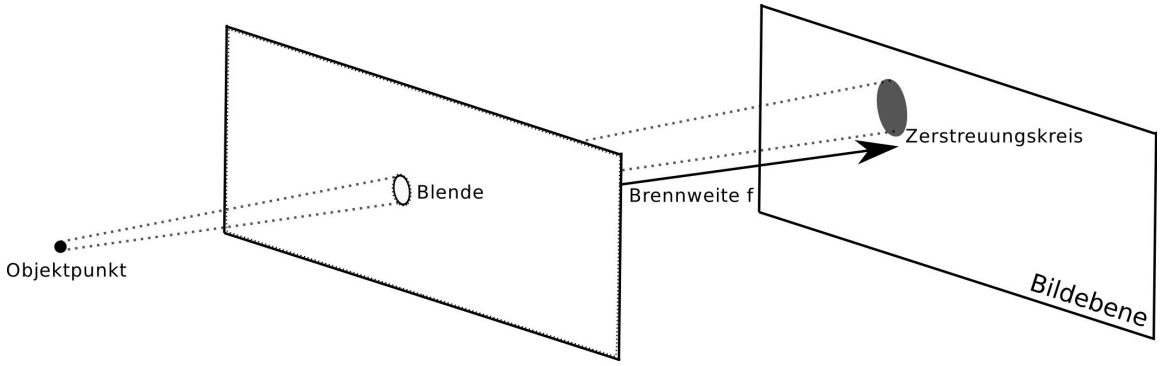


Abb. A.2.: Schematische Entstehung der Zerstreuungskreise auf der Bildebene. Ein Objektpunkt wird abgebildet, indem sein reflektierter Lichtkegel durch die Blende mit der Bildebene schneidet.

A.1.1. Perspektivische Projektion

Unter Berücksichtigung der Brennweite f, lässt sich für einen abzubildenden Objektpunkt $\mathbf{v} = (v_x, v_y, v_z)^T \in \mathbb{R}^3$ berechnen, auf welchen zweidimensionalen Punkt $\mathbf{p} = (p_x, p_y) \in \mathbb{R}^2$ der Bildebene er projiziert wird. Wie Abbildung A.3 zeigt, gilt durch Betrachtung ähnlicher Dreiecke folgender Sachverhalt:

$$p_x = f\frac{v_x}{v_z}, p_y = f\frac{v_y}{v_z} \tag{A.1}$$

Da im Moment keine weiteren Eigenschaften der Linse zu berücksichtigen sind, spricht man hier von einer *idealen Projektion*. Tatsächlich treten in der Praxis jedoch weitere Abbildungseinflüsse auf, auf die im Folgenden deshalb näher eingegangen werden soll. Für den Augenblick soll jedoch zunächst die einer idealen Projektion entsprechende *Projektionsmatrix* erläutert werden.

Homogene Koordinaten haben den Vorteil, dass sich aufeinanderfolgende Transformationen in einer einzigen, gemeinsamen Transformationsmatrix ausdrücken lassen. Um diesen Vorteil zu

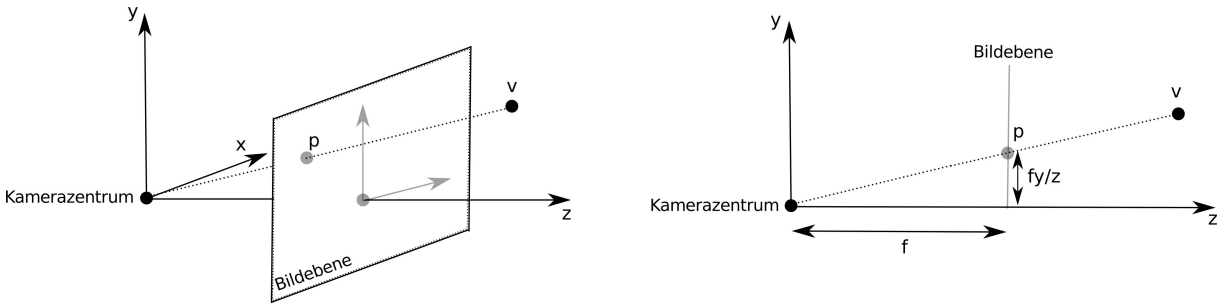


Abb. A.3.: Perspektivische Projektion eines Objektpunkts **v** auf seinen auf der Bildebene liegenden Repräsentanten **p**. Quelle: [HZ03].

nutzen, werden perspektivische Abbildungen deshalb häufig in dieser Form angegeben. Für Gleichung A.1 kann man deshalb äquivalent schreiben:

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \mapsto \begin{pmatrix} fv_x \\ fv_y \\ v_z \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \\ 1 \end{pmatrix} = \mathbf{Pv} \tag{A.2}$$

In diesem Zusammenhang wird die 3×4 Matrix **P** auch als *kanonische Projektionsmatrix* bezeichnet.

Bisher wird angenommen, dass der Nullpunkt des Koordinatensystems der Bildebene mittig angeordnet ist, so dass er genau im Schnittpunkt der orthogonalen Hauptachse durch die Blende liegt. Tatsächlich liegt er aus unterschiedlichen Gründen aber verschoben. Ein solcher Grund liegt in der digitalen Verarbeitung, in der die Bildebene durch ganzzahlige Pixelkoordinaten diskretisiert wird. Durch eine Verschiebung der Ursprungsachsen in eine der vier Bildecken, wird eine konsistente Koordinatenangabe der Pixelelemente erreicht. Für Gleichung A.1 gilt damit, dass eine notwendige Verschiebung t_x und t_y entlang der jeweiligen Achsen miteinbezogen werden muss. Aus $p_x = f\frac{v_x}{v_z} + t_x$, $p_y = f\frac{v_y}{v_z} + t_y$ lässt sich die kanonische Projektionsmatrix aus Gleichung A.2 erweitern zu:

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \mapsto \begin{pmatrix} fv_x + v_z t_x \\ fv_y + v_z t_y \\ v_z \end{pmatrix} = \begin{pmatrix} f & 0 & t_x & 0 \\ 0 & f & t_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \\ 1 \end{pmatrix} = \mathbf{Kv} \tag{A.3}$$

A.1.2. Kalibrierung

Matrix **K** aus Gleichung A.3 wird im Allgemeinen als die *Kalibrierungsmatrix* der Kamera bezeichnet. Sie bildet die Grundlage für das Miteinbeziehen weiterer Parameter, die Einfluss auf die Abbildung des Objektraums auf die Bildebene nehmen. Im Gesamten unterscheidet man

dabei die sogenannten extrinsischen Werte - jene die sich auf die äußerliche Geometrie der Kamera beziehen - von den intrinsischen. Letztere beziehen sich insbesondere auf die internen Abbildungseigenschaften der Kamera, des Objektivs, usw. Für die automatische Bestimmung dieser Parameter, haben sich im Laufe der Zeit unterschiedliche Kalibrierungsverfahren etabliert. Besonders hervorzuheben ist dabei das Verfahren nach Roger Tsai [Tsa86, Tsa87], das in vielen verfügbaren Softwarepaketen zur Kalibrierung eingesetzt wird und mittlerweile als Standard gilt. Auf einige der darin verdeutlichten Kalibrierungsparameter soll nun im folgenden kurz eingegangen werden.

Extrinsische Kameraparameter

Unter extrinsischen Parametern versteht man jene Werte, die den Zusammenhang zwischen den dreidimensionalen Weltkoordinaten und dem dreidimensionalen Kamerakoordinatensystem beschreiben. Dabei handelt es sich im Wesentlichen um die Orientierung der Kamera beziehungsweise deren Rotation $\mathbf{R}$ und Translation $\mathbf{t}$ bezüglich eines gegebenen Weltkoordinatensystems. Bei der Rotation handelt es sich im Allgemeinen um eine orthonormale 3×3 Transformationsmatrix mit drei Freiheitsgraden - den drei Eulerwinkeln, die die Orientierung des Kamerakoordinatensystems widerspiegeln. Die Translation $\mathbf{t}$ beschreibt dagegen die Verschiebung der Kamera als dreidimensionalen Vektor und weist damit ebenso drei Freiheitsgrade auf. Bezüglich bisher beschriebener Projektion, lässt sich damit eine notwendige Koordinatentransformation angeben, die Objektpunkte aus dem Weltbezug in Kamerakoordinaten der Kamera c überführt:

$$\mathbf{v}^c = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \mathbf{v} \tag{A.4}$$

Eingesetzt in Gleichung A.3 erhält man damit

$$\mathbf{p} = \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{v} \tag{A.5}$$

Matrix $[\mathbf{R}|\mathbf{t}]$ wird dabei häufig *extrinsische Kalibrierungsmatrix* genannt, wohingegen sich $\mathbf{K}$ auf die rein intrinsischen Eigenschaften der Kamera und Abbildung bezieht. Auf jene soll nun im folgenden näher eingegangen werden.

Intrinsische Parameter

Die intrinsischen Eigenschaften einer Kamera definieren die interne Abbildung zwischen dem dreidimensionalen Kamera- und dem zweidimensionalen Bildkoordinatensystem. Diese umfassen im Einzelnen die Brennweite der Linse, Translation der Bildmitte sowie die bisher unerwähnte Pixelskalierung bei einer digitaler Aufnahme, Verzerrungskoeffizienten von der durch das Objektiv hervorgerufenen radialen und tangentialen Verzerrung und schließlich die Koor-

dinaten des Verzerrungszentrums. Die radiale Verzerrung skaliert dabei den Vektor, der vom Verzerrungsmittelpunkt, dem sogenannten Fokus, zum unverzerrten Punkt zeigt. Die tangentiale Verzerrung verschiebt den Punkt entlang der Tangente, die orthogonal auf dem radialen Richtungsvektor liegt. Abbildung A.4 verdeutlicht diesen Sachverhalt.

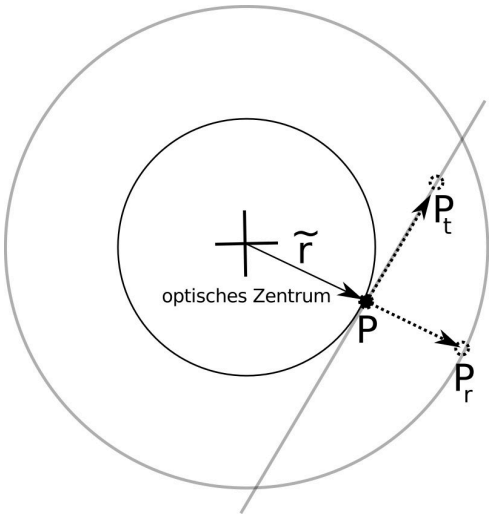


Abb. A.4.: Verzerrung die durch die intrinsischen Parameter hervorgerufen wird. Ein Objektpunkt wird tangential und radial auf der Bildebene verschoben.

Die radialen Verzerrungen stellen den Haupteinfluss auf die interne Abbildung dar. Tangentiale Einflüsse können häufig ignoriert werden, da sie oft nur minimal auftreten. Für einen abgebildeten Punkt $\mathbf{v} \mapsto \mathbf{p}$ gilt daher, dass er nicht an jene Stelle $(\tilde{p_x}, \tilde{p_y})^T$ projiziert wird, die ein ideales Kameramodell vorgeben würde, sondern folgendem Einfluss unterliegt:

$$\begin{pmatrix} p_x \\ p_y \end{pmatrix} = L(r) \begin{pmatrix} \tilde{p_x} \\ \tilde{p_y} \end{pmatrix} \tag{A.6}$$

Der Parameter $\tilde{r}$ bezeichnet dabei den Abstand $\sqrt{\tilde{p_x}^2 + \tilde{p_y}^2}$ vom Verzerrungszentrum zum idealen, unverzerrten Punkt $(\tilde{p_x}, \tilde{p_y})^T$, $L(r)$ dagegen einen Verzerrungsfaktor in Abhängigkeit dazu. Kennt man neben diesen beiden Parametern das Verzerrungszentrum, lässt sich die Verzerrung eines gemessenen Punktes leicht korrigieren. Für die Funktion $L(r)$ gilt dabei allerdings, dass sie nur für $r > 0$ definiert ist und daneben $L(0) = 1$ gilt. Beliebige Verzerrungsfunktionen können durch eine Taylor-Reihe in Form von $L(r) = 1 + \kappa_1 r + \kappa_2 r^2 + \kappa_3 r^3 + \ldots$ hinreichend genau approximiert werden. Neben den Koordinaten des Verzerrungszentrums werden die jeweiligen Koeffizienten $\kappa_1, \kappa_2, \ldots$ dabei während der Kalibrierung der internen Kameraparameter bestimmt. Hierfür sei an dieser Stelle allerdings auf weiterführende Literatur [HS97, Bro71, Bou04, FB86] hingewiesen.

A.1.3. 3D-Rekonstruktion durch Stereogeometrie

Für die Projektionsmatrix $\mathbf{P}$ gilt, dass ihre Spalteneinträge die Basisvektoren des abgebildeten Weltkoordinatensystems sind. Die Matrix ist damit invertierbar. Für jeden beliebigen Bildpunkt $\mathbf{p}$ kann so mit Hilfe der Inversen $\mathbf{P}^{-1}$ eine Gerade definiert werden, deren Punkte alle auf den Bildpunkt abgebildet werden. Diese Gerade nutzt als Ursprung die extrinsische Position der Kamera, so dass sich für die Abbildung $\mathbf{C} + \alpha\mathbf{P}^{-1}\mathbf{v} \mapsto \mathbf{p}$ zusammenfassen lässt.
Für auf $\mathbf{p}$ abgebildete Objektpunkte kann zwar nachvollzogen welcher Geraden sie entsprechen, das Wissen über ihre Tiefe im Weltbezug ging durch die Projektion allerdings verloren.
Geht man davon aus, dass mindestens zwei unterschiedliche Kameraansichten vorliegen, lässt sich die Tiefeninformation des ursprünglichen Objektpunktes $\mathbf{v}$ allerdings durch Triangulation rekonstruieren. Für beide Ansichten müssen hierzu die jeweiligen Projektionsmatrizen $\mathbf{P}$ und $\mathbf{P}'$ sowie die korrespondierenden Bildpunkte $\mathbf{p}$ und $\mathbf{p}'$ des abgebildeten Objektpunkts $\mathbf{v}$ bekannt sein. Dann lassen sich in beiden Fällen entsprechende Projektionsgeraden herleiten und miteinander schneiden. Der Schnittpunkt $\mathbf{C} + \alpha\mathbf{P}^{-1}\mathbf{p} = \mathbf{C}' + \alpha'\mathbf{P}'^{-1}\mathbf{p}'$ entspricht schließlich der dreidimensionalen Ursprungsposition $\mathbf{v}$ in Weltkoordinaten.

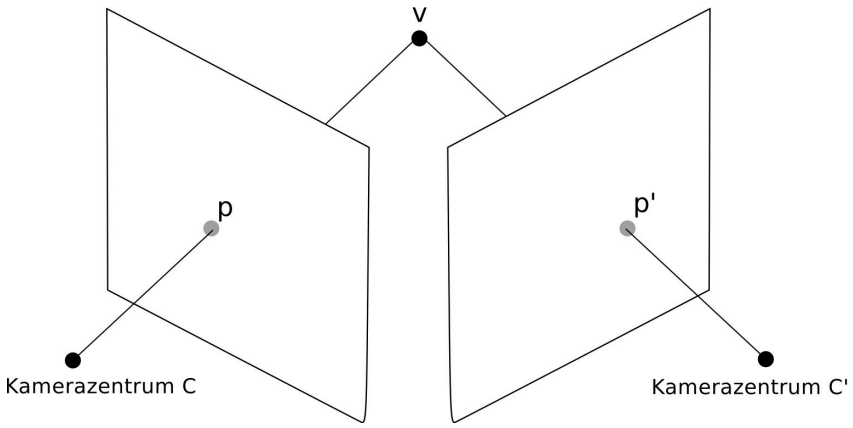


Abb. A.5.: Schematisches Vorgehen der Triangulation. Sichtgeraden die durch jeweils abgebildete Objektpunkte $\mathbf{p}$ und $\mathbf{p}'$ gelegt werden, schneiden im ursprünglichen Objektpunkt $\mathbf{v}$.

In der Praxis treten jedoch häufig Ungenauigkeiten beim Ermitteln der Bildpunkte auf, was zur Folge hat, dass sich die Geraden im dreidimensionalen Objektraum nicht mehr schneiden. Statt des direkten Bestimmens des Schnittpunktes, versucht man diesen deswegen mit Hilfe der Methode der kleinsten Quadrate anzunähern, so dass der minimale Abstand der beiden Geraden gefunden wird. Als Ergebnis erhält man schließlich die beiden Endpunkte des auf beiden Geraden stehenden Lots. Der gesuchte Ursprungspunkt $\mathbf{v}$ wird dann letztendlich als dessen Mittelpunkt interpretiert.
Die Vorteile dieses *Mittelpunktverfahrens* liegen dabei eindeutig in dessen Einfachheit und der schnellen Berechenbarkeit. Für feststehende Kameras mit bekannter Projektionsmatrix lässt sich damit eine in der Regel hinreichend genaue Rekonstruktion der Objektstruktur erreichen.

In anderen Fällen wo kein genaues Wissen über die Projektionsmatrix vorliegt, sondern lediglich die affine Transformation anhand der extrinsischen Parameter oder die Kalibrierungsmatrix anhand der intrinsischen bekannt ist, muss auf projektions- und affininvariante Methoden zurückgegriffen werden. Dafür sei an dieser Stelle aber auf Richard Hartleys Zusammenfassung möglicher Triangulationsmethoden hingewiesen [HS94], welches einen umfassenden Einblick und eine detailreiche Zusammenfassung hierüber hinausgehender Verfahren bietet.

B. Echtzeitimplementierung durch Parallelisierung

Der Vorteil parallelisierbarer Algorithmen liegt darin, voneinander unabhängige Rechenschritte auf unterschiedlichen Datenbereichen gleichzeitig auszuführen. Für einen alltäglichen Computereinsatz eher unerheblich, stellten rechenintensive Grafikanwendungen, und Rendering insbesondere, dabei klassische Domänen dar, in denen die gleichzeitige Ausführung von Prozessor-Instruktionen auf voneinander unabhängigen Speicherbereichen zu enormen Geschwindigkeitsvorteilen führen konnte. So erscheint es nicht überraschend, dass sich Grafikkarten und Koprozessoren in den letzten Jahren diesen Vorteil zunutze machten und sich mit der wachsenden Anzahl vorhandener Prozessorkerne und effizienteren Datenbusanbindungen mit ihrem Schwerpunkt auf Parallelisierung vereinzelter Rechenschritte spezialisierten.

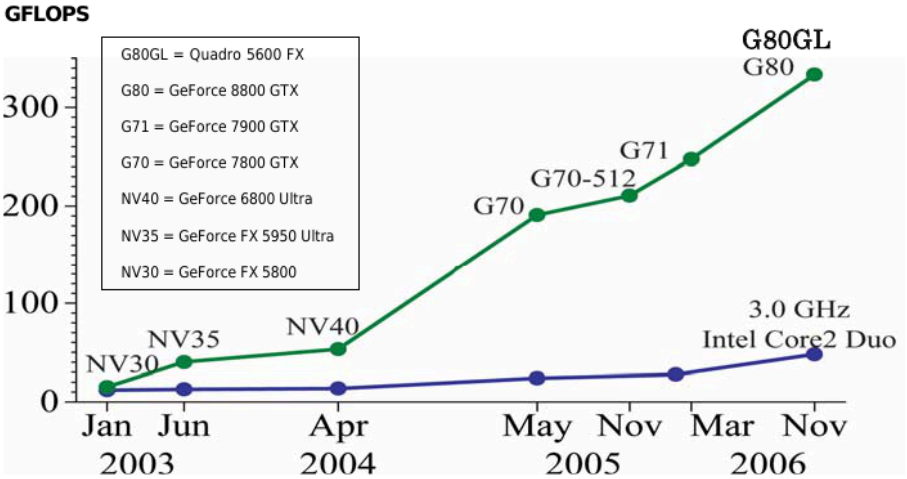


Abb. B.1.: Entwicklung der Anzahl an Gleitkommaoperationen pro Sekunde (engl. Floating-Point Operations per Second / FLOPS) für CPU und GPU. Quelle: [NVI10].

Inzwischen unterstützen viele gängige Grafikkarten, sogenannte *Graphics Processing Units* (GPUs), insbesondere die Datenparallelität. Dabei handelt es sich um das gleichzeitige Ausführung derselben Instruktionen auf unterschiedlichen Datenbereichen, nicht um Kontrollflüsse oder notwendiger Caching-Methoden um die Datenzugriffe zu beschleunigen. Im Rendering handelt es sich dabei um das gleichzeitige Verarbeiten von Eckpunkten und Pixeln, indem Untermengen davon auf parallele Threads abgebildet und bearbeitet werden. Für Filtertechniken auf Bildern oder Videoenkodierung kann das Bild in Unterbereiche segmentiert werden, die wiederum auf parallelen Threads individuell und unabhängig voneinander verarbeitet werden.

Dieses Prinzip lässt sich auf beliebige Anwendungsbereiche übertragen, in denen datenparallele Algorithmen zum Einsatz kommen können. Durch die Unabhängigkeit der Stützstellen voneinander, während der Approximation der Beobachtungswahrscheinlichkeiten in Kapitel 3.1.4, liegt hier eine solche Möglichkeit der parallelen Abarbeitung vor. Im diesem Kapitel soll daher ein Einblick über die implementierte Parallelisierung der Algorithmen und die dadurch erreichbare Geschwindigkeitssteigerung beschrieben werden.

B.1. NVIDIA CUDA

Die *Compute Unified Device Architecture*, im Akronym CUDA genannt, bezeichnet eine von der NVIDIA Corporation entwickelte Hard- und Software Architektur, die eine eigentliche Grafikkoprozessor-API abstrahiert und Zugriffe auf den Grafikspeicher generalisiert [NVI10]. Wie in Abbildung B.2 dargestellt, setzt CUDA dabei auf dem Hardware-Treiber auf und bietet, neben einer API und der dazugehörigen Laufzeitumgebung, Standardbibliotheken für allgemeine Funktionen. Für eine niedrige Lernkurve, ist die API in der Programmiersprache C gehalten Das CUDA-Modell fasst dabei datenparallele Berechnungen in sogenannte *Kernel* zusammen, die in Form von Funktionen implementiert, kompiliert und auf die Grafikhardware zur Laufzeit hochgeladen werden. Individuelle Threads führen diesen Kernel dann zur Laufzeit gleichzeitig

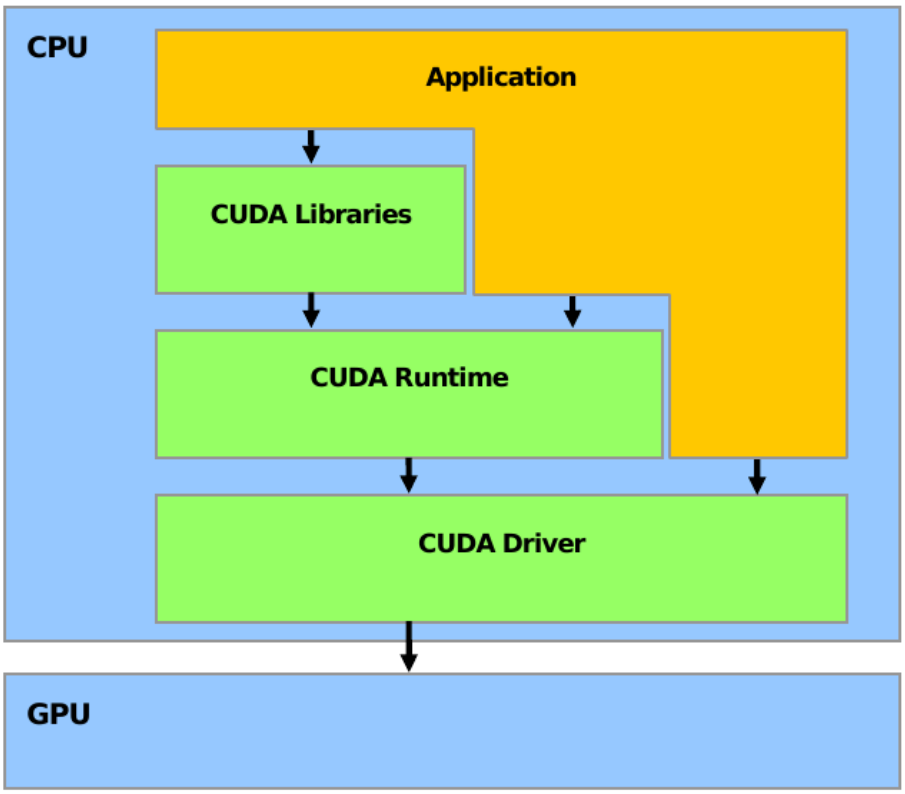


Abb. B.2.: Das CUDA Schichtenmodell. Quelle: [NVI10].

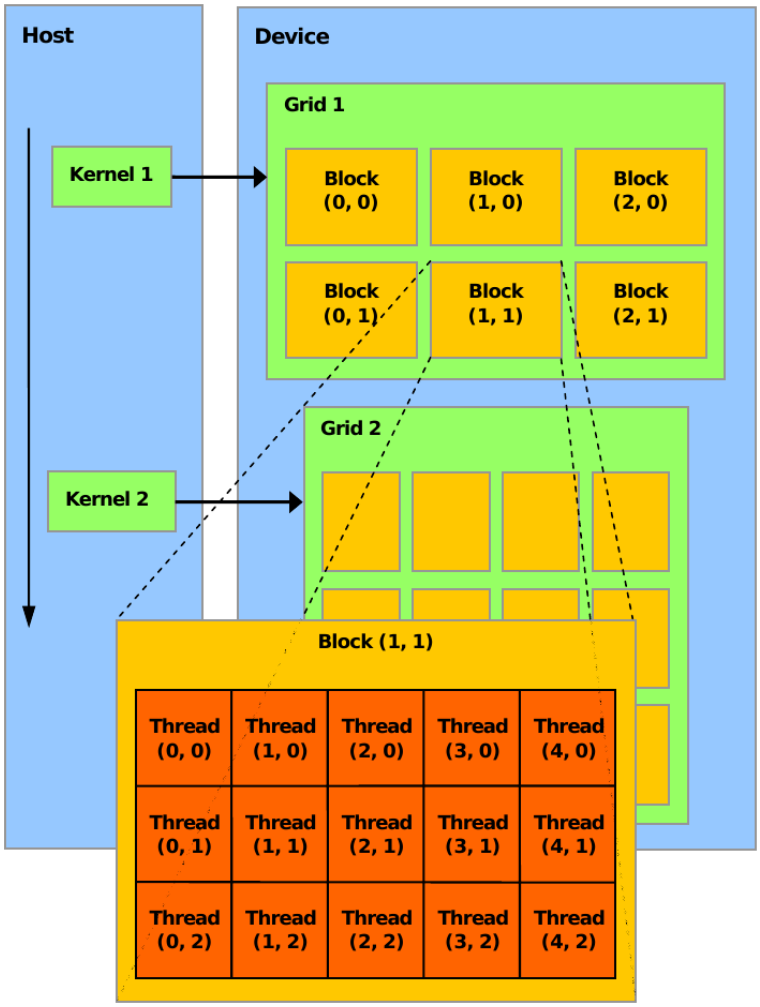


Abb. B.3.: Threadblöcke und deren Anordnung im Blockgitter. Quelle: [NVI10].

aus. In Blöcken organisiert können Threads desselben Blocks auf einen geteilten Datenbereich zugreifen und Daten austauschen sowie synchronisiert werden, indem an vereinzelten Stellen im Programmfluss des Kernels Stellen markiert werden, ab diesen alle Threads gleichzeitig angekommen und bereit für die Fortführung sein müssen. Jeder Thread wird durch einen eindeutigen Nummernbezeichner innerhalb seines Blocks adressiert. Zur Übertragbarkeit komplexer Berechnungen in Abhängigkeit der jeweiligen Threadnummerierung, kann die Adressierung auch in Form eines zwei- oder dreidimensionalen Arrays genutzt werden, wobei die entsprechende Elementbezeichnung intern auf eine eindimensionale Threadidentifikation umgerechnet wird. Zur flexiblen Unterstützung unterschiedlicher Grafikhardware, so dass der Kernel nicht neu kompiliert werden muss um auf anderen Karten ausgeführt zu werden, werden Threadblöcke unabhängig voneinander gehandhabt und bieten infolgedessen keine Möglichkeiten der blockübergreifenden Kommunikation darin beinhaltender Threads.

Auf leistungsstarken Karten können somit mehrere Threadblöcke parallel ausgeführt werden, wohingegen Grafikhardware im unteren Consumerbereich lediglich die sequentielle Ausfüh-

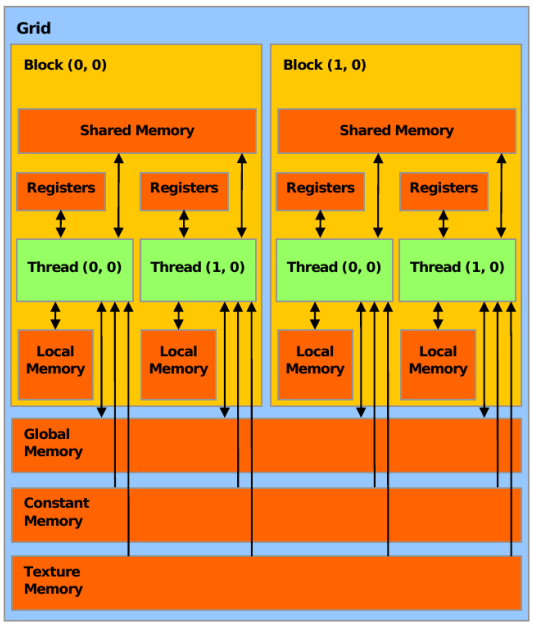


Abb. B.4.: Das CUDA Speichermodell. Quelle: [NVI10].

rung der jeweiligen Blöcke erlaubt. Die Organisation der Threadblöcke erfolgt dabei in einem sogenannten *Blockgitter*, einem zweidimensionalen Array beliebiger Größe, dessen Elemente die entsprechenden Threadblöcke adressieren.
Abbildungen B.4 und B.5 stellen den geregelten Speicherzugriff in CUDA und dessen Hardwareimplementierung pro Grafikkarte dar. Jeder verbaute Grafikprozessor bietet dabei eine feste Anzahl an Registern der Größe 32 Bit. Darüber hinaus teilen sich alle Prozessoren:

- Einen les- und beschreibbaren, parallelen Datencache der auf die einzelnen Threadblöcke zur threadübergreifenden Kommunikation darin verteilt wird.
- Einen lesbaren Cache für konstante Datenwerte, die bei Hochladen des Kernels initialisiert werden können.
- Einen lesbaren Cache für Texturen, der häufig dafür genutzt wird größere Datenmengen abzulegen, auf die nur lesend zugegriffen werden muss.
- Einen globalen les- und beschreibbaren Datenspeicher der Grafikhardware, dessen Größenvorteil durch langsame und ungecachte Zugriffe relativiert wird.

B.2. Parallelisieren der Beobachtungsevaluation

Der folgende Abschnitt befasst sich mit der Parallelisierung der in Kapitel 3.1.4 aufgeführten Beobachtungsmodelle. Im Einzelnen handelt es sich dabei um die Evaluation mit Gradientenhistogrammen sowie der Anwendung Neuronaler Netze zur Drehwinkelbewertung.

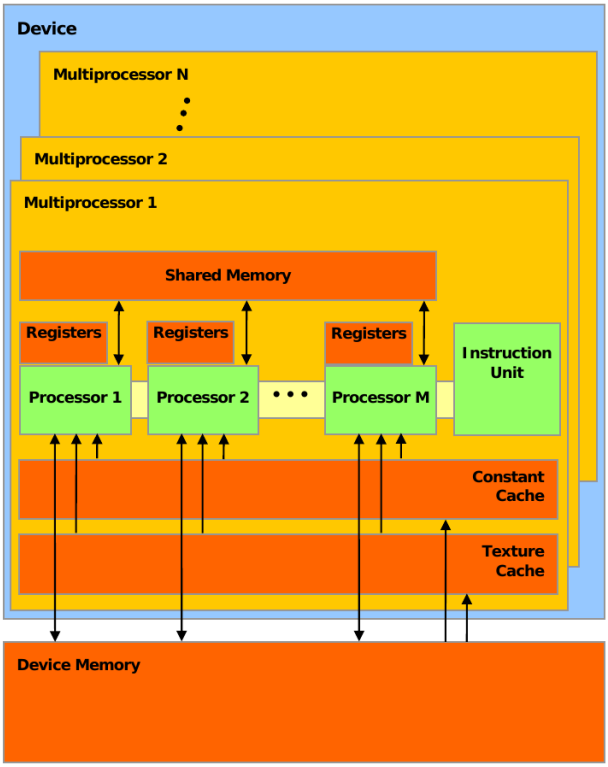


Abb. B.5.: Speicherorganisation bei Mehrprozessorarchitekturen. Quelle: [NVI10].

B.2.1. Gradientenhistogramme

Die Berechnung der Gradientenhistogramme zu gegebenen Motiven verlangt die Berücksichtigung der überlappenden Zellbereiche zu Blöcken. Zuvor müssen jedoch die lokalen Histogramme auf den Gradientenmagnituden des Motivbilds erstellt werden. Eine Parallelisierung ist hier möglich, wenn die Arbeitsschritte voneinander gelöst und diese im Einzelnen parallel ausgeführt werden.

- **Berechnung der Zellhistogramme**
  Ein Motiv wird in 8×8 Pixel große Zellen aufgeteilt, für die jeweils ein Histogramm bezüglich der Gradienten erstellt wird. Um Caching auszunutzen, werden hierzu pro Zellenzeile Histogramme berechnet, die anschließend akkumuliert die finale Zellhistogrammbeschreibung ergeben. Der Schritt wird in Abbildung B.6 visualisiert.
  Der Vorteil quadratischer Zellgrößen kann während der Akkumulierung der Histogrammtöpfe ausgenutzt werden: In einem ersten Schritt summieren 8 Threads die Töpfe der beiden Zeilenhistogramme 1 und 5, die Töpfe der Zeilenhistogramme 2 und 6, usw., bis die 8 Zeilen auf die Hälfte halbiert werden konnten. Anschließend wiederholt man den Prozess und halbiert die Histogrammmenge erneut. Ein dritter Schritt erstellt schließlich die endgültigen Zellhistogramme. Die Vorgehensweise hierzu wird im Detail in Abbildung B.7 dargestellt.

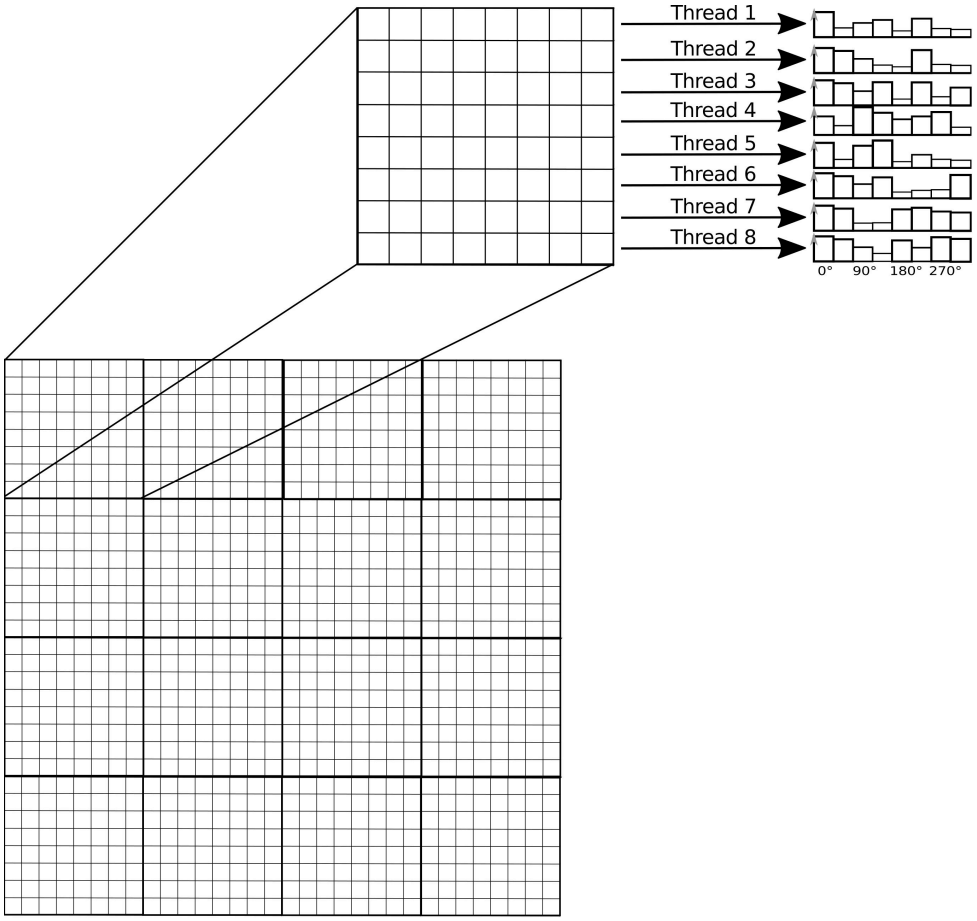


Abb. B.6.: Pro Thread wird ein Histogramm über acht Pixel im jeweiligen Block berechnet.

- **Berechnung der Blöcke**
  Im nächsten Schritt werden die Zellhistogramme zu Blöcken zusammengefasst. Ein Block bezieht sich dabei überlappend auf 3×3 Zellen, die konkateniert werden müssen. Bei einem beispielhaften Motiv der Größe 24×24 Pixel ergibt das einen einzelnen Block über 3×3 Zellen.
  Auch hier kann wieder der Vorteil der 8 Threads pro Zelle ausgenutzt werden. Wurden daneben so viele Threads gestartetem wie Zellen vorhanden sind, im Beispiel also 9, so kann die erste Dimension (8 Threads) über die 8 Töpfe der Histogramme, die zweite und dritte über die Zellzeilen und -spalten parallelisieren. Für Zelle $(1,1)$ wird das Histogramm durch 8 Threads so in einen gemeinsamen, threadübergreifenden Speicher geschrieben. Für die übrigen Zellen $(1,2),(1,3),(2,1),\ldots$ ebenso. Wie beim Übertragen der Zeilenhistogramme zu einem Zellhistogramm, kann anschließend vorgegangen werden, um alle Töpfe des Blockhistogramms zu akkumulieren und schließlich threadlastig zur Normalisierung der Einträge zu nutzen.

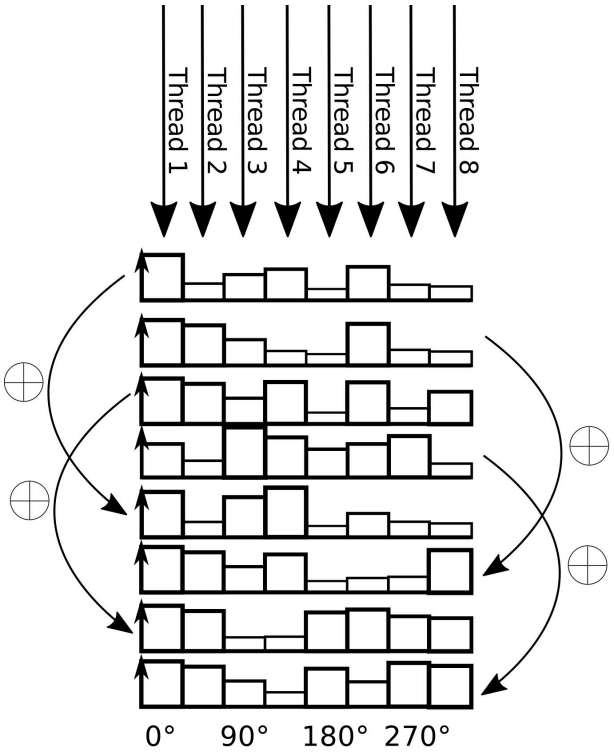


Abb. B.7.: Die acht Blockhistogramme werden zu einem einzigen zusammengefasst, indem jeder Thread einen Bin repräsentiert.

B.2.2. Künstliche Neuronale Netze

In vollständig vorwärtsgerichteten Netzen wird die Aktivierung von Neuronen allein durch ihre Eingabe durch die Vorgängerschicht bestimmt. Die Neuronen feuern dabei unabhängig voneinander und erlauben auch unabhängig voneinander bearbeitet zu werden, solange die Reihenfolge der Schichten berücksichtigt wird. Wie wir in Kapitel 3.2.1 ausgeführt haben, werden im Rahmen dieser Arbeit dreischichtige, vollständig vorwärts gerichtete Netze eingesetzt, um eine Wahrscheinlichkeitsfunktion über die Drehwinkelklassen zu berechnen. Jede Schicht kann damit als eigener CUDA-Kernel implementiert werden, der die Aktivierung der ihm entsprechenden Neuronen auf jeweilige Threads abbildet und parallel berechnet. Abbildung B.8 beschreibt die grundsätzliche Vorgehensweise. Der Ablauf wird dabei auf zwei Kernel aufgeteilt: Einen zur Berechnung der Aktivierung der zweiten, versteckten Schicht und einer zur Berechnung der Ausgabe durch die dritte Schicht. Zur Ausnutzung einer massiv parallelen Abarbeitung, werden die Merkmalsvektoren aller Zustandshypothesen in jedem Frame in den Texturspeicher der Grafikhardware geladen. Die Parameter der Verbindungsgewichte und Aktivierungsfunktionen dagegen nur initial bei Programmstart, da diese konstant bleiben und sich im Lauf der Ausführung nicht verändern. Die Threadadressierung geschieht zweidimensional: Die erste Komponente bezeichnet den Index in der Menge zu evaluierender Merkmalsvektoren und identifiziert diesen eindeutig im Texturspeicher. Die zweite bezieht sich auf das jeweilige Neuron

der momentan zu verarbeitenden Schicht. Für ein beispielhaftes Netz mit 80 Neuronen in der versteckten Schicht und 36 Neuronen in der Ausgabeschicht, werden so bei 100 Mustern durch 100×80 Threads die Ausgaben der zweiten Schicht parallel berechnet. Beschränkungen sind durch die Grafikhardware festgelegt, die die Anzahl der gleichzeitig ausführbaren Threads pro Gatterblock je nach Ausbau des Dies vorgibt. Für Spezifikationen diesbezüglich sei der Leser allerdings an weiterführende CUDA-Referenzen und Hardwarespezfikationen der NVIDIA Corporation verwiesen.
Zum Berechnen der endgültigen Netzausgaben, wird die Ausführung im entsprechenden CUDA-Kernel ebenfalls mit zwei Dimensionen beschrieben: Die erste bezieht sich ebenfalls auf den Index des zu bewertenden Motivs, die zweite nummeriert wieder die jeweiligen Neuronen der Ausgabeschicht durch. In obigem Beispiel führt dies zu 100×36 parallelen Threads. Die Zwischenergebnisse der Vorgängerschicht werden hierzu im Speicher der Grafikkarte gehalten. Die Neuronen iterieren während ihrer Abarbeitung dabei durch alle Vorgängereinheiten zu denen sie jeweils verbunden sind und akkumulieren die entsprechenden Ausgaben. Im Beispiel sammelt so jedes Ausgabeneuron die Ausgaben aller 80 Vorgänger und verarbeitet deren Summe in der parametrisierten Aktivierungsfunktion.

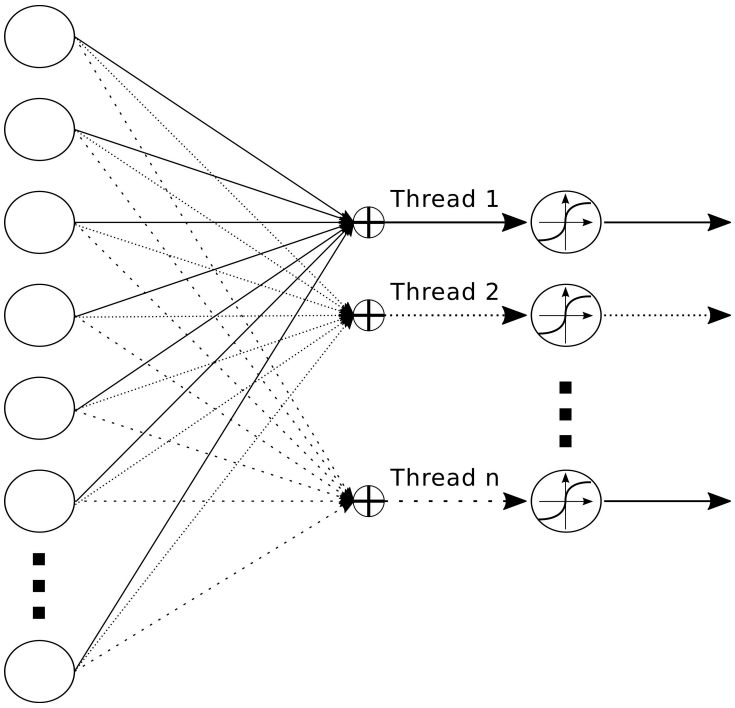


Abb. B.8.: Jedes Neuron wird von einem einzelnen Thread bearbeitet. Im ersten Schritt wird die Summer der Eingangssignale bestimmt, im zweiten Schritt die jeweiligen Neuronenaktivierung ermittelt.

C. Tabellen detaillierter Ergebnisse der Kopfdrehungsschätzung

In diesem Anhang sind die detaillierten Ergebnisse der Evaluation der Kopfdrehungsschätzung aufgeführt. Zunächst werden die mittleren Fehler des Gesamtsystems angegeben, das unter verschiedenen Parametereinflüssen evaluiert wurde (Winkeldiskretisierung, Anzahl Partikel, Standardabweichung des aufaddierten Rauschterms). Im Anschluss folgen die Ergebnisse des implementierten Referenzsystems, das eine vorangehende Kameraselektion durch Vorderkopfdetektion anwendet und auf den Frontalaufnahmen die Kopfdrehung einzeln schätzt und diese für eine finale Hypothese mittelt.

C.1. Mittlerer Fehler des Gesamtsystems

Klassenbreite [°]	#Partikel	$\sigma_{horiz.}$	$\sigma_{vert.}$	Fehler horiz. [°]	Fehler vert. [°]
1	250	10	3	16,2	12,6
		15	5	9,8	10,9
		20	10	**7,8**	**10,6**
		25	15	8,4	10,8
		30	20	8,5	11,1
	500	10	3	12,5	12,2
		15	5	7,0	10,8
		20	10	**6,4**	**10,4**
		25	15	7,4	10,5
		30	20	8,7	10,9
	1000	10	3	10,5	11,6
		15	5	6,6	10,5
		20	10	**6,5**	**10,3**
		25	15	7,7	10,4
		30	20	8,9	10,7

Tab. C.1.: Mittlerer Fehler des Gesamtsystems bei einer Winkeldiskretierung von 1°.

Klassenbreite [°]	#Partikel	$\sigma_{horiz.}$	$\sigma_{vert.}$	Fehler horiz. [°]	Fehler vert. [°]
3	250	10	3	19,3	11,9
		15	5	7,5	10,2
		20	10	**6,3**	**9,9**
		25	15	7,1	10,0
		30	20	8,6	10,2
	500	10	3	10,9	11,3
		15	5	8,4	10,1
		20	10	**6,4**	**9,6**
		25	15	6,6	9,8
		30	20	7,6	10,1
	1000	10	3	10,9	10,9
		15	5	**6,5**	9,8
		20	10	6,8	**9,5**
		25	15	7,2	9,7
		30	20	7,7	9,9

Tab. C.2.: Mittlerer Fehler des Gesamtsystems bei einer Winkeldiskretierung von 3°.

Klassenbreite [°]	#Partikel	$\sigma_{horiz.}$	$\sigma_{vert.}$	Fehler horiz. [°]	Fehler vert. [°]
5	250	10	3	15,4	11,1
		15	5	6,9	9,3
		20	10	6,8	**9,1**
		25	15	**6,5**	9,3
		30	20	6,6	9,5
	500	10	3	10,0	10,8
		15	5	**5,8**	9,1
		20	10	5,9	**8,9**
		25	15	6,4	9,0
		30	20	7,1	9,2
	1000	10	3	10,8	9,9
		15	5	5,6	8,9
		20	10	**5,4**	**8,7**
		25	15	6,6	8,9
		30	20	7,0	9,1

Tab. C.3.: Mittlerer Fehler des Gesamtsystems bei einer Winkeldiskretierung von 5°.

Klassenbreite [°]	#Partikel	$\sigma_{horiz.}$	$\sigma_{vert.}$	Fehler horiz. [°]	Fehler vert. [°]
10	250	10	3	12,3	11,2
		15	5	6,7	9,7
		20	10	6,9	**9,4**
		25	15	6,6	9,6
		30	20	**6,4**	9,7
	500	10	3	9,6	10,7
		15	5	6,2	9,4
		20	10	**5,7**	**9,1**
		25	15	6,2	9,3
		30	20	6,8	9,5
	1000	10	3	8,1	10,2
		15	5	5,9	9,2
		20	10	**5,3**	**8,9**
		25	15	6,0	9,2
		30	20	6,3	9,3

Tab. C.4.: Mittlerer Fehler des Gesamtsystems bei einer Winkeldiskretierung von 10°.

Klassenbreite [°]	#Partikel	$\sigma_{horiz.}$	$\sigma_{vert.}$	Fehler horiz. [°]	Fehler vert. [°]
15	250	10	3	10,4	11,6
		15	5	7,0	9,8
		20	10	**6,1**	**9,3**
		25	15	6,9	9,5
		30	20	6,7	9,7
	500	10	3	9,3	10,9
		15	5	6,2	9,3
		20	10	**5,7**	**8,9**
		25	15	6,2	9,2
		30	20	6,5	9,4
	1000	10	3	7,9	10,4
		15	5	5,8	9,1
		20	10	**5,7**	**8,9**
		25	15	5,9	9,0
		30	20	6,4	9,2

Tab. C.5.: Mittlerer Fehler des Gesamtsystems bei einer Winkeldiskretierung von 15°.

C.2. Mittlerer Fehler des Referenzsystems

Personenvideo	Fehler Ref.		Fehler Ref. (Kalman)		Fehler Gesamtsystem	
	horiz. [°]	vert. [°]	horiz. [°]	vert. [°]	horiz. [°]	vert. [°]
01	34,3	22,1	39,7	21,3	4,6	11,1
02	39,6	31,2	46,2	29,4	4,6	7,3
03	64,8	27,6	70,7	26,5	1,5	8,7
04	39,9	21,9	41,8	21,1	3,9	4,0
05	48,6	17,4	61,9	16,6	3,4	13,3
06	33,4	32,2	41,9	34,1	18,0	12,3
07	50,5	25,1	55,1	24,1	2,0	4,1
08	37,9	38,5	39,8	38,6	5,4	9,6
09	73,9	36,1	71,1	30,7	1,5	9,1
10	39,9	26,8	50,5	26,5	7,5	7,8
11	49,6	24,3	53,3	25,2	6,6	6,2
12	69,4	25,2	70,2	25,3	5,2	9,6
13	49,4	42,5	56,4	45,8	4,7	19,5
14	54,7	49,5	62,4	50,3	6,5	8,3
15	43,9	30,3	51,7	29,6	3,6	5,5
Durchschnitt	48,3	30,3	53,9	30,1	5,3	9,0

Tab. C.6.: Detaillierte Gegenüberstellung des mittleren Fehlers des in der Arbeit entworfenen Gesamtsystems zum Referenzsystem. Für das Gesamtsystem wurden die Ergebnisse für 1000 Partikel, einer Winkeldiskretisierung von 10° und einem Rauschen von $(20^\circ, 10^\circ)$ aufgelistet.

D. Tabellen detaillierter Ergebnisse der Zuwendung der Aufmerksamkeit

In diesem Anhang sind die detaillierten Korrektklassifikationsraten der Aufmerksamkeitszuwendungsschätzung aufgeführt. Die Ergebnisse werden hierfür für alle Besprechungsvideos im Datensatz (10 Stück) und unter Bezug auf die Annotationen aller Annotatoren (3 Annotatoren) angegeben. Aus Gründen der Vollständigkeit wird die Korrektklassifikationsrate ebenfalls im Hinblick auf lediglich jenen Frames in den Videos berechnet, in denen alle Annotatoren übereinstimmten.
Für Person 3 - jenem Teilnehmer während der Besprechungen der immer an der Südseite des Besprechungstisches Platz nahm und einen Sensor auf dem Kopf trug um die tatsächliche Kopfdrehung zu protokollieren - werden die Ergebnisse daneben bezüglich der Winkelprotokolle aufgeführt, statt nur auf den automatisch geschätzten Kopfdrehwinkeln.

D.1. Referenzansatz: Geometrisches Schließen

			KKR pro Besprechungsszenario										KKR ∅
			1	2	3	4	5	6	7	8	9	10	
Winkelschätzungen	Person 1	A1	61,7	75,9	66,6	59,1	68,2	50,5	-	61,0	-	55,8	62,1
		A2	61,6	70,7	59,9	55,9	67,5	50,3	-	58,9	-	57,1	60,4
		A3	64,8	76,6	62,3	60,2	71,0	52,2	-	58,3	-	60,0	63,2
		konf.	71,6	86,2	72,1	64,8	76,5	60,5	-	67,2	-	67,4	70,9
	Person 2	A1	76,5	74,3	64,3	46,7	34,0	34,4	64,1	43,7	73,2	36,5	56,7
		A2	72,4	80,0	67,8	47,1	39,5	38,5	64,3	43,5	66,9	35,1	56,8
		A3	75,6	75,9	67,5	44,3	34,4	41,5	69,4	43,3	77,1	39,3	58,7
		konf.	86,5	83,4	74,5	49,2	37,2	37,9	74,2	49,1	91,2	38,5	65,2
	Person 3	A1	49,2	56,6	34,7	41,6	45,7	24,2	30,8	43,5	40,9	51,0	42,5
		A2	49,5	47,3	35,5	36,4	35,7	23,0	31,1	45,5	35,6	49,4	39,9
		A3	55,9	56,1	34,4	38,1	45,0	20,6	25,8	47,1	44,7	55,7	43,5
		konf.	56,3	58,1	38,1	42,4	49,4	26,7	31,7	51,4	48,6	57,9	46,9
	Person 4	A1	59,7	-	54,0	50,7	-	56,2	63,2	46,0	47,7	39,8	52,8
		A2	62,4	-	54,2	41,7	-	56,3	60,5	46,4	44,7	41,6	52,1
		A3	62,3	-	55,1	45,1	-	61,8	61,3	44,3	49,0	42,6	53,6
		konf.	65,0	-	57,1	52,2	-	66,2	69,6	53,2	54,6	42,8	58,5
W.protokolle	Person 3	A1	60,1	62,0	38,9	35,9	43,5	24,0	32,5	46,7	45,9	44,8	45,0
		A2	59,6	57,0	36,0	34,0	38,6	21,0	31,6	47,3	43,7	49,6	43,6
		A3	61,4	55,7	35,4	31,7	44,0	21,5	27,7	53,4	51,1	48,5	44,7
		konf.	68,7	66,2	39,5	37,5	47,4	25,6	33,5	57,3	59,1	55,3	50,9

Tab. D.1.: Korrektklassifikationsrate in [%] wenn die Kopfdrehung geometrisch als Blickrichtung interpretiert wird. Zielmenge: Personen, Tisch, Leinwand (insg. 5-7 Ziele, da untersch. Anzahl Personen pro Video).

			KKR pro Besprechungsszenario										KKR ∅
			1	2	3	4	5	6	7	8	9	10	
Winkelschätzungen	Person 1	A1	28,3	46,0	35,8	31,9	37,7	23,4	-	33,6	-	24,6	32,2
		A2	30,3	41,3	30,4	30,3	36,8	24,7	-	31,0	-	24,3	31,0
		A3	31,5	45,6	32,4	31,8	40,4	25,1	-	31,0	-	27,2	33,0
		konf.	33,7	53,2	37,2	35,2	42,8	29,5	-	36,1	-	30,3	37,2
	Person 2	A1	50,8	57,5	43,8	29,2	23,4	21,6	41,9	17,5	44,3	26,8	36,9
		A2	47,4	60,6	46,1	28,1	26,3	23,2	38,0	17,3	39,6	24,4	36,0
		A3	50,5	59,7	46,9	28,9	23,5	25,3	44,4	16,9	47,0	28,4	38,3
		konf.	57,0	66,5	53,8	32,4	26,1	23,5	48,4	19,4	56,5	27,7	43,0
	Person 3	A1	37,5	46,0	23,4	23,0	31,1	15,2	15,3	28,3	27,9	34,2	29,2
		A2	36,9	36,3	24,4	18,4	24,6	15,4	15,3	30,3	22,9	33,9	26,9
		A3	44,3	47,0	23,6	18,7	31,5	14,2	12,8	31,9	29,0	38,6	30,6
		konf.	39,6	45,9	25,8	21,1	39,2	18,5	15,8	31,8	32,1	41,7	31,9
	Person 4	A1	35,8	-	32,2	33,8	-	16,8	21,2	22,1	19,0	16,1	25,4
		A2	36,9	-	32,8	28,4	-	17,9	19,1	20,7	18,2	17,0	24,7
		A3	36,9	-	33,7	30,9	-	19,6	20,4	19,4	20,2	18,6	25,7
		konf.	38,0	-	33,7	36,3	-	21,5	24,6	25,9	22,5	14,2	28,1
W.protokolle	Person 3	A1	42,6	39,4	30,6	17,7	36,6	12,1	26,9	30,9	28,0	31,9	30,6
		A2	41,7	37,2	28,5	18,3	33,7	9,9	27,3	30,3	26,9	35,7	29,9
		A3	43,1	36,6	29,1	15,3	36,0	12,0	26,3	38,3	31,6	36,6	31,5
		konf.	46,7	43,7	34,9	18,7	47,0	13,9	31,1	38,0	37,6	41,8	36,2

Tab. D.2.: Korrektklassifikationsrate in [%] wenn die Kopfdrehung geometrisch als Blickrichtung interpretiert wird. Zielmenge: Personen, alle Objekte und Mobiliar (insg. 38-40 Ziele, da untersch. Anzahl Personen pro Video).

D.2. Systemergebnisse mit gesamtem Kontextbezug

			KKR pro Besprechungsszenario										KKR ∅
			1	2	3	4	5	6	7	8	9	10	
Winkelschätzungen	Person 1	A1	61,7	77,3	68,7	61,9	70,6	63,1	-	67,6	-	60,1	66,0
		A2	60,1	70,2	62,0	58,9	69,6	66,4	-	65,1	-	59,8	63,8
		A3	62,0	77,4	64,9	63,3	75,3	65,0	-	65,3	-	64,3	67,0
		konf.	73,1	85,5	75,9	67,7	79,4	72,8	-	73,7	-	70,1	74,8
	Person 2	A1	70,4	72,1	63,2	47,7	42,6	39,4	64,7	70,7	75,8	56,6	62,0
		A2	69,2	78,6	67,0	49,2	49,3	45,2	64,0	70,2	66,8	57,0	62,7
		A3	69,1	75,8	66,2	46,4	42,5	46,1	65,7	70,2	78,4	58,7	63,5
		konf.	80,3	81,6	74,4	53,8	48,5	43,9	71,7	79,5	91,3	66,0	71,8
	Person 3	A1	59,3	56,6	46,0	61,0	58,3	39,2	49,0	49,4	64,1	61,8	54,8
		A2	57,2	52,4	43,9	59,2	41,9	40,5	48,0	53,7	57,3	61,6	52,0
		A3	49,6	54,1	44,9	57,8	56,4	38,0	41,0	48,1	66,6	67,3	52,3
		konf.	62,3	56,9	48,1	64,3	61,1	46,7	50,7	57,9	80,0	71,4	60,0
	Person 4	A1	64,1	-	53,6	65,9	-	52,8	64,6	50,7	49,3	54,1	57,3
		A2	65,7	-	56,0	49,1	-	49,5	64,4	49,8	47,6	55,0	55,7
		A3	67,2	-	54,8	60,9	-	57,4	64,0	50,8	51,9	57,8	58,7
		konf.	70,7	-	55,7	64,0	-	58,9	71,7	57,3	56,0	59,9	62,6
W.protokolle	Person 3	A1	69,3	60,6	57,0	65,1	75,4	36,8	53,4	66,2	56,9	54,8	59,9
		A2	66,2	66,3	52,9	62,8	59,1	36,7	50,0	64,7	56,7	55,4	58,0
		A3	56,3	59,6	52,8	60,0	72,6	33,7	46,4	61,6	62,1	57,8	56,3
		konf.	76,0	67,0	60,3	69,6	81,6	41,3	56,7	77,1	73,4	64,8	66,9

Tab. D.3.: Korrektklassifikationsrate des Gesamtsystems inkl. Kontextbezug in [%]. Zielmenge: Personen, Tisch, Leinwand (insg. 5-7 Ziele, da untersch. Anzahl Personen pro Video).

			KKR pro Besprechungsszenario										KKR ∅
			1	2	3	4	5	6	7	8	9	10	
Winkelschätzungen	Person 1	A1	43,0	55,1	45,5	46,1	48,2	51,4	-	49,2	-	42,6	47,3
		A2	45,8	50,0	39,2	41,9	45,6	50,6	-	47,8	-	41,9	45,3
		A3	47,0	55,8	41,9	46,9	50,8	49,5	-	46,7	-	45,8	48,0
		konf.	47,8	62,0	45,0	48,3	54,5	61,2	-	54,7	-	50,2	52,9
	Person 2	A1	54,4	50,5	47,0	30,2	23,1	27,8	50,2	39,7	57,3	36,0	43,2
		A2	56,3	54,6	49,6	35,5	28,5	31,3	50,1	39,7	48,2	41,9	44,8
		A3	55,9	53,6	49,6	31,6	22,7	32,4	51,7	38,2	58,8	41,4	45,1
		konf.	66,4	59,6	58,1	36,4	26,3	30,0	57,3	45,3	66,5	45,9	51,6
	Person 3	A1	47,6	47,8	27,2	45,3	42,2	31,8	29,3	36,6	32,0	47,8	39,1
		A2	46,7	45,6	26,5	41,6	33,6	32,9	31,3	41,0	25,8	48,5	37,8
		A3	37,1	46,4	25,9	41,7	41,4	28,8	24,4	36,4	35,1	51,7	36,9
		konf.	48,4	48,7	27,8	46,3	51,5	35,7	30,1	41,8	33,6	56,9	41,8
	Person 4	A1	43,8	-	40,7	42,3	-	42,1	43,9	42,4	30,8	37,9	40,7
		A2	45,5	-	43,7	33,1	-	42,5	42,9	42,6	29,3	39,6	40,4
		A3	45,3	-	41,6	40,5	-	46,5	42,1	41,5	32,0	40,6	41,4
		konf.	47,7	-	42,7	45,0	-	49,4	47,1	49,2	35,0	42,2	45,0
W.protokolle	Person 3	A1	47,1	45,4	33,2	40,7	52,5	29,6	28,2	47,9	22,1	36,2	38,5
		A2	46,8	48,7	29,8	38,8	45,1	29,7	28,3	44,4	25,3	40,0	38,2
		A3	38,1	43,5	30,6	36,4	50,8	26,3	24,4	49,7	26,6	39,9	36,6
		konf.	49,9	51,5	35,6	41,3	63,8	33,5	29,7	53,9	27,1	46,4	42,8

Tab. D.4.: Korrektklassifikationsrate des Gesamtsystems inkl. Kontextbezug in [%]. Zielmenge: Personen, alle Objekte und Mobiliar (insg. 38-40 Ziele, da untersch. Anzahl Personen pro Video).

D.3. Einfluß des Kontextbezugs

D.3.1. Ergebnisse ohne Kontextbezug

			KKR pro Besprechungsszenario										KKR ∅
			1	2	3	4	5	6	7	8	9	10	
Winkelschätzungen	Person 1	A1	24,3	10,2	22,2	18,4	22,9	22,5	-	21,7	-	10,5	18,9
		A2	28,2	13,4	23,9	18,1	22,1	25,4	-	21,9	-	11,5	20,5
		A3	22,9	14,3	23,1	16,9	20,5	22,8	-	20,8	-	10,1	18,8
		konf.	22,7	8,9	25,1	20,9	20,0	26,2	-	20,6	-	10,9	19,1
	Person 2	A1	31,7	26,7	31,0	32,4	34,0	31,6	21,2	14,9	17,5	20,1	25,0
		A2	32,8	28,3	36,0	32,1	34,3	26,4	22,0	16,2	6,8	29,9	25,8
		A3	32,7	27,5	34,6	36,2	33,3	27,4	21,9	13,7	17,1	24,8	26,1
		konf.	37,2	26,9	37,1	38,9	34,7	33,2	26,4	14,0	3,1	22,6	25,7
	Person 3	A1	27,7	22,0	28,5	43,7	34,0	25,9	25,4	23,1	26,7	25,9	27,7
		A2	29,2	31,0	31,9	42,1	28,3	25,6	28,6	24,3	18,7	28,4	28,7
		A3	19,4	24,9	30,4	39,9	31,0	21,1	25,5	22,1	26,0	25,1	25,8
		konf.	22,6	19,3	29,8	43,7	39,8	22,9	24,7	21,9	13,1	28,0	25,3
	Person 4	A1	36,9	-	29,4	42,0	-	29,0	24,5	32,7	21,1	28,5	30,6
		A2	36,4	-	32,4	49,3	-	32,9	26,3	32,5	10,4	29,9	31,0
		A3	39,3	-	26,3	39,6	-	31,1	25,0	37,7	21,2	26,4	30,9
		konf.	37,2	-	26,5	43,0	-	31,8	22,5	29,5	8,2	30,8	29,1
W.protokolle	Person 3	A1	30,3	21,5	27,0	42,1	46,3	23,5	27,8	29,7	25,6	27,2	29,6
		A2	32,8	32,6	25,1	40,9	38,5	24,2	30,0	29,2	20,4	28,3	30,7
		A3	22,8	26,0	23,9	37,3	42,8	18,7	27,3	26,5	23,8	23,6	26,7
		konf.	27,1	20,2	27,1	41,6	57,6	20,4	26,7	27,0	14,1	25,6	27,9

Tab. D.5.: Korrektklassifikationsrate des Gesamtsystems ohne Kontextbezug in [%]. Zielmenge: Personen, Tisch, Leinwand (insg. 5-7 Ziele, da untersch. Anzahl Personen pro Video).

			KKR pro Besprechungsszenario										KKR ∅
			1	2	3	4	5	6	7	8	9	10	
Winkelschätzungen	Person 1	A1	13,1	8,6	12,9	14,5	18,6	11,2	-	14,0	-	7,5	12,3
		A2	11,6	6,6	11,5	14,1	17,4	10,5	-	14,6	-	8,2	11,6
		A3	10,9	7,3	11,8	12,9	16,6	9,1	-	12,9	-	7,1	10,9
		konf.	10,3	7,6	14,4	16,7	16,5	11,0	-	12,8	-	7,7	11,9
	Person 2	A1	34,7	31,8	36,1	45,7	57,2	18,3	14,3	15,1	11,8	27,0	28,4
		A2	43,0	33,6	39,7	44,9	63,8	17,8	13,9	16,3	11,6	35,2	31,5
		A3	36,1	30,1	36,6	41,7	60,9	19,8	14,8	14,6	11,5	31,2	28,9
		konf.	37,4	29,2	38,0	45,9	65,2	22,1	18,6	12,9	2,6	28,1	28,8
	Person 3	A1	40,5	34,3	37,6	49,6	56,8	44,5	10,9	23,9	17,5	38,4	34,6
		A2	40,8	39,7	37,9	45,7	50,5	41,7	11,0	26,5	13,8	39,6	34,4
		A3	27,8	31,7	35,7	43,3	53,5	39,0	11,5	22,1	18,4	36,5	30,8
		konf.	36,5	31,7	35,2	48,4	72,7	50,7	11,7	23,6	7,3	40,5	33,3
	Person 4	A1	34,0	-	29,7	43,2	-	30,8	10,2	23,8	9,2	38,1	27,3
		A2	34,4	-	35,7	49,5	-	31,4	11,7	22,8	12,8	37,8	29,0
		A3	38,4	-	28,6	39,8	-	32,2	11,6	21,7	12,0	36,1	27,6
		konf.	34,3	-	30,5	44,8	-	36,8	9,8	18,6	5,6	38,5	27,7
W.protokolle	Person 3	A1	40,8	34,4	39,2	50,7	58,4	45,3	10,8	24,1	17,4	38,0	35,1
		A2	41,1	40,8	38,7	46,9	52,5	40,5	10,6	26,1	14,1	38,7	34,7
		A3	28,2	32,2	36,9	43,9	54,4	37,3	11,4	21,5	18,3	35,6	30,8
		konf.	36,9	32,4	36,7	49,5	74,5	49,8	11,4	22,2	7,4	39,0	33,4

Tab. D.6.: Korrektklassifikationsrate des Gesamtsystems ohne Kontextbezug in [%]. Zielmenge: Personen, alle Objekte und Mobiliar (insg. 38-40 Ziele, da untersch. Anzahl Personen pro Video).

D.3.2. Ergebnisse mit Bewegung als einzigem Kontextbezug

			KKR pro Besprechungsszenario										KKR ∅
			1	2	3	4	5	6	7	8	9	10	
Winkelschätzungen	Person 1	A1	62,2	76,3	68,3	61,7	69,9	62,6	-	67,8	-	60,2	65,8
		A2	60,6	69,8	62,0	58,5	69,1	65,6	-	65,4	-	59,8	63,6
		A3	63,4	76,7	64,9	63,1	74,6	64,5	-	65,4	-	64,4	67,0
		konf.	74,0	84,4	75,8	67,4	78,8	72,3	-	73,9	-	70,4	74,7
	Person 2	A1	70,6	71,7	63,9	47,5	41,7	38,2	65,0	70,5	75,3	55,5	61,8
		A2	68,9	78,5	68,1	49,3	48,9	43,4	63,9	70,0	66,3	55,1	62,3
		A3	69,8	75,6	67,8	46,5	41,9	45,1	66,0	70,0	77,9	57,4	63,4
		konf.	80,7	81,8	75,9	53,8	47,9	41,8	71,9	79,3	90,7	64,0	71,5
	Person 3	A1	58,5	56,6	46,0	60,6	57,8	38,8	48,8	49,6	64,0	62,6	54,6
		A2	57,0	51,8	44,1	59,1	41,6	40,2	47,9	52,6	57,1	63,0	51,8
		A3	51,5	54,5	45,0	57,8	55,9	37,7	40,8	47,1	66,7	69,0	52,6
		konf.	62,4	56,8	48,4	64,0	61,0	46,2	50,5	57,4	79,8	73,5	60,1
	Person 4	A1	64,1	-	55,5	65,4	-	52,3	64,9	50,2	49,0	54,1	57,4
		A2	65,5	-	57,6	48,7	-	49,4	64,8	49,7	47,0	54,9	55,8
		A3	67,1	-	56,5	60,4	-	57,0	64,1	50,3	51,3	57,8	58,6
		konf.	70,6	-	58,1	63,3	-	58,6	72,3	56,6	55,4	60,0	62,8
W.protokolle	Person 3	A1	69,9	63,6	55,9	64,7	74,2	37,1	52,5	67,5	58,3	54,2	60,3
		A2	66,8	66,5	52,0	62,7	58,6	35,7	49,6	65,9	57,1	55,2	58,0
		A3	58,8	61,8	52,1	60,1	72,5	34,0	46,8	63,6	63,7	57,3	57,3
		konf.	77,6	70,2	59,2	69,6	81,4	41,1	56,4	79,2	75,0	64,4	67,7

Tab. D.7.: Korrektklassifikationsrate des Gesamtsystems mit Bewegung als einzigem Kontextbezug in [%]. Zielmenge: Personen, Tisch, Leinwand (insg. 5-7 Ziele, da untersch. Anzahl Personen pro Video).

			KKR pro Besprechungsszenario										KKR ∅
		1	2	3	4	5	6	7	8	9	10		
Winkelschätzungen	Person 1	A1	42,0	53,0	44,2	45,0	45,2	48,1	-	45,2	-	39,7	45,0
		A2	45,0	48,0	37,9	40,6	42,8	46,9	-	43,8	-	38,8	42,9
		A3	46,3	53,6	40,7	45,4	47,9	46,3	-	42,4	-	42,8	45,6
		konf.	46,7	59,1	43,6	46,9	51,3	57,6	-	50,0	-	46,9	50,2
	Person 2	A1	52,3	48,2	46,9	28,8	21,6	26,0	48,5	35,9	54,2	34,0	41,2
		A2	53,9	52,0	49,0	34,0	26,9	29,0	47,9	35,7	45,4	39,6	42,6
		A3	54,2	52,0	49,3	30,3	21,3	30,1	49,9	34,3	55,9	39,1	43,2
		konf.	63,9	57,9	58,0	34,9	24,8	28,1	55,4	40,7	62,6	43,4	49,3
	Person 3	A1	46,3	46,2	26,3	43,8	39,6	30,5	27,8	34,3	31,4	46,4	37,6
		A2	45,4	42,5	25,9	40,3	32,3	31,5	30,0	39,3	25,5	47,4	36,4
		A3	37,3	45,1	25,0	40,4	39,1	27,9	23,2	34,8	34,4	50,7	35,8
		konf.	47,9	46,2	26,9	44,7	50,1	34,9	28,2	38,8	33,3	55,7	40,4
	Person 4	A1	40,0	-	38,5	39,8	-	39,5	39,5	40,1	27,4	35,7	37,6
		A2	42,2	-	41,3	31,2	-	39,9	38,2	40,1	26,3	37,3	37,5
		A3	41,4	-	38,9	38,3	-	43,7	37,8	39,2	28,7	38,5	38,4
		konf.	43,9	-	40,4	42,2	-	46,7	42,6	46,1	31,1	39,7	41,7
W.protokolle	Person 3	A1	45,0	44,2	31,8	39,6	49,0	28,9	27,0	45,7	22,0	35,2	37,0
		A2	44,5	46,2	28,4	37,5	43,0	28,3	27,2	42,6	24,8	39,2	36,7
		A3	36,9	42,0	29,3	35,1	47,7	25,9	23,5	48,3	26,1	39,2	35,4
		konf.	47,3	50,0	33,9	39,8	60,8	32,5	28,4	51,7	27,1	45,2	41,3

Tab. D.8.: Korrektklassifikationsrate des Gesamtsystems mit Bewegung als einzigem Kontextbezug in [%]. Zielmenge: Personen, alle Objekte und Mobiliar (insg. 38-40 Ziele, da untersch. Anzahl Personen pro Video).

E. Abbildungsverzeichnis

F. Tabellenverzeichnis

G. Verzeichnis eigener Veröffentlichungen

[LBC+09] O. Lanz, R. Brunelli, P. Chippendale, M. Voit und R. Stiefelhagen: *Extracting Interaction Cues: Focus of Attention, Body Pose, and Gestures*. In: A. Waibel und R. Stiefelhagen (Herausgeber): *Computers in the Human Interaction Loop*, Seiten 87–93. Springer, 2009.

[NEVS06] K. Nickel, H. K. Ekenel, M. Voit und R. Stiefelhagen: *Audio-Visual Perception of Humans for a Humanoid Robot*. In: *Proc. 2nd Int'l Workshop on Human-Centered Robotic Systems*, 2006.

[SBE+06] R. Stiefelhagen, K. Bernardin, H. K. Ekenel, J. McDonough, K. Nickel, M. Voit und M. Woelfel: *Audio-Visual Perception of a Lecturer in a Smart Seminar Room*. Signal Processing - Special Issue on Multimodal Interfaces, 86(12):3518–3533, 2006.

[SBEV08] R. Stiefelhagen, K. Bernardin, H. Ekenel und M. Voit: *Tracking Identities and Attention in Smart Environments - Contributions and Progress in the CHIL Project*. In: *Proc. 8th Int'l Conf. on Face and Gesture Recognition*, 2008.

[SEF+07] R. Stiefelhagen, H. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit und A. Waibel: *Enabling Multimodal Human-Robot Interaction for the Karlsruhe Humanoid Robot*. IEEE Transactions on Robotics - Special Issue on Human-Robot Interaction, 23(5):840–851, 2007.

[VGCF+09] M. Voit, N. Gourier, C. Canton-Ferrer, O. Lanz, R. Stiefelhagen und R. Brunelli: *Estimation of Head Pose*. In: A. Waibel und R. Stiefelhagen (Herausgeber): *Computers in the Human Interaction Loop*, Seiten 33–42. Springer, 2009.

[VNS05a] M. Voit, K. Nickel und R. Stiefelhagen: *Estimating the Lecturer's Head Pose in Seminar Scenarios - A Multi-View Approach*. In: *Proc. 2nd Int'l Workshop on Machine Learning for Multimodal Interaction*, Seiten 230–240, 2005.

[VNS05b] M. Voit, K. Nickel und R. Stiefelhagen: *Multi-View Head Pose Estimation Using Neural Networks*. In: *Proc. 2nd Canadian Conf. on Computer and Robot Vision*, Seiten 347–352, 2005.

[VNS06] M. Voit, K. Nickel und R. Stiefelhagen: *Neural Network-Based Head Pose Estimation and Multi-View Fusion*. In: *Multimodal Technologies for Perception of Humans - Proc. First Int'l Evaluation Workshop on Classification of Events, Activities and Relationships*, Seiten 291–298, 2006.

[VNS08] M. Voit, K. Nickel und R. Stiefelhagen: *Head Pose Estimation in Single- and Multi-view Environments - Results on the CLEAR'07 Benchmarks*. In: *Multimodal Technologies for Perception of Humans - Proc. Int'l Evaluation Workshops CLEAR 2007 and RT 2007*, Seiten 307–316, 2008.

[VS06a] M. Voit und R. Stiefelhagen: *A Bayesian Approach for Multi-View Head Pose Estimation*. In: *Proc. Int'l Conf. on Multisensor Fusion and Integration for Intelligent Systems*, Seiten 31–34, 2006.

[VS06b] M. Voit und R. Stiefelhagen: *Tracking Head Pose and Focus of Attention with Multiple Far-Field Cameras*. In: *Proc. Int'l Conf. on Multimodal Interfaces*, Seiten 281–286, 2006.

[VS08a] M. Voit und R. Stiefelhagen: *Deducing the Visual Focus of Attention From Head Pose Estimation in Dynamic Multi-View Meeting Scenarios*. In: *Proc. Int'l Conf. on Multimodal Interfaces*, 2008.

[VS08b] M. Voit und R. Stiefelhagen: *Visual Focus of Attention in Dynamic Meeting Scenarios*. In: *Proc. 5th Int'l Workshop on Machine Learning for Multimodal Interaction*, Seiten 1–13, 2008.

[VS10] M. Voit und R. Stiefelhagen: *3D User-Perspective, Voxel-Based Estimation of Visual Focus of Attention in Dynamic Meeting Scenarios*. In: *Proc. 12th Int'l Conf. on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction*, 2010.

H. Literaturverzeichnis

[All87] A. Allport: *Selection for Action: Some Behavioral and Neurophysiological Considerations of Attention and Action*. In: H. Heuer und H. F. Sanders (Herausgeber): *Perspectives on Perception and Action*, Seiten 395–419. Erlbaum, Hillsdale, NJ, USA, 1987.

[AMGC02] M. S. Arulampalam, S. Maskell, N. Gordon und T. Clapp: *A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking*. IEEE Transactions on Signal Processing, 50(2):174–188, 2002.

[BAMPS98] J. Bruske, E. Abraham-Mumm, J. Pauli und G. Sommer: *Head-pose Estimation From Facial Images with Subspace Neural Networks*. In: *Proc. Int'l Neural Network and Brain Conf.*, Seiten 528–531, 1998.

[Bey94] D. J. Beymer: *Face Recognition under Varying Pose*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 756–761, 1994.

[BHO09] S. Ba, H. Hung und J. M. Odobez: *Visual Activity Context for Focus of Attention Estimation in Dynamic Meetings*. In: *Proc. Int'l Conf. on Multimedia and Expo*, Seiten 1424–1427, 2009.

[BI05] P. Baldi und L. Itti: *Attention: Bits versus Wows*. In: *Proc. IEEE Int'l Conf. on Neural Networks and Brain*, Seiten PL–56–PL–61, 2005.

[Bir97] S. Birchfield: *An Elliptical Head Tracker*. In: *Proc. 31st Asilomar Conf. on Signals, Systems, and Computers*, Seiten 1710–1714, 1997.

[Bir98] S. Birchfield: *Elliptical Head Tracking Using Intensity Gradients and Color Histograms*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 232–237, 1998.

[Ble64] W. W. Bledsoe: *The Model Method in Facial Recognition, Technical Report PRI 15*. Technischer Bericht, Panoramic Research, Inc., Palo Alto, California, 1964.

[BMR05] O. Brdiczka, J. Maisonnasse und P. Reignier: *Automatic Detection of Interaction Groups*. In: *Proc. 7th Int'l Conf. on Multimodal Interfaces*, Seiten 32–36, New York City, 2005.

[BMT08] T. Bader, A. Meissner und R. Tscherney: *Digital Map Table with Fovea-Tablett®: Smart Furniture for Emergency Operation Centers*. In: *Proc. 5th Int'l Conf. on Information Systems for Crisis Response and Management*, Seiten 679–688, 2008.

[BO04] S. O. Ba und J. M. Odobez: *A Probabilistic Framework for Joint Head Tracking and Pose Estimation*. In: *Proc. 17th Int'l Conf. on Pattern Recognition*, Seiten 264–267, 2004.

[BO05] S. O. Ba und J. M. Odobez: *A Rao-Blackwellized Mixed State Particle Filter for Head Pose Tracking in Meetings*. In: *Proc. 7th ACM-ICMI Workshop on Multimodal Multiparty Meeting Processing*, Seiten 9–16, 2005.

[BO06] S. O. Ba und J. M. Obodez: *A Study on Visual Focus of Attention Recognition From Head Pose in a Meeting Room*. In: *Proc. 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Seiten 75–87, 2006.

[BO07a] S. O. Ba und J. M. Odobez: *Head Pose Tracking and Focus of Attention Recognition Algorithms in Meeting Rooms*. In: *Multimodal Technologies for Perception of Humans - Proc. First Int'l Evaluation Workshop on Classification of Events, Activities and Relationships*, Seiten 345–357, 2007.

[BO07b] S. O. Ba und J. M. Odobez: *Probabilistic Head Pose Tracking Evaluation in Single and Multiple Camera Setups*. In: *Multimodal Technologies for Perception of Humans - Proc. Int'l Evaluation Workshops CLEAR 2007 and RT 2007*, Seiten 276–286, 2007.

[BO08a] S. O. Ba und J. M. Obodez: *Visual Focus of Attention Estimation From Head Pose Posterior Probability Distributions*. In: *Proc. of IEEE Int'l Conf. on Multimedia and Expo*, Seiten 53–56, 2008.

[BO08b] S. O. Ba und J. M. Odobez: *Multi-Party Focus of Attention Recognition in Meetings From Head Pose and Multimodal Contextual Cues*. In: *Proc. Int'l Conf. on Acoustics, Speech, and Signal Processing*, 2008.

[BO08c] S. O. Ba und J. M. Odobez: *Multi-Person Visual Focus of Attention From Head Pose and Meeting Contextual Cues*. Technischer Bericht, IDIAP Research Institute, 2008.

[BO09] S. O. Ba und J. M. Odobez: *Recognizing Visual Focus of Attention From Head Pose in Natural Meetings*. IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics, 39(1):16–33, 2009.

[BO11] S.O. Ba und J. M. Odobez: *Multi-Person Visual Focus of Attention From Head Pose and Meeting Contextual Cues*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33:101–116, 2011.

[Bou04] J. Y. Bouget: *Camera Calibration Toolbox for Matlab*. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html, 2004. Stand: 23.09.2011.

[BR08] B. Benfold und I. Reid: *Colour Invariant Head Pose Classification in Low Resolution Video*. In: *Proc. 19th British Machine Vision Conf.*, 2008.

[Bra00] G. Bradski: *The OpenCV Library*. Dr. Dobb's Journal of Software Tools, 25(11):120, 122–125, 2000.

[Bro71] D. C. Brown: *Close-Range Camera Calibration*. Photogrammetric Engineering, 37(8):855–866, 1971.

[Bro01] L. M. Brown: *3D Head Tracking Using Motion Adaptive Texture-Mapping*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 998–1003, 2001.

[BT02] L. M. Brown und Y. L. Tian: *Comparative Study of Coarse Head Pose Estimation*. In: *Proc. IEEE Workshop on Motion and Video Computing*, Seiten 125–130, 2002.

[Bux03] H. Buxton: *Learning and Understanding Dynamic Scene Activity: A Review*. Image and Vision Computing, 21:125–136, 2003.

[CF09] C. Canton-Ferrer: *Human Motion Capture with Scalable Body Models*. Dissertation, Universitat Politecnica de Catalunya, 2009.

[CFCP06] C. Canton-Ferrer, J. R. Casas und M. Pardas: *Head Pose Detection Based on Fusion of Multiple Viewpoint Information*. In: *Multimodal Technologies for Perception of Humans - Proc. First Int'l Evaluation Workshop on Classification of Events, Activities and Relationships*, Seiten 305–310, Southampton, 2006.

[CFCP07] C. Canton-Ferrer, J. R. Casas und M. Pardas: *Head Orientation Estimation using Particle Filtering in Multiview Scenarios*. In: *Multimodal Technologies for Perception of Humans - Proc. Int'l Evaluation Workshops CLEAR 2007 and RT 2007*, Seiten 317–324, 2007.

[CFSP+08] C. Canton-Ferrer, C. Segura, M. Pardas, J. R. Casas und J. Hernando: *Multimodal Real-Time Focus of Attention Estimation in SmartRooms*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 1–8, 2008.

[CHF+05] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang und F. Quek: *A Multimodal Analysis of Floor Control in Meetings*. In: *Proc. Workshop on Machine Learning for Multimodal Interaction*, Seiten 36–49, 2005.

[CZH+03] L. Chen, L. Zhang, Y. Hu, M. Li und H. Zhang: *Head Pose Estimation Using Fisher Manifold Learning*. Proc. IEEE Int'l Workshop on Analysis and Modeling of Faces and Gestures, Seiten 203–207, 2003.

[DE82] J. F. Dovidio und S. L. Ellyson: *Decoding Visual Dominance: Attributions of Power Based on Relative Percentages of Looking while Speaking and Looking while Listening*. Social Psychology Quarterly, 45:106–113, 1982.

[DGA00] A. Doucet, S. Godsill und C. Andrieu: *On Sequential Monte Carlo Sampling Methods for Bayesian Filtering*. Statistics and Computing, 10(3):197–208, 2000.

[DH04] H. Dee und D. Hogg: *Detecting Inexplicable Behaviour*. In: *Proc. British Machine Vision Conf.*, Seiten 477–486, 2004.

[DMP96] T. Darrell, B. Maghaddam und A. P. Pentland: *Active Face Tracking and Pose Estimation in an Interactive Room*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 67–72, 1996.

[DT05] N. Dalal und B. Triggs: *Histograms of Oriented Gradients for Human Detection*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 886–893, 2005.

[EBMM03] A. A. Efros, A. C. Berg, G. Mori und J. Malik: *Recognizing Action at a Distance*. In: *Proc. 9th IEEE Int'l Conf. on Computer Vision*, Seiten 726–733, 2003.

[FB86] J. G. Fryer und D. C. Brown: *Lens Distortion for Close-Range Photogrammetry*. Photogrammetric Engineering and Remote Sensing, 52(1):51–58, 1986.

[FHY09] G. Friedland, H. Hung und C. Yeo: *Multi-Modal Speaker Diarization of Real-World Meetings Using Compressed-Domain Video Features*. In: *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing*, Seiten 4069–4072, 2009.

[FS08] E. G. Freedman und D. L. Sparks: *Eye-Head Coordination During Head-Unrestrained Gaze Shifts in Rhesus Monkeys*. Journal of Neurophysiology, 77(5):2328–2348, 2008.

[FTB+09] M. Farenzena, A. Tavano, L. Bazzani, D. Tosato, G. Paggetti, G. Menegaz, V. Murino und M. Cristani: *Social Interactions by Visual Focus Of Attention in a*

Three-Dimensional Environment. In: *Proc. 11th Int'l Conf. of the Italian Association for Artificial Intelligence*, 2009.

[GBBO09] G. Garau, S. Ba, H. Bourlard und J. M. Odobez: *Investigating the Use of Visual Focus of Attention for Audio-Visual Speaker Diarisation*. In: *Proc. 17th ACM Int'l Con. on Multimedia*, Seiten 681–684, 2009.

[GC94] A. Gee und R. Cipolla: *Determining the Gaze of Faces in Images*. Image and Vision Computing, 12(March):639–647, 1994.

[GCSR03] A. Gelman, J. B. Carlin, H. S. Stern und D. B. Rubin: *Bayesian Data Analysis*. CRC Press, 2003.

[GHC04] N. Gourier, D. Hall und J. L. Crowley: *Estimating Face Orientation From Robust Detection of Salient Facial Structures*. In: *Proc. Pointing 2004 - International Workshop on Visual Observation of Deictic Gestures*, Seite 17–25, 2004.

[GP06] D. Gatica-Perez: *Analyzing Group Interactions in Conversations: a Review*. In: *Proc. IEEE Int'l Conf. on Multisensor Fusion and Integration for Intelligent Systems*, Seiten 41–46, 2006.

[GPMB05] D. Gatica-Perez, I. McCowan und S. Bengio: *Detecting Group Interest-Level in Meetings*. In: *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, Seiten 489–492, 2005.

[GPZB05] D. Gatica-Perez, D. Zhang und S. Bengio: *Extracting Information From Multimedia Meeting Collections*. In: *Proc. 7th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval*, Seiten 245–252, 2005.

[GSRL98] W. E. L. Grimson, C. Stauffer, R. Romano und L. Lee: *Using Adaptive Tracking to Classify and Monitor Activities in a Site*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 2, 1998.

[Gur97] K. Gurney: *An Introduction to Neural Networks*. Routledge, London, United Kingdom, 1997.

[HF08] H. Hung und G. Friedland: *Towards Audio-Visual On-line Diarization Of Participants In Group Meetings*. In: *Proc. Workshop on Multi-Camera and Multi-Modal Sensor Fusion Algorithms and Applications in conjunction with ECCV*, 2008.

[HGP08] H. Hung und D. Gatica-Perez: *Identifying Dominant People in Meetings From Audio-Visual Sensors*. In: *Proc. 8th IEEE Int'l Conf. on Automatic Face & Gesture Recognition*, Seiten 1–6, 2008.

[HGP10] H. Hung und D. Gatica-Perez: *Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior*. IEEE Transactions on Multimedia, 12(6):563–575, 2010.

[HGPHF08] H. Hung, D. Gatica-Perez, Y. Huang und G. Friedland: *Estimating The Dominant Person In Multi-Party Conversations Using Speaker Diarization Strategies*. In: *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, Seiten 2197–2200, 2008.

[HGSR06] M. W. Hoffman, D. B. Grimes, A. P. Shon und R. P. N. Rao: *A Probabilistic Model of Gaze Imitation and Shared Attention*. Neural Networks - 2006 Special issue: The brain mechanisms of imitation learning, 19:299–310, 2006.

[HHGP08] H. Hung, Y. Huang und D. Gatica-Perez: *Associating Audio-Visual Activity Cues in a Dominance Estimation Framework*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Seiten 1–6, 2008.

[HJB+08] H. Hung, D. B. Jayagopi, S. O. Ba, J. M. Odobez und D. Gatica-Perez: *Investigating Automatic Dominance Estimation in Groups From Visual Attention and Speaking Activity*. In: *Proc. 10th Int'l Conf. on Multimodal Interfaces*, Seiten 233–236, 2008.

[HJY+07] H. Hung, D. B. Jayagopi, C. Yeo, G. Friedland, S. O. Ba, J. M. Odobez, K. Ramchandran, N. Mirghafori und D. Gatica-Perez: *Using Audio and Video Features to Classify the Most Dominant Person in a Group Meeting*. In: *Proc. 15th International Conf. on Multimedia*, Seiten 835–838, 2007.

[Hor91] K. Hornik: *Approximation Capabilities of Multilayer Feedforward Networks*. Neural Networks, 4:251–257, 1991.

[HS94] R. I. Hartley und P. Sturm: *Triangulation*. In: *Proc. ARPA Image Understanding Workshop*, Seiten 957–966, 1994.

[HS97] J. Heikkilä und O. Silvén: *A Four-step Camera Calibration Procedure with Implicit Image Correction*. In: *Proc. Conf. on Computer Vision and Pattern Recognition*, Seiten 1106–1112, 1997.

[HU90] D. P. Huttenlocher und S. Ullman: *Recognizing Solid Objects by Alignment with an Image*. International Journal of Computer Vision, 5(2):195–212, 1990.

[HYD96] T. Horprasert, Y. Yacoob und L. S. Davis: *Computing 3-D Head Orientation From a Monocular Image Sequence*. In: *Proc. 2nd Int'l Conf. on Automatic Face and Gesture Recognition*, Seiten 242–247, 1996.

[HZ98] J. Heinzmann und A. Zelinsky: *3-D Facial Pose and Gaze Point Estimation Using a Robust Real-Time Tracking Paradigm*. In: *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, Seiten 142–147, Nara, 1998.

[HZ03] R. Hartley und A. Zisserman: *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[IB98] M. Isard und A. Blake: *Condensation - Conditional Density Propagation for Visual Tracking*. International Journal of Computer Vision, 29(1):5–28, 1998.

[IB05] L. Itti und P. Baldi: *A Principled Approach to Detecting Surprising Events in Video*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 631–637, 2005.

[IB09] L. Itti und P. Baldi: *Bayesian Surprise Attracts Human Attention*. Vision research, 49(10):1295–1306, 2009.

[IK01] L. Itti und C. Koch: *Computational Modelling of Visual Attention*. Nature reviews Neuroscience, 2(3):194–203, 2001.

[IS10] J. IJsselmuiden und R. Stiefelhagen: *Towards High-Level Human Activity Recognition through Computer Vision and Temporal Logic*. In: *Proc. 33rd Annual German Conf. on Advances in Artificial Intelligence*, Seiten 426–435, 2010.

[Jay03] E. T. Jaynes: *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[JH95] N. Johnson und D. Hogg: *Learning the Distribution of Object Trajectories for Event Recognition*. In: *Proc. British Machine Vision Conf.*, Seiten 583–592, 1995.

[JHYGP08] D. B. Jayagopi, H. Hung, C. Yeo und D. Gatica-Perez: *Predicting the Dominant Clique in Meetings Through Fusion of Nonverbal Cues*. In: *Proc. 16th ACM Int'l Conf. on Multimedia*, Seiten 809–812, 2008.

[JHYGP09] D. B. Jayagopi, H. Hung, C. Yeo und D. Gatica-Perez: *Modeling Dominance in Group Conversations Using Nonverbal Activity Cues*. IEEE Transactions on Audio, Speech, and Language Processing, 17(3):501–513, 2009.

[KBS00] V. Kruger, S. Bruns und G. Sommer: *Efficient Head Pose Estimation with Gabor Wavelet Networks*. In: *Proc. British Machine Vision Conf.*, Seiten 12–14, 2000.

[KHDM98] J. Kittler, M. Hatef, R. P. W. Duin und J. Matas: *On Combining Classifiers*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20:226–239, 1998.

[KP03] H. L. Kundel und M. Polansky: *Measurement of Observer Agreement*. Radiology, 228(2):303–308, 2003.

[KSS+06] Y. Kobayashi, D. Sugimura, Y. Sato, K. Hirasawa, N. Suzuki, H. Kage und A. Sugimoto: *3D Head Tracking using the Particle Filter with Cascaded Classifiers*. In: *Proc. 17th British Machine Vision Conf.*, Seiten 37–46, 2006.

[LB07] O. Lanz und R. Brunelli: *Joint Bayesian Tracking of Head Location and Pose From Low-Resolution Video*. In: *Multimodal Technologies for Perception of Humans - Proc. Int'l Evaluation Workshops CLEAR 2007 and RT 2007*, Seiten 287–296, 2007.

[LGSL00] Y. Li, S. Gong, J. Sherrah und H. Liddell: *Multi-View Face Detection Using Support Vector Machines and Eigenspace Modelling*. In: *Proc. of Int'l Conf. on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, Seite 241–244, 2000.

[LWB00] S. R. H. Langton, R. J. Watt und V. Bruce: *Do the Eyes Have It? Cues to the Direction of Social Attention*. Trends in Cognitive Sciences, 4(2):50–59, 2000.

[MBGP+03] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner und H. Bourlard: *Modeling Human Interaction in Meetings*. In: *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, Seiten 748–751, 2003.

[MCDT07] E. Murphy-Chutorian, A. Doshan und M. M. Trivedi: *Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation*. In: *Proc. IEEE Conf. on Intelligent Transportation Systems*, Seiten 709–714, 2007.

[MCT08a] E. Murphy-Chutorian und M. M. Trivedi: *3D Tracking and Dynamic Analysis of Human Head Movements and Attentional Targets*. In: *Proc. IEEE Int'l Conf. on Distributed Smart Cameras*, Seiten 1–8, California, 2008.

[MCT08b] E. Murphy-Chutorian und M. M. Trivedi: *HyHOPE: Hybrid Head Orientation and Position Estimation for Vision-based Driver Head Tracking*. In: *Proc. IEEE Intelligent Vehicles Symposium*, Seite 512–517, 2008.

[MCT09] E. Murphy-Chutorian und M. M. Trivedi: *Head Pose Estimation in Computer Vision: A Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(4):607–626, 2009.

[ME02] D. Makris und T. Ellis: *Spatial and Probabilistic Modelling of Pedestrian Behaviour*. In: *Proc. British Machine Vision Conf.*, Seiten 557–566, 2002.

[MGPB+05] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard und D. Zhang: *Automatic Analysis of Multimodal Group Actions in Meetings*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(3):305–317, 2005.

[NF96] S. Niyogi und W. T. Freeman: *Example-Based Head Tracking*. In: *Proc. 2nd Int'l Conf. on Automatic Face and Gesture Recognition*, Seiten 374–378, 1996.

[NG99] J. Ng und S. Gong: *Multi-View Face Detection and Pose Estimation Using A Composite Support Vector Machine across the View Sphere*. In: *Proc. Int'l Workshop on Recognition, Analsysis, and Tracking of Faces and Gestures*, Seiten 14–21, 1999.

[NI05] V. Navalpakkam und L. Itti: *Modeling the Influence of Task on Attention*. Vision research, 45(2):205–231, 2005.

[NI06] V. Navalpakkam und L. Itti: *Top-Down Attention Selection is Fine Grained*. Journal of Vision, 6(11):1180–1193, 2006.

[Nic08] K. Nickel: *Visuelle Benutzermodellierung mit Tracking und Zeigegestenerkennung für einen humanoiden Roboter*. Dissertation, Universität Karlsruhe, 2008.

[NP00] A. Nikolaidis und I. Pitas: *Facial Feature Extraction and Determination of Pose*. Pattern Recognition, 33(11):1783–1791, 2000.

[NVI10] NVIDIA Corporation: *NVIDIA CUDA: Compute Unified Device Architecture - Programming Guide Version 2.0*, 2010.

[OB07] J. M. Odobez und S. O. Ba: *A Cognitive and Unsupervised MAP Adaptation Approach to the Recognition of Focus of Attention From Head Pose*. In: *Proc. Int'l Conf. on Multimedia and Expo*, Seiten 1379–1382, 2007.

[OB09] K. Oberauer und S. Bialkova: *Accessing Information in Working Memory: can the Focus of Attention Grasp Two Elements at the Same Time?* Journal of Experimental Psychology: General, 138(1):64–87, 2009.

[OGX09] J. Orozco, S. Gong und T. Xiang: *Head Pose Classification in Crowded Scenes*. In: *Proc. British Machine Vision Conf.*, Seiten 1–11, 2009.

[OHG02] N. Oliver, E. Horvitz und A. Garg: *Layered Representations for Human Activity Recognition*. In: *Proc. 4th IEEE Int'l Conf. on Multimodal Interfaces*, Seiten 3–8, 2002.

[OOF+05] J. Ou, L. M. Oh, S. R. Fussell, T. Blum und J. Yang: *Analyzing and Predicting Focus of Attention in Remote Collaborative Tasks*. In: *Proc. 7th Int'l Conf. on Multimodal Interfaces*, Seiten 116–123, 2005.

[OTYH05] K. Otsuka, Y. Takemae, J. Yamato und M. Hiroshi: *A Probabilistic Inference of Multiparty-Conversation Structure Based on Markov-Switching Models of Gaze Patterns, Head Directions, and Utterances*. In: *Proc. 7th Int'l Conf. on Multimodal Interfaces*, Seiten 191–198, 2005.

[OYTM06] K. Otsuka, J. Yamato, Y. Takemae und H. Murase: *Conversation Scene Analysis with Dynamic Bayesian Network Based on Visual Head Tracking*. In: *Proc. 2006 IEEE Int'l Conf. on Multimedia and Expo*, Seiten 949–952, 2006.

[PA00] S. Park und J. K. Aggarwal: *Head Segmentation and Head Orientation in 3D Space for Pose Estimation of Multiple People*. In: *Proc. 4th IEEE Southwest Symposium on Image Analysis and Interpretation*, Seiten 192–196, 2000.

[PB98] R. Pappu und P. A. Beardsley: *A Qualitative Approach to Classifying Gaze Direction*. In: *Proc. 3rd Int'l Conf. on Automatic Face and Gesture Recognition*, Seite 160–165, Nara, 1998.

[Pos80] M. I. Posner: *Orienting of Attention*. The Quarterly Journal of Experimental Psychology, 32(1):3–25, 1980.

[PV05] P. Pérez und J. Vermaak: *Bayesian Tracking with Auxiliary Discrete Processes. Application to Detection and Tracking of Objects with Occlusions*. In: *Proc. IEEE ICCV Workshop on Dynamical Vision*, Seiten 190–202, Santa Barbara, 2005.

[PZ07] G. Potamianos und Z. Zhang: *A Joint System for Single-Person 2D-Face and 3D-Head Tracking in CHIL Seminars*. In: *Multimodal Technologies for Perception of Humans - Proc. Int'l Evaluation Workshops CLEAR 2007 and RT 2007*, Seiten 105–118, 2007.

[Res] SR Research: *EyeLink II*. http://www.sr-research.com/EL_II.html. Stand: 23.09.2011.

[RH05] R. Rienks und D. Heylen: *Dominance Detection in Meetings Using Easily Obtainable Features*. In: H. Bourlard und S. Renals (Herausgeber): *Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Seiten 76–86, 2005.

[RHE03] R. Ruddarraju, A. Haro und I. A. Essa: *Fast Multiple Camera Head Pose Tracking*. In: *Proc. Int'l Conf. on Vision Interfaces*, Halifax, 2003.

[RR98] R. Rae und H. J. Ritter: *Recognition of Human Head Orientation Based on Artificial Neural Networks*. IEEE Transactions on Neural Networks, 9(2):257–65, 1998.

[RR06] N. Robertson und I. Reid: *Estimating Gaze Direction From Low-Resolution Faces in Video*. In: *Proc. European Conf. on Computer Vision*, Seiten 402–415, 2006.

[RS61] H. Raiffa und R. Schlaifer: *Applied Statistical Decision Theory*. Harvard University, 1961.

[Sav72] L. J. Savage: *The Foundations of Statistics*. Dover Publications, 1972.

[SB02] S. Srinivasan und K. L. Boyer: *Head Pose Estimation Using View Based Eigenspaces*. In: *Proc. 16th Int'l Conf. on Pattern Recognition*, Seiten 302–305, 2002.

[SBB+07] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel und J. Garofolo: *The CLEAR 2007 Evaluation*. In: *Multimodal Technologies for Perception of Humans - Proc. Int'l Evaluation Workshops CLEAR 2007 and RT 2007*, Seiten 3–34, 2007.

[SBE+06] R. Stiefelhagen, K. Bernardin, H. K. Ekenel, J. McDonough, K. Nickel, M. Voit und M. Woelfel: *Audio-Visual Perception of a Lecturer in a Smart Seminar Room*. Signal Processing - Special Issue on Multimodal Interfaces, 86, 2006.

[SBGPO06] K. Smith, S. O. Ba, D. Gatica-Perez und J. M. Odobez: *Tracking the Multi Person Wandering Visual Focus of Attention*. In: *Proc. 8th Int'l Conf. on Multimodal Interfaces*, Seiten 265–272, New York, New York, USA, 2006.

[SBGPO07] K. Smith, S. O. Ba, D. Gatica-Perez und J. M. Odobez: *Tracking Attention for Multiple People: Wandering Visual Focus of Attention Estimation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30:1212–1229, 2007.

[Sch73] H. Schmidtke: *Ergonomie 1 - Grundlagen menschlicher Arbeit und Leistungen*. Carl Hanser Verlag, 1973.

[SFYW98] R. Stiefelhagen, M. Finke, J. Yang und A. Waibel: *From Gaze to Focus of Attention*. In: *Proc. Workshop on Perceptual User Interfaces*, Seiten 25–30, 1998.

[SFYW99] R. Stiefelhagen, M. Finke, J. Yang und A. Waibel: *From Gaze To Focus Of Attention*. In: *Proc. 3rd Int'l Conf. on Visual Information Systems*, 1999.

[SG00] C. Stauffer und W. E. L. Grimson: *Learning Patterns of Activity Using Real-Time Tracking*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):747–757, 2000.

[SG10] R. Stiefelhagen und J. Geisler: *Der Smart-Control-Room*. Jahresbericht 2009/2010, Fraunhofer-Insitut für Informations- und Datenverarbeitung IITB, 2010.

[SK00] H. Schneiderman und T. Kanade: *A Statistical Method for 3D Object Detection Applied to Faces and Cars*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 2–7, 2000.

[SMW+03] M. Siracusa, L. P. Morency, K. Wilson, J. Fisher und T. Darrell: *A Multi-Modal Approach for Determining Speaker Location and Focus*. In: *Proc. 5th Int'l Conf. on Multimodal interfaces*, 2003.

[SNS04] E. Seemann, K. Nickel und R. Stiefelhagen: *Head Pose Estimation Using Stereo Vision For Human-Robot Interaction*. In: *Proc. 6th Int' Conf. on Automatic Face and Gesture Recognition*, Seiten 626–631, 2004.

[SRF10] B. Schauerte, J. Richarz und G. A. Fink: *Saliency-Based Identification and Recognition of Pointed-At Objects*. In: *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, Seiten 4638–4643, 2010.

[Sti02] R. Stiefelhagen: *Tracking and Modeling Focus of Attention in Meetings*. Dissertation, Universität Karlsruhe, 2002.

[Sti04] R. Stiefelhagen: *Estimating Head Pose with Neural Networks-Results on the Pointing04 ICPR Workshop Evaluation Data*. In: *Proc. Pointing 2004 - International Workshop on Visual Observation of Deictic Gestures*, 2004.

[SvdCIS09] A. Schick, F. van de Camp, J. IJsselmuiden und R. Stiefelhagen: *Extending Touch: Towards Interaction with Large-scale Surfaces*. In: *Proc. ACM Int'l Conf. on Interactive Tabletops and Surfaces*, Seiten 127–134, 2009.

[SYW98] R. Stiefelhagen, J. Yang und A. Waibel: *Towards Tracking Interaction Between People*. In: *Proc. AAAI Spring Symposium on Intelligent Environments*, Seiten 123–127, 1998.

[SYW01a] R. Stiefelhagen, J. Yang und A. Waibel: *Estimating Focus of Attention Based on Gaze and Sound*. In: *Proc. 2001 Workshop on Perceptive User Interfaces*, Seiten 1–9, 2001.

[SYW01b] R. Stiefelhagen, J. Yang und A. Waibel: *Tracking Focus of Attention for Human-Robot Communication*. In: *Proc. IEEE-RAS Int'l Conf. on Humanoid Robots*, 2001.

[TBC+03] Y. Tian, L. Brown, J. Connell, S. Pankanti, A. Hampapur, A. Senior und R. Bolle: *Absolute Head Pose Estimation From Overhead Wide-Angle Cameras*. In: *Proc. IEEE Int'l Workshop on Analysis and Modeling of Faces and Gestures*, Seiten 92–99, 2003.

[Tsa86] R. Y. Tsai: *An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 364–374, 1986.

[Tsa87] R. Y. Tsai: *A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses*. IEEE Journal of Robotics and Automation, 3:323–344, 1987.

[vdCVS10] F. van de Camp, M. Voit und R. Stiefelhagen: *Efficient Person Identification Using Active Cameras in a Smartroom*. In: *Proc. Workshop on Multimodal Pervasive Video Analysis*, Seiten 17–22, 2010.

[VJ01] P. Viola und M. Jones: *Robust Real-time Object Detection*. International Journal of Computer Vision, 57:137–154, 2001.

[vTTBE05] K. van Turnhout, J. Terken, I. Bakx und B. Eggen: *Identifying the Intended Addressee in Mixed Human-Human and Human-Computer Interaction From Non-Verbal Features*. In: *Proc. 7th Int'l Conf. on Multimodal Interfaces*, Seiten 175–182, 2005.

[WS05] M. T. Wenzel und W. H. Schiffmann: *Head Pose Estimation of Partially Occluded Faces*. In: *Proc. 2nd Canadian Conf. on Computer and Robot Vision*, Seiten 353–360, Victoria, 2005.

[WS09] A. Waibel und R. Stiefelhagen (Herausgeber): *Computers in the Human Interaction Loop*. Springer, 2009.

[WT00] Y. Wu und K. Toyama: *Wide-Range, Person- and Illumination-Insensitive Head Orientation Estimation*. In: *Proc. 4th IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, Seiten 183–188, 2000.

[WW92] A. Watt und M. Watt: *Advanced Animation and Rendering Techniques*. Addison-Wesley Professional, 1992.

[WWLC00] H. R. Wilson, F. Wilkinson, L. M. Lin und M. Castillo: *Perception of Head Orientation*. Vision Research, 40(5):459–472, 2000.

[YZF+08] S. Yan, Z. Zhang, Y. Fu, Y. Hu, J. Tu und T. Huang: *Learning a Person-Independent Representation for Precise 3D Pose Estimation*. In: *Multimodal Technologies for Perception of Humans - Proc. Int'l Evaluation Workshops CLEAR 2007 and RT 2007*, Seiten 297–306, 2008.

[ZGPB+04] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan und G. Lathoud: *Modeling Individual and Group Actions in Meetings: A Two-Layer HMM Framework*. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition - Workshop on Event Mining in Video*, Seiten 117–125, 2004.

[ZGPBM06] D. Zhang, D. Gatica-Perez, S. Bengio und I. McCowan: *Modeling Individual and Group Actions in Meetings with Layered HMMs*. IEEE Transactions on Multimedia, 8(3):509–520, 2006.

[Zha07] G. P. Zhang: *Avoiding Pitfalls in Neural Network Research*. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 37(1):3–16, 2007.

[ZHLH06] Z. Zhang, Y. Hu, M. Liu und T. Huang: *Head Pose Estimation in Seminar Room using Multi View Face Detectors*. In: *Multimodal Technologies for Perception of Humans - Proc. First Int'l Evaluation Workshop on Classification of Events, Activities and Relationships*, Seiten 299–304, 2006.

[ZPC02] L. Zhao, G. Pingali und I. Carlbom: *Real-Time Head Orientation Estimation Using Neural Networks*. Proc. Int'l Conf. on Image Processing, Seiten 297–300, 2002.

[ZSA09] X. Zabulis, T. Sarmis und A. A. Argyros: *3D Head Pose Estimation From Multiple Distant Views*. In: *Proc. British Machine Vision Conf.*, London, 2009.

[ZZLZ02] Z. Zhang, L. Zhu, S. Z. Li und H. Zhang: *Real-Time Multi-View Face Detection*. In: *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, Seiten 142–147, 2002.