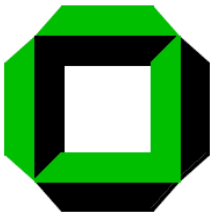

Universität Karlsruhe
Fakultät für Informatik
Institut für Logik, Komplexität und Deduktionssysteme
Prof. Dr. A. Waibel

Visuelle Schätzung der horizontalen Kopfdrehung in Multikameraumgebungen

Diplomarbeit

Michael Voit

Dezember 2004

Betreuer: Prof. Dr. A. Waibel
Dr. R. Stiefelhagen
Dipl.-Inform. K. Nickel

Hiermit versichere ich, dass die vorliegende Arbeit ohne fremde Hilfe erstellt, keine anderen als die angegebenen Quellen benutzt und die den benutzten Quellen wörtlich entnommenen Stellen als solche kenntlich gemacht wurden.

Michael Voit

Karlsruhe, 31. Dezember 2004

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

Das Bedürfnis die Umwelt menschlichen Eigenschaften anzupassen fördert eine Technisierung die über den aktuellen Zustand möglicher Mensch-Maschine-Interaktion hinausgeht und setzt aktuellen Arbeiten das Ziel Eingabemodalitäten der menschlichen Perzeption anzupassen. Die bislang fehlende Möglichkeit an den Bewusstseinszustand eines Interaktionspartners anzuknüpfen, hindert aktuelle Computersysteme daran, als vollwertige Kommunikationspartner angesehen zu werden. Von passiv in die Umgebung eines Benutzers integrierten Agenten erhofft man sich daher eine *ubiquitäre*[1] Modellierung dessen Bewusstseinszustands um auf Handlungen des Menschen entsprechend agieren und reagieren zu können.

Multimodale Benutzerschnittstellen werden durch die Verwendung mehrerer Eingabemodalitäten definiert, durch die ein Interaktionspartner mit einem Computersystem in Kontakt treten kann. Diese interdisziplinäre Form der Informationsübertragung bedient sich dabei unterschiedlicher Sensoren und vernetzt sonst autonome Systeme um die Vielzahl an anfallenden Aufgaben zu bewerkstelligen. Um solchen Systemen eine passive Rolle in der Umgebung ihrer Benutzer zuzuweisen, muss eine jederzeit mögliche Aussage über den aktuellen Kontext einer eventuellen Kontaktaufnahme abrufbar sein. Entsprechend dem menschlichen Vorbild wird versucht durch Sehen und Hören die Gleichstellung der Interaktionspartner zu gewährleisten.

Die Kognition und Verarbeitung der visuellen Wahrnehmung stellt heutige Computersysteme noch vor eine Vielzahl an zu lösenden Aufgaben. Objek-

[1] allgegenwärtig

te müssen erkannt, Bewegungen analysiert und Referenzierungen innerhalb eines Dialogs korrekt zugewiesen und berücksichtigt werden. Im Fall einer Mensch-Computer-Kommunikation muss das dahinterliegende System erkennen wann es angesprochen wird und welches Ziel der Kommunikationspartner durch seine Gestik und Mimik verfolgt.

1.2 Verwandte Arbeiten

Nach heutigem Wissensstand bietet die Blickrichtung eines Menschen einen wesentlichen Aufschluss über dessen Aufmerksamkeit [15]. Soll durch Computersehen die Blickrichtung einer Person ermittelt werden, so bieten sich intuitiv zwei Ansätze an: zum einen das Verfolgen der Pupillen, zum anderen das Modellieren der beobachteten Kopfdrehung der jeweiligen Person. Die für erstes Verfahren notwendige Aufnahme der Augen stellt zwar durch ihre hochauflösende Darstellung eine sehr gute Grundlage zur Erkennung der Blickrichtung dar, erfordert aber eine entsprechend nahe Positionierung der Kameras an den Kopf der zu beobachtenden Person, was eine diskrete Einbettung etwaiger Computersysteme in die Umwelt unmöglich werden lässt. In den letzten Jahren entstand daher eine Vielzahl an Arbeiten, deren gemeinsames Ziel darin bestand die Stabilität einer Beobachtung der Kopfdrehung weiter zu erhöhen. Im wesentlichen lassen sich dabei *modellbasierte* und *bildbasierte* Vorgehensweisen unterscheiden.

1.2.1 Modellbasierte Schätzung

Im allgemeinen verfolgt eine modellbasierte Schätzung das Auffinden mehrerer Gesichtsmerkmale (Augen, Nasenlöcher, Mundwinkel, usw.), deren relative Anordnung zueinander zu einer Bildung bzw. Verifizierung eines dreidimensionalen Kopfmodells ähnlichen Maßes genutzt werden [6], [23], [5].

Der Vorteil modellbasierter Hypothesenbildung liegt im gleichzeitigen Erkennen von Gesichtsverformungen und Bewegungen einzelner Gesichtsmerkmale [23]. Die Sichtbarkeit aller hierfür notwendigen Merkmale über jedes Kopfbild hinweg bildet dabei jedoch sowohl Voraussetzung als auch Nachteil dieses Ansatzes. Eine Anbringung mehrerer Kameras resultiert zwar in unterschiedlichen Ansichten und einem dadurch möglichen Erfassen derjenigen Merkmale die sonst verloren gehen würden, verhindert jedoch nicht, dass besonders feine Merkmale wie Augen oder Nasenlöcher bei ungünstiger Beleuchtung oder aussergewöhnlicher und schneller Kopfdrehung selbst durch den Einsatz mehrerer Kameras nur schwer verfolg- und erkennbar sind [20].

1.2.2 Bildbasierte Schätzung

Vorliegende Kopfbilder als Ganzes zu analysieren bzw. statt der Modellbildung das Bild als eigenes Objekt zur Schätzung zu benutzen wird durch den Ansatz der ansichtbasierten bzw. bildbasierten Schätzung beschrieben.

Der Einsatz neuronaler Netze zur Schätzung der Kopfdrehung wird unter anderem in [13], [19] sowie [21] vorgestellt. Die Schätzung erfolgt dabei durch eine von dem neuronalen Netz vorgenommene Funktionsapproximation auf dem Wertebereich der gewünschten Ausgabehypothesen.

Das in [17] beschriebene Verfahren nutzt sieben aus Beispielbildern unterschiedlicher Kopfdrehungen durch Hauptkomponentenanalyse erstellte Eigenräume, die jeweils eine unterschiedliche Drehklasse repräsentieren. Neue Kopfbilder werden anschließend auf den entsprechenden Eigenraum abgebildet, indem ihre Zugehörigkeit bezüglich eines Eigenraums durch ein Bewertungskriterium ermittelt wird. Die hierfür notwendige Diskretisierung der Drehungen in die Winkelklassen der jeweiligen Eigenräume, stellt jedoch den offensichtlichen Nachteil des Verfahrens durch dessen nur sehr grobe Schätzleistung dar.

Eine Klassifikation durch Implementierung einer Nächster-Nachbar-Suche wird in [10] vorgestellt: Kopfbilder werden auf 3×4-elementige Raster reduziert, indem für jedes Element der durchschnittliche Helligkeitswert aller von diesem Rasterbereich eingeschlossenen Pixel berechnet wird. Die so entstehenden, zwölfdimensionalen Vektoren werden zu einer Clusterbildung innerhalb des Merkmalraums genutzt. Neue Kopfbilder bzw. deren jeweilige Vektoren werden anschließend durch eine Nächster-Nachbar-Suche über alle Cluster des Merkmalraums bewertet und entsprechend zugewiesen.

1.2.3 Fazit

Den vorgestellten Verfahren ist allen gemein, dass dem Interaktionspartner Beschränkungen hinsichtlich seiner erlaubten Bewegung vorgeschrieben werden. Eine freie Bewegung im Raum ist nicht oder nur eingeschränkt möglich. Wünschenswert dagegen wäre die Erfassung jeglicher Kopfdrehungen an jeder beliebigen Position. Eine Anbringung mehrerer Kameras, so dass ein Raum von allen Seiten betrachtet und aufgenommen wird würde helfen auch Kopfansichten zu erfassen, die Einzel- oder Stereokamerasystemen sonst rückseitig zugewandt sind. Im Rahmen dieser Arbeit soll ein solches Verfahren weiter erörtert werden.

1.3 Aufgabenstellung

Ziel dieser Arbeit ist es ein System zu entwickeln, das zu jedem Zeitpunkt anhand von vier verschiedenen, zeitgleichen Kameraansichten einer beliebigen Person, eine Hypothese bezüglich deren Kopfdrehung abgibt.
Hierzu sind vier Kameras[2] in den oberen Ecken eines rechteckigen Raums, gemäß Abbildung 1.1 installiert. Die von ihnen übermittelten Bilder besitzen eine Auflösung von 640 auf 480 Pixel, ein Gaußfilter bewirkt eine zusätzliche Weichzeichnung. Die Pixel sind durch RGB-Farbwerte beschrieben. Abbildung 1.3 zeigt eine beispielhafte, aufgenommene Szene einer im Raum stehenden Person.

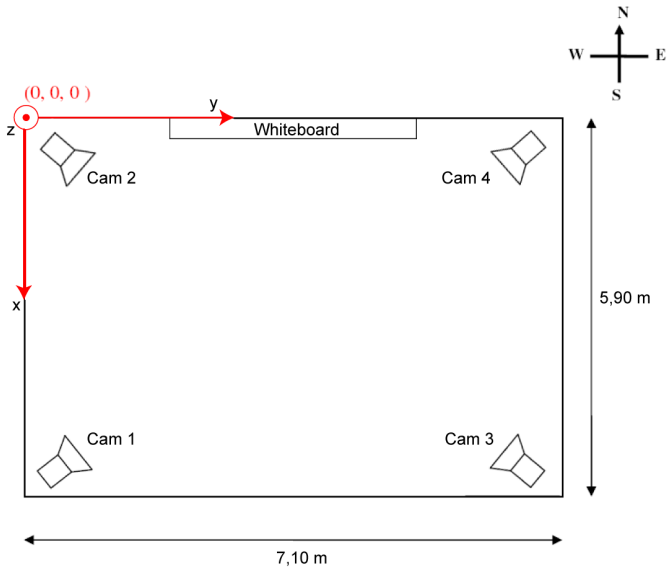

Abbildung 1.1. Die vier Kameras sind in den oberen Ecken des Raums installiert und erfassen damit jede Position an der Personen anzutreffen sein können.

Die tatsächlichen Kopfdrehungen werden mit einem magnetischen *Motion Capture System*[3] transkribiert um eine Überprüfung der vom System erstellten Hypothesen bezüglich der tatsächlich vorliegenden Kopfdrehung zu ermöglichen: Angebracht auf dem Kopf der zu beobachtenden Person, wird der Zustand eines Magnetsensors relativ zu einem zugehörigen Emitter bestimmt. Der Zustand ist definiert durch die Drehwinkel *pan*, *tilt* und *roll*

[2]Die Kameras sind jeweils vom Typ Sony DFW-V500
[3]Zum Einsatz kam das Motion Tracking System *flock of birds* der Firma *Ascension Technologies Corporation*

(siehe Abbildung 1.2), sowie die dreidimensionale Position x, y, und z in Abhängigkeit zur Position des Emitters.
Der Magnetsensor wird mit einem durchsichtigen Haarreif auf dem Kopf des entsprechenden Interaktionspartners befestigt. Während der Aufnahmen darf sich die Person frei im Raum bewegen.

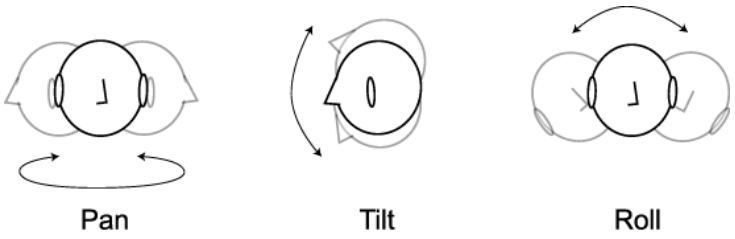

Abbildung 1.2. Die drei möglichen Drehwinkel pan, tilt und roll eines Kopfes.

Die geforderte Ausgabe des Systems soll dem horizontalen Drehwinkel[4] des erfassten Kopfes entsprechen. Der Winkel soll dabei in Bezug auf den in Abbildung 1.1 vorgestellten Raumaufbau mathematisch korrekt zwischen 0° und 360° angegeben werden.

1.4 Inhaltsübersicht

Das zu entwickelnde System lässt sich in fünf Stufen einteilen, deren Details in Kapitel 2 erläutert werden:

1. Kopfsuche
2. Vorverarbeitung
3. Vorderkopf-Hinterkopf-Segmentierung
4. Schätzen der einzelnen Dreh-Hypothesen
5. Fusion der Hypothesen zu einer endgültigen Schätzung

Die Kopfsuche in Abschnitt 2.2 befasst sich dabei mit der Aufgabe einen geeigneten Kopfkandidaten zur Drehschätzung in allen vorliegenden Kamerabildern zu finden und zu extrahieren. Die dabei beschriebene Vorgehensweise nutzt eine farbbasierte Segmentierung um Kopfkandidaten zu ermitteln die auf ihre Ähnlichkeit einer ellipsoiden Form hin untersucht werden. Der

[4]Der horizontale Drehwinkel entspricht dabei dem in Abbildung 1.2 aufgeführten Drehwinkel pan

zusätzliche Einsatz einer adaptiven Vordergrundsegmentierung ermöglicht dabei farbähnliche Hintergrundobjekte zu ignorieren und den Suchraum weiter einzuschränken.

Die anschließend auf den extrahierten Kopfbildern stattfindende Vorverarbeitung, welche in Abschnitt 2.3 erläutert wird, dient dazu ungünstige Lichtverhältnisse und Schattierungen zu minimieren indem die Kopfbilder in Helligkeitsbilder transformiert und im Kontrast normiert werden. Ferner werden vertikale als auch horizontale Kantenbilder erzeugt die der Drehschätzung zusätzliche Informationen bezüglich der vorliegenden Kopfdrehung bieten sollen.

Die vorverarbeiteten Kopfbilder werden durch die in Abschnitt 2.4 aufgeführte Vorderkopf-Hinterkopf-Klassifikation auf das Vorhandensein etwaiger Hinterkopfansichten geprüft und gegebenenfalls entfernt, so dass die eigentliche Drehschätzung in Abschnitt 2.5 lediglich aufgrund vorliegender Vorderkopfansichten geschieht. Durch Ignorieren jener Kopfansichten, die nur wenige Merkmale zur Bestimmung der Drehung aufweisen, soll die Stabilität des Systems erhöht werden. Die Klassifikation bezüglich Vorder- und Hinterköpfen und die eigentliche Drehschätzung erfolgen dabei durch Implementierung neuronaler Netze und beziehen sich auf jeweils jede Kameraansicht eines extrahierten Kopfes. Die so entstehenden, jeweiligen Einzelschätzungen pro vorliegender Kopfansicht, werden abschließend in Abschnitt 2.6 zu einer Gesamthypothese fusioniert und als Systemausgabe zurückgeliefert.

Eine Auflistung und Analyse ermittelter Ergebnisse des Vorderkopf-Hinterkopf-Klassifikators sowie der Schätzung der Drehhypothesen erfolgt in Kapitel 3. Unterschiedliche Topologien der eingesetzten Klassifikatoren werden evaluiert und auf ihren Einfluss auf den Gesamtfehler des Systems hin untersucht.

Abbildung 1.3. Die vier Kameraansichten einer Person.

Kapitel 2

Schätzen der Kopfdrehung

Im folgenden Kapitel soll das in dieser Arbeit entwickelte Vorgehen zur Schätzung der Kopfdrehung erläutert werden.
Das hierbei entworfene System folgt dabei durch den Einsatz *neuronaler Netze* dem bildbasierten Ansatz zur Hypothesenbildung der Kopfdrehung. Für den auf diesem Gebiet unerfahrenen Leser sei daher an dieser Stelle auf Anhang A hingewiesen, welcher eine kurze Einführung über die Grundlagen neuronaler Netze bietet.

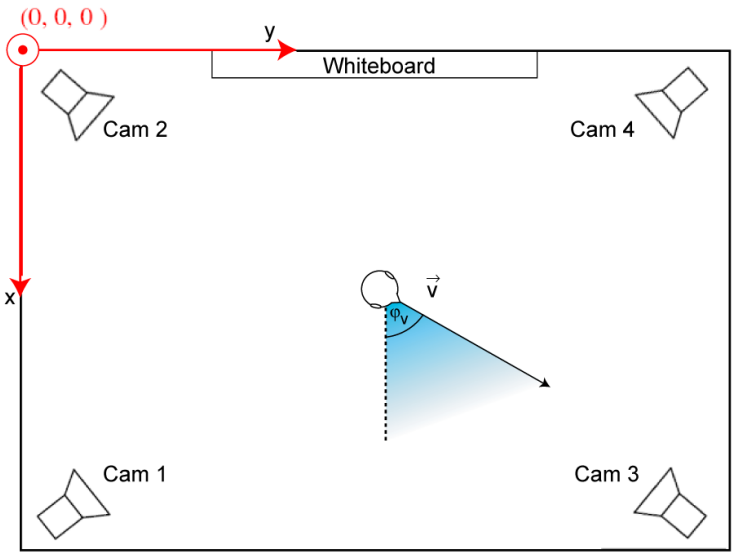

Abbildung 2.1. Die Aufgabe des Systems ist es anhand vier unterschiedlicher Kameraansichten eines Kopfes, dessen horizontalen Drehwinkel φ_v und damit dessen Blickrichtung $\vec{v}$ zu bestimmen.

2.1 Übersicht

Ziel dieser Arbeit ist ein System, das anhand vier verschiedener Ansichten des Kopfes eine Winkelhypothese über dessen *horizontale Drehung* abgibt. Der Winkel soll dabei durch einen Wertebereich von $[0°, 360°)$ mathematisch korrekt, entgegen dem Uhrzeigersinn beschrieben werden (siehe Abbildung 2.1).
Im ersten Schritt wird hierzu in den vorliegenden Kameraansichten der (entsprechende) Kopf gesucht und extrahiert. Das Finden von Kopfkandidaten erfolgt dabei durch die Erstellung eines Suchraums auf dem ein Ellipsen-Suchverfahren die elliptische Form des Kopfes approximiert. Die Erstellung des Suchraums geschieht durch eine farbbasierte Modellierung des zu verfolgenden Kopfes sowie einer adaptiven Vordergrund-Segmentierung. Letzteres Verfahren dient zur Erhöhung der Stabilität durch ein Erkennen statischer Objekte die zwar dem Kopf farbähnlich erscheinen aber dem Hintergrund zugewiesen werden können.

Die extrahierten Kopfbilder werden im zweiten Schritt vorverarbeitet um entsprechende Eingabemuster für die jeweiligen neuronalen Netze bereitzustellen. Die Vorverarbeitung beinhaltet dabei das Normieren des Kontrasts sowie das Extrahieren horizontaler und vertikaler Kanten der Kopfansicht.

Da Hinterköpfe nur mangelhafte Merkmale zur Schätzung der Kopfdrehung aufweisen, findet im dritten Schritt ein Ausfiltern etwaiger Hinterkopfansichten statt: Ein hierzu trainiertes neuronales Netz klassifiziert angelegte Kopfbilder als Vorder- bzw. Hinterkopf und entfernt damit Kameraansichten, die bei der darauffolgenden Hypothesenbildung der Kopfdrehung unangemessen wären.
Das Schätzen des Winkels selbst geschieht wie auch bei der Vorderkopf-Hinterkopf-Klassifikation durch Einsatz eines auf Kopfbildern trainierten neuronalen Netzes. Die Ausgabe des Netzwerks beschreibt dabei den auf dem angelegten Kopfbild gesehenen Drehwinkel und bildet damit eine Hypothese pro zur Verfügung stehender Kameraansicht. Die Erstellung einer finalen Drehangabe erfolgt in Schritt fünf durch Fusion der einzelnen Winkelschätzungen.

Abbildung 2.2. Die Hypothesenbildung bezüglich der horizontalen Kopfdrehung erfolgt in fünf Schritten.

2.2 Kopfextraktion

Um optimale Bedingungen für die Schätzung der Kopfdrehung anhand vier verschiedener Kameraansichten zu schaffen, muss der Kopf in allen vier Ansichten gefunden und extrahiert werden. Der folgende Abschnitt erläutert hierzu das in dieser Arbeit eingesetzte Verfahren einer Kombination aus farbbasierter Modellierung, adaptiver Vordergrund-Segmentierung und Ellipsen-Suche nach Kopfkandidaten über einem durch die beiden erstgenannten Verfahren erstellten Suchraum.

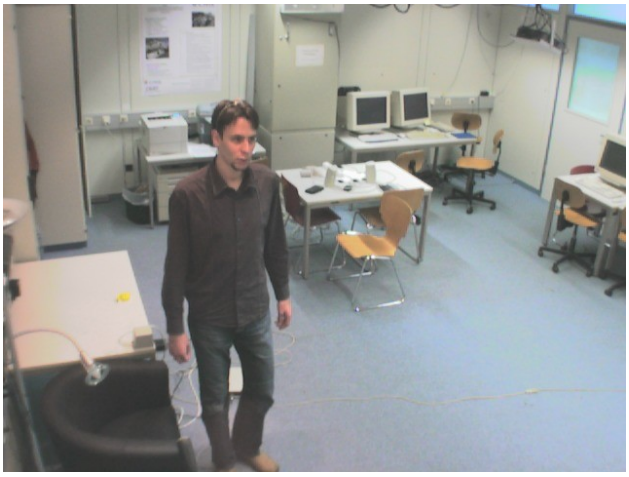

Abbildung 2.3. Kameraansicht einer Person deren Kopfdrehung geschätzt werden soll.

2.2.1 Farbmodellierung

Besteht, wie im Rahmen dieser Arbeit, völlige Bewegungsfreiheit für den entsprechenden Interaktionspartner, impliziert dies sowohl das Auftreten dessen Vorder- als auch Hinterkopfes in den aufgenommenen Bildsequenzen, was den Einsatz einer gern in ähnlichen Arbeiten beschriebenen Hautfarbsegmentierung zur Gesichtsverfolgung ausschließt ([8], [22], [11], [12]). Soll dennoch ein farbbasierter Ansatz zur Verfolgung eingesetzt werden, muss eine Erweiterung des Farbmodells auf Haarfarben stattfinden, so dass auch Bereiche als dem Kopf zugehörig klassifiziert werden, die keine direkten Gesichtsmerkmale aufweisen, was z.B. bei der Ansicht eines Hinterkopfes der Fall ist.

Das in der vorliegenden Arbeit zum Einsatz kommende Verfahren der farbbasierten Segmentierung nutzt statt eines globalen a-priori Farbmodells aller zu unterstützenden Personen ein Referenz-Histogramm des entsprechenden Interaktionspartners. Im Gegensatz zu parametrischen Modellen wie Gauß-

oder Gaußmischverteilungen, ermöglicht der hier gewählte Ansatz der nichtparametrischen Darstellung den Ausschluss von Farbwerten, die nicht explizit im zu verfolgenden Bild des Kopfes vorkommen. Ferner ist durch Einsatz eines Referenz-Histogramms mit der sogenannten *Histogramm-Rückprojektion*[1] ein weit verbreitetes, oft in der Objekterkennung eingesetztes Verfahren der Farbsegmentierung verfügbar.

Um innerhalb eines vorliegenden Bildes ein zu suchendes Objekt O aufzufinden, bestimmt die Histogramm-Rückprojektion das Verhältnis zwischen dem normalisierten Referenz-Histogramm H_O des zu suchenden Objekts[2] und dem Histogramm H_I des vorliegenden Bildes I, indem das jeweilige Verhältnis der entsprechenden Histogramm-Töpfe miteinander berechnet wird. Für beispielhafte, zweidimensionale Histogramme mit $n_1 \times n_2$ Töpfen gilt damit:

$$\forall i \in [0, n_1] \forall j \in [0, n_2] : H_R(i,j) = min\left(\frac{H_O(i,j)}{H_I(i,j)}, 1\right) \tag{2.1}$$

Die Zugehörigkeit eines Pixels im vorliegenden Bild zu dem gesuchten Objekt kann somit durch den entsprechenden Wert des jeweiligen Topfes (i, j) des Verhältnishistogramms H_R ausgedrückt werden, auf den das Pixel verweist. Ersetzt man jedes Pixel im Bild durch seine berechnete Farbähnlichkeit, entsteht daraus ein *Zugehörigkeitsbild* in Form eines Intensitätenbildes entsprechend Abb. 2.4. Dabei werden Pixel, deren Farbwerte gar nicht oder nur selten im gesuchten Objekt vorkommen unterdrückt, Pixel mit stark gesuchten Farbwerten entsprechend hervorgehoben.

Erstellen eines Referenz-Histogramms

Die Initialisierung des Referenz-Histogramms kann entweder vorab durch Erstellung eines a-priori Modells[3] geschehen, indem einzelne Bildausschnitte farbähnlicher Köpfe zur Modellbildung herangezogen werden, oder durch eine Modellbildung zur Laufzeit. Letztere Methode verspricht dabei ein Farbmodell des Kopfes zu generieren, das der jeweiligen, aktuellen Situation und Beleuchtung angepasst ist - das Verfolgen des Kopfes wird robuster.

Zur Anpassung an sich ändernde Umstände und zur Erfassung eines Modells über den kompletten Kopf wird in dieser Arbeit das Referenz-Histogramm

[1] engl. histogram backprojection

[2] Hier des zu suchenden Kopfes

[3] Der Einsatz einer Datenbank ist denkbar, in der Muster-Histogramme aller zu unterstützenden Personen gespeichert sind

Abbildung 2.4. Ergebnis einer Histogramm-Rückprojektion. Die dunkel eingefärbten Pixel weisen eine hinreichende Farbähnlichkeit zum gesuchten Kopf auf.

jeweils über einen beliebigen, aber festen Zeitrahmen vorhergehender Videobilder und nur mit den als Kopf klassifizierten Bildausschnitten erstellt und wiederholend aktualisiert. Farbwerte aus Kopfbildern die länger zurückliegen als das vorgegebene Zeitfenster werden aus dem Histogramm entfernt, neue, dem Kopf zugeordnete Pixelwerte dem Histogramm hinzugefügt: Die Folge ist eine baldige Neubildung des Farbmodells nach etwaiger Fehlklassifikation des Kopfes.
Für den hierfür zugrundeliegenden Farbraum wird der ursprüngliche RGB-Raum der Aufnahmen beibehalten.

2.2.2 Vordergrundsegmentierung

Hintergrundmodellierung ermöglicht es, bewegte Objekte als solche zu erkennen und von einem relativ statischen Hintergrund zu segmentieren.

In [18] wird ein Verfahren vorgestellt, welches in Echtzeit Pixel durch Gaußmischverteilungen beschreibt, um durch die Beobachtung der Parameteränderungen eine Klassifikation bezüglich der beiden Klassen *Hintergrund* oder *Vordergrund* durchzuführen.
Das in Abbildung 2.5 gezeigte Binärbild stellt die Ausgabe einer Implementierung des oben genannten Verfahrens im Rahmen dieser Arbeit dar. Durch eine Kombination mit dem durch Histogramm-Rückprojektion resultierenden Zugehörigkeitsbild der Farbsegmentierung soll im folgenden Abschnitt eine Suche nach möglichen Kopfkandidaten geschehen.

Abbildung 2.5. Resultierendes Binärbild der Vordergrund-Hintergrund-Segmentierung nach [18]. Hier schwarz eingefärbte Pixel wurden dabei dem Vordergrund zugewiesen, weiße Pixel dagegen dem Hintergrund. Durch die langsame Bewegung der beobachteten Person ziehen einige Pixel nach und wurden noch nicht korrekt als Hintergrund klassifiziert: ein schattenhaftes Nachziehen der Silhouette entsteht.

2.2.3 Suchen von Kopfkandidaten

Farbbasierte Verfolgung und Vordergrundsegmentierung liefern jeweils Intensitätenbilder, deren entsprechend als positiv klassifizierten Pixel einen Suchraum für Kopfkandidaten vorgeben.

Eine vereinfachende Annahme, dass alle Köpfe elliptische Form besitzen und im Rahmen dieser Arbeit stets aufgerichtet vorzufinden sind, ermöglicht es die Suche auf ein Ellipsen-Suchverfahren nach adäquaten Kopfkandidaten zu reduzieren [1]: Pixelcluster werden auf ihre Ähnlichkeit zu einer ellipsoiden Form hin untersucht und gemäß einer Gütefunktion bewertet.

Im folgenden wird ein Kopfkandidat einer Ellipse gleichgesetzt, deren Zustand durch $s = (x, y, \sigma, \tau)$ beschrieben werden kann: x definiert die horizontale, y die vertikale Position des Ellipsenzentrums im Bild, σ den horizontalen und τ den vertikalen Radius der Ellipse.

Eine Gütefunktion für eine Ellipse bzw. einen Kopfkandidaten mit Zustand s_i kann auf den durch die vorigen Abschnitte erstellten Ergebnisbildern durch

$$\Phi_{s_i} = \phi_{H_R}(s_i) + \phi_V(s_i) \tag{2.2}$$

definiert werden, wobei $\phi_{H_R}(s_i)$ eine Bewertung der Ellipse über dem Ergebnis der Histogramm-Rückprojektion, $\phi_V(s_i)$ eine Bewertung der Ellipse über

dem Resultat der Vordergrundsegmentierung darstellen.
Ziel ist es, einen Kopfkandidaten mit Zustand s^* zu finden, so dass gilt:

$$\begin{aligned} s^* &= \arg\max_{s_i \in S} \Phi_{s_i} \\ &= \arg\max_{s_i \in S} \{\phi_{H_R}(s_i) + \phi_V(s_i)\} \end{aligned} \tag{2.3}$$

Das System versucht also eine Ellipse derart zu finden, dass die Summe aus ihrer Bewertung über dem Segmentierungsergebnis der farbbasierten Verfolgung und der Bewertung über der Vordergrundsegmentierung maximal ist.

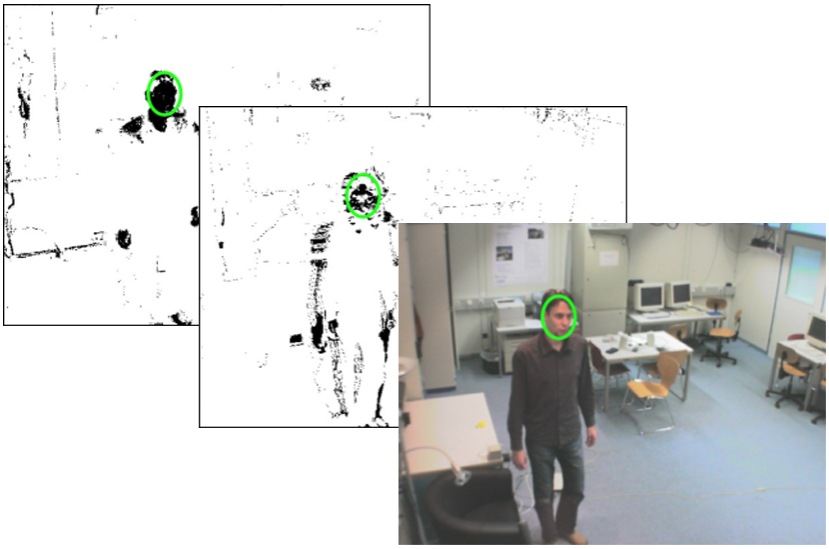

Abbildung 2.6. Die Suche nach Kopfkandidaten erfolgt durch eine Approximation des Kopfes durch eine Ellipse. Dabei werden Ellipsen verschiedener Größen über das Suchfenster geschoben und entsprechend der Farbsegmentierung und Vordergrundklassifikation bewertet. Die Ellipse mit der besten Bewertung entspricht dem wahrscheinlichsten Kopfkandidaten und das darunterliegende Kopfbild kann entsprechend den Ausmaßen der Ellipse extrahiert werden.

Sei nun P die Menge aller Pixel, die von der Ellipse mit Zustand $s = (x, y, \sigma, \tau)$ eingeschlossen werden, sowie $f(p)$ der entsprechende Zugehörigkeitswert eines Pixels $p \in P$ nach den oben beschriebenen Verfahren, dann lässt sich die Bewertung $\phi(s_i)$ wie folgt beschreiben:

$$\phi(s_i) = \sum_{p \in P} \zeta(p), \text{ mit } \zeta(p) = \begin{cases} c & \text{falls } f(p) = 0 \\ f(p) & \text{sonst} \end{cases} \tag{2.4}$$

c stellt dabei einen konstanten Faktor dar, der als Strafe für die Einbeziehung ungenügender Pixel fungiert. Im Fall der Vordergrundsegmentierung

entspräche dies Punkten die eindeutig dem Hintergrund zugeordnet werden können, im Fall der Farbsegmentierung Pixeln die nicht hinreichende Ähnlichkeit zum Referenzhistogramm aufweisen. In der Praxis hat sich für c ein einfacher Wert von -1 als ausreichend erwiesen.

Suchraumreduktion durch Triangulation

Eine darüberhinausgehende Reduktion des Suchraums wird in dieser Arbeit durch Verwendung des in Anhang B.2 beschriebenen Verfahrens der *Triangulation* erreicht. Das durch Triangulation mögliche Ermitteln der dreidimensionalen Position der zuletzt beobachteten Kopfposition erlaubt es Kopfkandidaten auf einen festen Bereich um diesen Punkt herum einzuschränken. Eine Rückprojektion des Suchbereichs auf die jeweiligen projizierten Kameraansichten ermöglicht die Minimierung der Suche auf ein Fenster variabler Größe (siehe Abbildung 2.7).

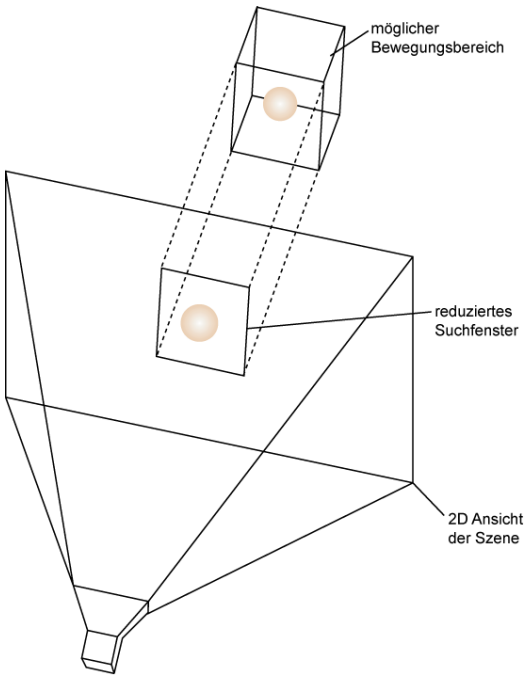

Abbildung 2.7. Durch Rückprojektion eines erlaubten Bewegungsbereichs des zu verfolgenden Kopfes auf die projizierte Kameraansicht kann das Suchfenster nach Kopfkandidaten auf diesen Bereich reduziert werden.

2.3 Vorverarbeitung

Die in den Aufnahmesequenzen gefundenen Kopfpositionen, ermöglichen es Kopfbilder aus den Sequenzen zu extrahieren. Die Kopfbilder sollen im späteren Verlauf dazu dienen Einzelhypothesen der Kopfdrehung pro vorliegender Vorderkopfansicht zu generieren und diese bei der Fusion zu einer Gesamthypothese bereitzustellen. Hierzu müssen die extrahierten Kopfbilder in ihrer Größe angepasst und im Kontrast normiert werden um eine optimale Ansicht des Kopfes für die Schätzung zu bieten. Ferner soll durch zusätzliches Wissen über die horizontale und vertikale Kantenstruktur innerhalb des extrahierten Kopfbildes der Schätzung zusätzliche Informationen über die vorliegende Kopfstellung zur Verfügung gestellt werden. Der folgende Abschnitt erläutert die dafür eingesetzten Vorgehensweisen.

Abbildung 2.8. Die Vorverarbeitung der Kopfbilder besteht aus einer Skalierung der extrahierten Bilder, welche anschließend in ein Helligkeitsbild transformiert und im Kontrast normiert werden. Ferner werden jeweils ein vertikales als auch horizontales Kantenbild aus dem normierten Helligkeitsbild generiert.

2.3.1 Skalierung

Neuronale Netze verlangen eine konstante Anzahl an zugrundeliegenden Neuronen. Die feste Anzahl an Eingabeneuronen verlangt daher eine Skalierung der Eingabebilder auf eine der Eingabeschicht entsprechenden Größe. Durch Einhalten eines konstanten Seitenverhältnisses während der Extraktion der Kopfbilder sowie während des Skalierens wird verhindert, dass verzerrte Kopfansichten durch die Vergrösserung bzw. Verkleinerung entstehen würden.

Die Bilder werden durch eine *bilineare Interpolation* skaliert: Dabei werden Farbwerte für Bildpunkte an nicht ganzzahligen Positionen (x, y) durch eine gewichtete Interpolation der vier ganzzahligen Nachbarpixel $p(m, n)$, $p(m+1, n)$, $p(m, n+1)$ und $p(m+1, n+1)$ ermittelt. Die Gewichte setzen sich dabei entsprechend der Abstände zu den Nachbarpixel zusammen (siehe Abb. 2.9):

$$\begin{aligned} p(x,y) &= h \cdot v \cdot p(m,n) \\ &\quad +(1-h) \cdot v \cdot p(m+1,n) \\ &\quad +h \cdot (1-v) \cdot p(m,n+1) \\ &\quad +(1-h) \cdot (1-v) \cdot p(m+1,n+1) \end{aligned} \tag{2.5}$$

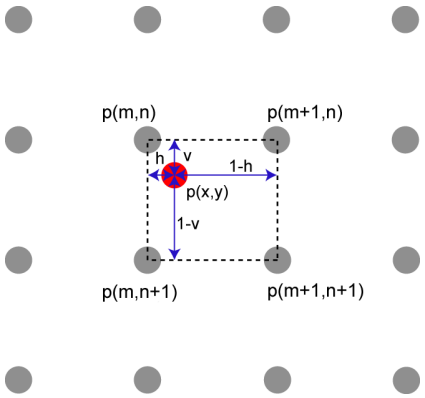

Abbildung 2.9. Bilineare Interpolation des Farbwertes eines neuen Pixels durch die Farbwerte der vier umliegenden, ganzzahligen Nachbarpixel. h gibt dabei den horizontalen Abstand, v den vertikalen Abstand zum nächstliegenden Nachbarpixel an.

2.3.2 Kontrastnormierung

Die ausgeschnittenen Kopfbilder liegen in der originalen Beleuchtung vor, die etwaige feingranulare Merkmale des Kopfes nicht optimal hervorhebt bzw. sogar komplett unterdrückt. Um diesen Effekt zu vermindern werden die extrahierten Kopfbilder in Grauwert- bzw. Helligkeitsbilder transformiert und anschließend histogramm-normalisiert was dazu führt dass der Kontrast erhöht wird (siehe Abbildung 2.10). Man spricht von einer sogenannten *Kontrastnormierung*.
Der Kontrast ist definiert durch den Unterschied zwischen hellstem und dunkelstem Wert eines Bildes. In einem 256- stufigem Grauwertbild, äussert sich

Abbildung 2.10. Beispiel einer Kontrasterhöhung: Ein extrahiertes Kopfbild (links) wird in ein Helligkeitsbild transformiert (mitte) und anschließend im Kontrast erhöht (rechts).

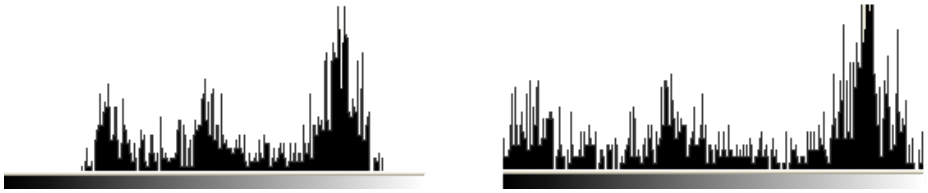

Abbildung 2.11. Resultat der Kontrastnormierung: das Histogramm des nicht normierten Grauwertbildes aus Abbildung 2.10 (links) nutzt den zur Verfügung stehenden Wertebereich an Graustufen nicht optimal aus. Im Gegensatz dazu nutzt das Histogramm des normierten Kopfbildes aus Abbildung 2.10 (rechts) den kompletten Wertebereich und erhöht damit den Kontrast, was zu einer besser erkennbaren Darstellung des Gesichts führt.

dieser Sachverhalt durch einen Wertebereich der zwar 256 Werte umfasst, diese aber nicht immer optimal belegt (siehe Abbildung 2.11).
Um den Kontrast generell zu erhöhen gibt es verschiedene Ansätze:

- *Lineare Streckung*: das Histogramm eines Bildes wird linear gestreckt. Optimal ist diese Anwendung bei Bildern mit gaußähnlicher Histogrammverteilung, d.h. wenn die vorkommenden Helligkeitswerte in einem relativ kleinen Areal normal um einen Mittelwert verteilt sind.

- *Histogrammangleichung*: eine nicht lineare Methode, welche die Helligkeitswerte eines Bildes so verändert, dass das enstehende Histogramm einem Referenz-Histogramm ähnelt: die Form eines bestehenden Histogramms wird einer neuen, gewünschten Form angepasst.

In dieser Arbeit kommt das erst genannte Verfahren zum Einsatz. Die Histogramme der Eingabebilder werden linear gestreckt, so dass ihr kompletter Wertebereich umfasst wird. Auf einem Grauwertbild mit dem minimalen Grauwert g_{min} und dem maximalen Grauwert g_{max} lässt sich die Abbildung eines Grauwertes g auf seinen neuen, linear gestreckten Wert $\varphi(g)$ dabei wie folgt definieren:

$$\varphi(g) = \hat{g}_{min} + \frac{\hat{g}_{max} - \hat{g}_{min}}{g_{max} - g_{min}} * (g - g_{min}) \tag{2.6}$$

wobei $\hat{g}_{min}$ der gewünschte minimale Graustufenwert und $\hat{g}_{max}$ der gewünschte maximale Graustufenwert des angepassten Histogramms ist.

2.3.3 Kantenfilter

Als zusätzliche Information bezüglich der Kopfdrehung soll neben dem im Kontrast normierten Helligkeitsbild des Kopfes auch ein horizontales sowie vertikales Kantenbild zum Einsatz kommen. Die Kanten sollen dabei weitere Hinweise auf die Kopfstellung liefern, insbesondere durch Verdeutlichung des Kopfumrisses und der Muster die durch Augen und Nasenlinien hervorgehoben werden. Die Kanten werden dabei durch den *Sobel-Operator* berechnet.

Der Sobel-Operator für Kantenextraktion ist definiert als die Faltung eines zweidimensionalen Bildes $I(m,n)$ mit einer Sobel-Matrix $M(m,k)$:

$$I_f(i,j) = \frac{1}{S} \sum_{m=-1}^{1} \sum_{k=-1}^{1} I(i+m, j+k) M(m,k) \tag{2.7}$$

wobei

$$S = \sum_{m=-1}^{1} \sum_{k=-1}^{1} M(m,k)$$

Die Sobel-Matrix $M(m,k)$ entspricht dabei für

- *horizontale Kanten*: $M(m,k) = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}$
- *vertikale Kanten*: $M(m,k) = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}$

2.4 Vorderkopf-Hinterkopf-Klassifikation

Die Aufgabe einer *Vorderkopf-Hinterkopf-Klassifikation* besteht im Rahmen dieser Arbeit darin, Hinterkopfansichten zu entfernen, so dass die eigentliche Schätzung der Kopfdrehung nur auf Vorderkopfbildern geschieht.
Durch die zufällige Musterbildung der Haare in Hinterkopfansichten, stellen solche Aufnahmen keine ausreichende Anzahl an zuverlässigen Merkmalen zur Verfügung, so dass durch Trainieren eines neuronalen Netzes zur

Kopfdrehungsschätzung Hinterkopfansichten einen nachteiligen Effekt einbringen würden. Im vorliegenden Abschnitt wird daher der Einsatz eines neuronalen Netzes vorgestellt, dessen stetige Ausgabe zwischen 0 und 1 die a-posteriori Wahrscheinlichkeit eines beobachteten Kopfbildes als Vorderkopfansicht einschätzt. Der Begriff *Vorderkopf* wird dabei so definiert dass sich das Gesicht in einem Bereich von $-90°$ bis $+90°$ relativ zur entsprechenden Sichtlinie der Kamera befindet, deren Aufnahme zur Schätzung vorliegt (siehe Abb. 2.12).

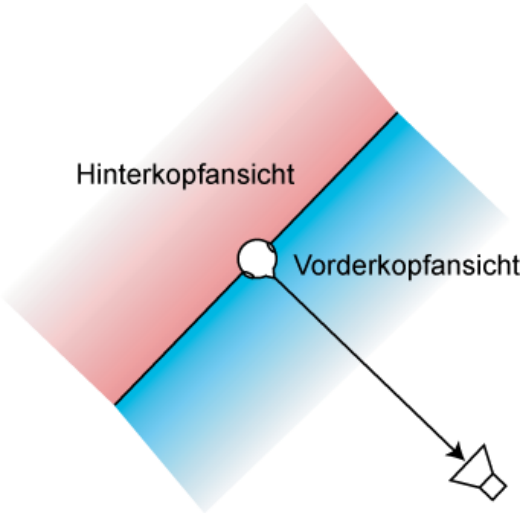

Abbildung 2.12. Eine Vorderkopfansicht liegt genau dann vor, wenn sich das Gesicht in einem Bereich von $-90°$ bis $+90°$ um den Sichtvektor der Kamera befindet.

2.4.1 Topologie

Zur Auffindung einer optimalen Netztopologie für die vorgestellte Aufgabe werden dreischichtige, vollständig vorwärts gerichtete Perzeptronen unterschiedlicher Anzahl an Neuronen der versteckten Schicht sowie der Eingabeschicht implementiert und evaluiert. Die geforderte Ausgabe ist jeweils eine stetige Schätzung zwischen 0 und 1, die die Wahrscheinlichkeit einer Vorderkopfansicht in den angelegten Bilddaten darstellt (Abbildung 2.13). Ein Schwellwertvergleich klassifiziert das angelegte Kopfbild aufgrund seiner Wahrscheinlichkeit als Vorderkopf bzw. Hinterkopf.

Als Eingabe dient dem Netz jeweils ein im Kontrast normiertes Intensitätenbild sowie ein jeweils horizontales und vertikales Kantenbild der Kopfansicht (siehe Abbildung 2.13). Die notwendige Anzahl der Eingabeneuronen korreliert dabei mit der beliebigen, aber festen Größe der Eingabebilder. Die

optimale Anzahl an Neuronen der Eingabeschicht wird durch verschiedene Bildgrößen empirisch ermittelt, deren Seitenverhältnisse jeweils dem der ursprünglich extrahierten Kopfbilder entsprechen. Zum Einsatz kommen dabei Bildgrößen von 15×20, 20×26 und 25×33 Pixel, was bei den drei beschriebenen Eingabebildern 900, 1560 und 2475 Eingabeneuronen enstpricht.
In Anlehnung an [20] werden pro Bildmaß Netze mit mittleren Schichten unterschiedlichen Umfangs ausgewertet. Zum Einsatz kommen dabei jeweils 50, 60, 70, 80, sowie 100 versteckte Neuronen. Die jeweiligen Ergebnisse sowie deren Interpretation finden sich in Kapitel 3.

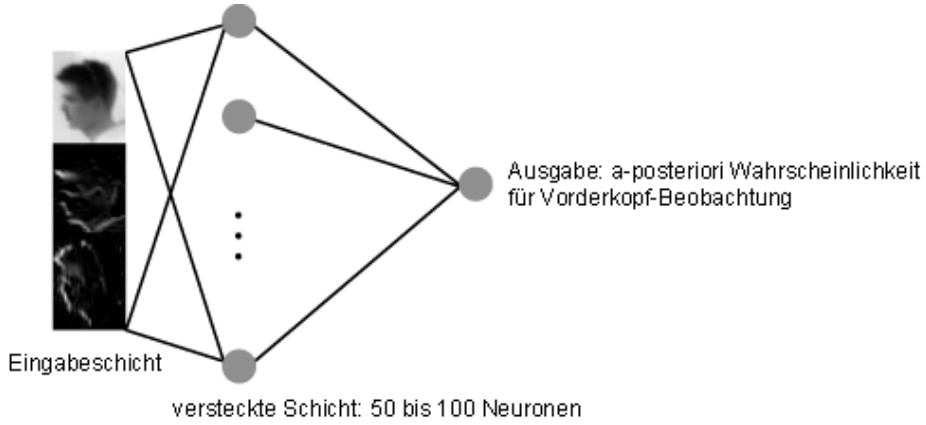

Abbildung 2.13. Ein neuronales Netz zur Vorderkopf-Hinterkopf-Klassifikation. Die Ausgabe besitzt den stetigen Wertebereich $[0, 1]$ und bezeichnet die a-posteriori Wahrscheinlichkeit für das Vorhandensein einer Vorderkopfansicht.

2.4.2 Training

Das neuronale Netz wird mit Kopfbildern trainiert, die jeweils einen Vorder- oder Hinterkopf zeigen. Aufbauend auf den transkribierten Daten der durch den Magnetsensor gemessenen Kopfdrehung, wird die gewünschte Soll-Ausgabe des Netzes durch Berechnung des Drehwinkels relativ zur Sichtlinie der entsprechenden Kamera berechnet, deren vorverarbeitetes Kopfbild dem Netz angelegt wird. Für die hierfür eingesetzte Umrechnung der im Raumkoordinatensystem transkribierten Winkel sei auf Anhang B.3 hingewiesen.
Für Kopfbilder deren relativer Drehwinkel im Bereich von $-90°$ bis $+90°$ liegt, entspricht die Sollausgabe dem Wert 1, bei Hinterköpfen entsprechend 0. Das Netzwerk wird mit 100 Trainingszyklen eingelernt, durch ein *cross evaluation set* an Kopfbildern wird darunter jenes Netz extrahiert dessen Trainingsiteration ein lokales Minimum bezüglich des Ausgabefehlers darstellte.

2.5 Drehschätzung

Durch die in den vorhergehenden Abschnitten beschriebenen Verfahren wird ein zu beobachtender Kopf in vier Kameraansichten verfolgt, pro Ansicht wird ein Kopfbild extrahiert und entsprechend vorverarbeitet. Etwaige Hinterkopfbilder werden durch den Einsatz eines Vorderkopf-Hinterkopf-Klassifikators in Form eines neuronalen Netzes entfernt. In diesem Abschnitt wird nun ein weiteres neuronales Netz vorgestellt, dessen Aufgabe darin besteht, die jeweils gesehene Kopfdrehung pro vorliegender Vorderkopfansicht zu schätzen. Die geforderte Ausgabe soll dabei einer Winkelangabe im stetigen Bereich von $-90°$ bis $+90°$ entsprechen und somit innerhalb des Bezugsystems der jeweiligen Kameras beschrieben werden. Durch die dadurch entstehende Schätzung pro Kameraansicht, bleibt das trainierte Netz für unterschiedliche Vorderkopfansichten nutzbar und ermöglicht den Einsatz beliebig vieler, weiterer Kameras, ohne das System neu einlernen zu müssen. Für die hierfür notwendige Umrechnung der Winkel in kamerarelative Angaben sowie das Berechnen des Sichtvektors des beobachteten Kopfes anhand kamerarelativer Winkelangaben sei auch an dieser Stelle auf Anhang B.3 verwiesen.

2.5.1 Topologie

Äquivalent zu Abschnitt 2.4, muss eine geeignete Topologie des Netzes gefunden werden, welche eine möglichst optimale Vorhersage des Drehwinkels erlaubt. Auch hier bietet sich ein dreischichtiges, vollständig vorwärts gerichtetes Perzeptron an, dessen Umfang der Eingabe- sowie Mittelschicht bestimmt werden muss.

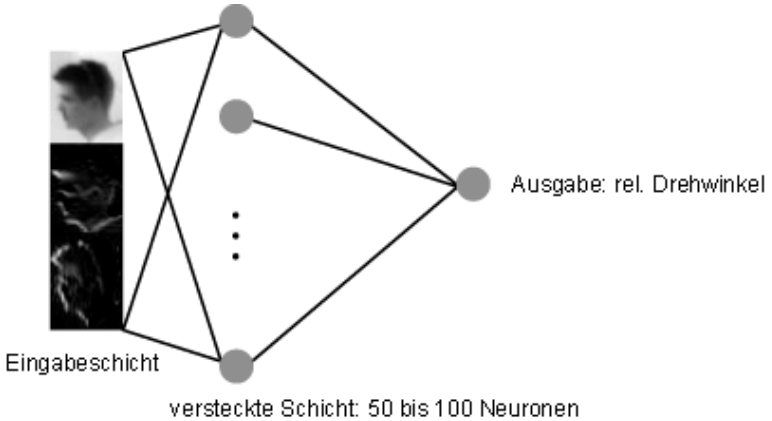

Abbildung 2.14. Das neuronale Netz schätzt die stetige, relative Drehung des Kopfes anhand der drei Eingabebilder.

Entsprechend dem vorigen Abschnitt der Vorderkopf-Hinterkopf-Klassifikation findet die Evaluation auf jeder der drei Bildgrößen von 15×20, 20×26 und 25×33 Pixel statt, durch die die Anzahl an notwendigen Eingabeneuronen bestimmt wird. Pro Bildgröße werden mittlere Schichten mit 50, 60, 70, 80 und 100 Neuronen implementiert und getestet.

Abbildung 2.14 verdeutlicht die eingesetzte Netz-Topologie.

2.5.2 Training

Das Netzwerk wird ausschließlich mit vorverarbeiteten Eingabebildern von Vorderkopfansichten trainiert. Die gewünschte Sollausgabe entspricht dabei dem relativen horizontalen Drehwinkel zur Sichtlinie der jeweiligen Kamera, deren vorverarbeitetes Kopfbild dem Netzwerk als Eingabe vorliegt. Durch die stattfindende Transkribierung der Kopfdrehung im Bezugsystem des Raums gemäß Abbildung 1.1 durch den eingesetzten Magnetsensor wird hierzu eine Umrechnung in das Bezugsystem der jeweiligen Kameras notwendig.

Entsprechend dem in Abschnitt 2.4 vorgestellten Vorderkopf-Hinterkopf-Klassifikator wird auch dieses Netzwerk mit 100 Trainingszyklen eingelernt. Durch ein *cross evaluation set* an Vorderkopfansichten werden die Lernzyklen überwacht und jenes Netz extrahiert dessen Trainingsiteration ein lokales Minimum bezüglich des Ausgabefehlers darstellte. Für Ergebnisse der einzelnen Netzwerke in Abhängigkeit vom Umfang der Eingabe- sowie Mittelschicht sei auf Kapitel 3 hingewiesen.

2.6 Fusion der Hypothesen

Um die Stabilität eines Systems durch eine Benutzung mehrerer Signalquellen zu erhöhen, muss eine geeignete Fusion der unterschiedlichen Messungen zu einer Gesamthypothese der Beobachtung gefunden werden. Im Fall der Beobachtung einer Kopfdrehung durch vier Kameras entspricht dies dem Vereinen vierer verschiedener Schätzungen zu einer finalen Annahme des tatsächlichen Drehwinkels. Eine solche Fusion kann im wesentlichen auf zwei unterschiedlichen Ebenen geschehen:

- *Signalebene*: Die vier unterschiedlichen Kameraansichten werden in eine gemeinsame Darstellung vereint. Die Schätzung der Kopfdrehung erfolgt anschließend aufbauend auf der gemeinsamen Repräsentation.

- *Hypothesenebene*: Das System erstellt für jede Kameraansicht eine eigene Schätzung der Kopfdrehung. Die unterschiedlichen Hypothesen werden daraufhin zu einer Gesamthypothese vereint.

Durch Implementierung letzterer genannter Fusionsebene, soll das in dieser Arbeit entwickelte System die einfache Möglichkeit bieten zusätzliche Kameras zur Beobachtung hinzuzuziehen um eine eventuelle Leistungssteigerung zu gewährleisten.

Der vorliegende Abschnitt soll nun das in dieser Arbeit eingesetzte Verfahren zur Hypothesenfusion verdeutlichen.

2.6.1 Hypothesenfusion

Im folgenden sei $\Theta = \{\theta_i\}$, mit $0° \leq \theta_i < 360°$ und $0 < i$ die Menge aller möglichen Drehwinkel einer Person gemäß Abbildung 2.1 gegeben.
Zu jedem Zeitpunkt t liegen dem System n verschiedene Einzelhypothesen $\{h_1, h_2, \ldots, h_n\}$ an Kopfdrehungen vor. Die Hypothesen wurden dabei auf den, durch den Vorderkopf-Klassifikator als Vorderkopfansicht erkannten Kopfbildern durch das neuronale Netz der Drehschätzung erstellt.
Sei ferner durch $P(\theta_i|h_j)$ die diskrete a-posteriori Wahrscheinlichkeit beschrieben, dass die tatsächliche Kopfdrehung θ_i beim Beobachten einer Schätzung h_j zugrundelag, so lässt sich eine finale Drehhypothese $\hat{\theta}$ der Kopfdrehung durch einen Maximum-Likelihood-Ansatz wie folgt definieren:

$$\hat{\theta} = \arg\max_{\theta_i \in \Theta} \prod_{j=1}^{n} P(\theta_i|h_j) \tag{2.8}$$

Durch die eingesetzte Multiplikation strebt die Wahrscheinlichkeit einer Drehhypothese θ_i gegen 0 je mehr Kameraansichten bzw. Vorderkopfansichten zur Fusion eingesetzt werden. Aufgrund der jedoch gewünschten Zunahme der Wahrscheinlichkeit bei mehreren Vorderkopfansichten[4] wurde die Multiplikation der einzelnen Wahrscheinlichkeiten $P(\theta_i|h_j)$ daher durch eine Summe ersetzt und implementiert:

$$\hat{\theta} = \arg\max_{\theta_i \in \Theta} \sum_{j=1}^{n} P(\theta_i|h_j) \tag{2.9}$$

[4]Zugrundeliegend wird angenommen dass eine geschätzte Drehung umso wahrscheinlicher ist, je mehr Einzelhypothesen diesem Ergebnis entsprechen

Das dabei notwendige Konfidenzmaß zum Ermitteln der a-posteriori Wahrscheinlichkeiten $P(\theta_i|h_j)$ wurde im Rahmen dieser Arbeit durch die Erstellung einer *Konfusionsmatrix* pro Kamera gefunden. Der folgende Abschnitt soll deren Erstellung und Einsatz näher beschreiben.

2.6.2 Konfidenzmaß

Die Elemente einer Konfusionsmatrix K spiegeln die Häufigkeiten wider, mit denen korrekte oder falsche Klassifikationen bezüglich einer gegebenen Menge diskreter Klassen beobachtet wurden. Die Zeilen entsprechen dabei der jeweils tatsächlich vorliegenden, die Spalten der jeweils geschätzten Klasse. Die Diagonalelemente beschreiben demnach die Häufigkeit einer korrekten Klassifikation zu einer Klasse C. Zur Verdeutlichung des Zusammenhangs sei auf Tabelle 2.1 verwiesen.

Eine Normierung jedes Spaltenvektors der Matrix erstellt eine diskrete Wahrscheinlichsverteilung pro beobachtbarer Klasse über dem durch die Zeilen der Matrix definierten Wertebereich möglicher Ausprägungen, wodurch die a-posterori Wahrscheinlichkeiten einer tatsächlich vorliegenden Klasse C_i unter Vorliegen einer Beobachtung C_j direkt aus der Matrix ablesbar werden. Für Elemente einer Spalte j gilt damit:

$$\hat{k}_{ij} = \frac{k_{ij}}{\sum_m k_{mj}} \tag{2.10}$$

Die Wahrscheinlichkeit $P(C_i|C_j)$ einer Klasse C_i gegeben der Beobachtung C_j ist damit definiert durch das entsprechende Element desjenigen normierten Spaltenvektors, der der Klasse C_j entspricht:

$$P(C_i|C_j) = \hat{k}_{ij} = \frac{k_{ij}}{\sum_m k_{mj}} \tag{2.11}$$

Die Erstellung einer Konfusionsmatrix erfolgt im Rahmen dieser Arbeit durch das *cross evaluation set* an Kopfbildern, welches benutzt wird, um das Training des neuronalen Netzes zur Schätzung der Kopfdrehung bezüglich des Klassifikationsfehlers zu überwachen. Der Hypothesenraum möglicher Kopfdrehungen entspricht zum Zeitpunkt der Schätzung einem Wertebereich von $-90°$ bis $+90°$, da jeweils versucht wird die relative Kopfdrehung zu ermitteln, die auf den vorverarbeiteten Bildern dem Netzwerk als Eingabe dient. Eine Einteilung in äquidistante Klassen diskretisiert diesen Wertebereich und ermöglicht die Erstellung oben genannter Konfusionsmatrix, indem bezüglich des *cross evaluation set* jede durch das Netz entstandene Hypothese ihrer

Klasse	-90	-60	-30	0	30	60	90
-90	58	82	22	4	5	8	2
-60	56	158	80	23	6		
-30		55	208	109	5	1	
0		2	90	433	72	4	
30			7	107	203	49	1
60		1	11	25	68	163	62
90	1	3	7	7	17	62	69

Tabelle 2.1. Eine, während der Evaluation beobachtete Konfusionsmatrix über Aufnahmen einer einzelnen Kamera. Der Wertebereich von $-90°$ bis $+90°$ der möglichen Kopfdrehungen wurde in 7 Klassen eingeteilt, von denen jede einen Bereich von $30°$ umfasst. Die Matrizenelemente beschreiben die Anzahl der beobachteten Zuordnungen zu den jeweiligen Klassen. Die Spalten geben dabei die geschätzten Klassen, die Zeilen die Sollklassen an.

entsprechenden Klasse zugewiesen und das jeweilige Matrizenelement inkrementiert wird. Hierfür hat sich eine Klasseneinteilung in $30°$ Schritten als ausreichend erwiesen.

Durch die Kalibrierung der jeweiligen Kamera, deren Kopfbilder zur Erstellung der Konfusionsmatrix dienten, kann durch das in Anhang B.3 aufgeführte Verfahren jederzeit von einer Winkelklasse der Matrix auf ihre entsprechende, absolute Winkeldarstellung bezüglich des Raumkoordinatensystems geschlossen werden.

Kapitel 3

Ergebnisse und Auswertung

Berücksichtigt man, dass die mögliche Anzahl an Vorderkopfansichten sowohl von der Anordnung der eingesetzten Kameras, als auch von der Position der beobachteten Person im Raum abhängt, so erfordert eine Evaluation die zusätzliche Kenntnisnahme der Anzahl eingesetzter Vorderkopfansichten, die zur Fusion einer Gesamthypothese beigetragen haben. Eine Messung der Leistungsfähigkeit des entwickelten Systems muss demnach hinsichtlich des mittleren Fehlers der fusionierten Gesamthypothese unter Berücksichtigung der Anzahl an Kameras geschehen, deren Kopfansicht und Schätzung hierfür eingesetzt werden. Dabei wird im folgenden zwischen dem Fall *bekannter Personen* und dem Fall *unbekannter Personen* unterschieden. Erst genannte Methode evaluiert das System mit Kopfansichten der Personen, von denen auch Kopfbilder zum Einlernen des Systems benutzt wurden. Hierzu wurden alle aus den Aufnahmen extrahierten Kopfbilder aller Personen in eine gemeinsame Trainingsmenge[1] und Testmenge eingeteilt. Das Verhältnis der Anzahl zugewiesener Bilder entsprach dabei $2:1$.
Der Fall einer Evaluierung des Systems anhand unbekannter Personen geschieht entsprechend dem *Leave-one-out* Prinzip, bei dem aus der vorliegenden Menge an Daten jeweils eine Person dem Training vorenthalten wird, so dass eine ausschließliche Evaluation auf dieser Person eine Bewertung hinsichtlich neuer, bisher ungesehener Kopfbilder garantiert.

3.1 Datensammlung

Die Aufnahmen geschahen in einem möblierten Zimmer um eine möglichst realistische Hintergrundszenerie nachzustellen (siehe Abbildung 1.3). Sieben

[1] Im folgenden sei das *cross evaluation set* zur Überwachung des Netzfehlers durch die Trainingsmenge miteinbezogen und soll nicht explizit aufgeführt werden

Personen wurden jeweils gebeten während der Aufnahmen ihren Kopf auf eine möglichst natürliche Art und Weise zu bewegen. Dabei durften sie frei im Raum umherlaufen. Die einzige Beschränkung wurde durch den ca. $1,5$ Meter großen Empfangsradius des Magnetsensors auferlegt, der zur Transkribierung der tatsächlichen Kopfdrehung eingesetzt und mit einem Haarreif auf dem Kopf der aufgenommenen Personen angebracht wurde.

Pro Kamera und Person wurden jeweils einminütige Videosequenzen aufgenommen. Im Fall von drei der sieben Personen wurde der Ausgangspunkt im Raum geändert und eine zweite Aufnahme aufgezeichnet um mehrere Positionen im Raum abzudecken. In allen Aufnahmen wurde die Position des Kopfzentroiden manuell markiert um erstens eine eventuelle Korrektur der Kopfextraktion zu ermöglichen und zweitens ein Entfernen mangelhafter Kopfansichten vorzunehmen, die durch eine Verdeckung durch den eigenen Körper (Hände vor das Gesicht), eine Verdeckung durch vorbeigehende Personen, eine Verdeckung durch Interieur des Raums oder durch Verlassen des Sichtbereichs der Kamera auftreten können.

3.2 Der Fall bekannter Personen

Da die Leistungsfähigkeit, insbesondere die Generalisierungsfähigkeit eines Systems hinsichtlich neuer, unbekannter Daten nur schlecht anhand Schätzungen ermittelt werden kann, die auf Aufnahmen gemacht wurden auf die das System während der Einlernphase trainiert wurde, soll im folgenden durch die Evaluation auf diesen bekannten Personen eine optimale Topologie der neuronalen Netze gefunden werden um damit den personenunabhängigen Fall zu realisieren. Hierzu wurden drei unterschiedliche Eingabebildgrößen von 15×20, 20×26 und 25×33 Pixel sowie mittlere Schichten von jeweils 50, 60, 70, 80 und 100 versteckten Neuronen implementiert und getestet.

Insbesondere die Menge an Trainingsdaten ist entscheidend für ein Einlernen eines stabilen neuronalen Netzes. Zur Erhöhung der Anzahl wurden alle Kopfbilder innerhalb der Trainingsmenge horizontal gespiegelt und somit verdoppelt. Die Verteilung der Trainingsdaten (siehe Abbildung 3.1) erscheint daher symmetrisch. Die Kopfbilder, die zur Evaluation der trainierten Netze benutzt wurden, blieben ungespiegelt um die zeitliche Abfolge der Aufnahmen und die genaue Anzahl der vier Kameraansichten unverändert bestehen zu lassen. Die nicht symmetrische Verteilung der Evaluationsdaten verdeutlicht dies in Abbildung 3.2. Insgesamt entstanden damit 28939 extrahierte Kopfbilder, wovon 22152 dem Training zugewiesen wurden, 6787 zur Evalu-

ation zur Verfügung standen.

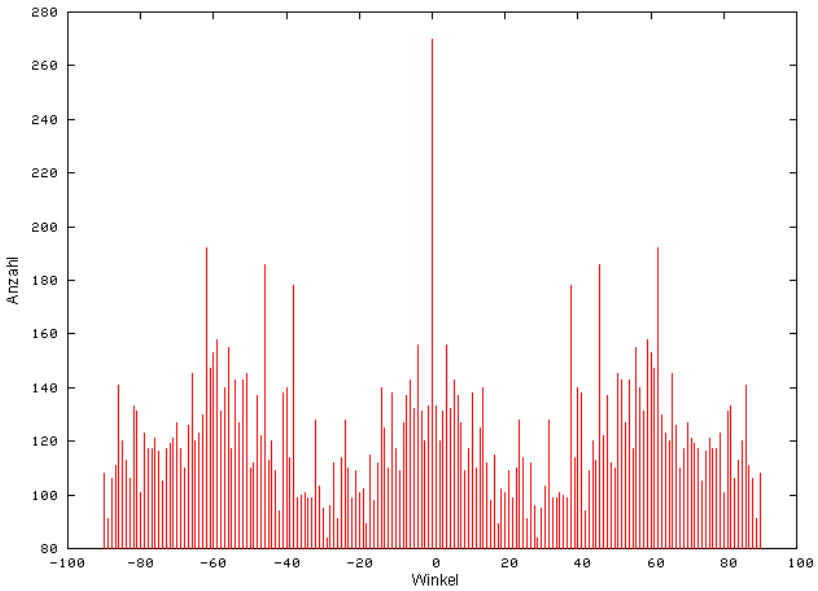

Abbildung 3.1. Die Verteilung der in der Trainingsmenge enthaltenen Kopfdrehungen erscheint symmetrisch, da alle extrahierten Vorderkopfansichten horizontal gespiegelt wurden um die Menge an Daten zu verdoppeln.

3.2.1 Analyse

Wie aus den Ergebnissen der Vorderkopf-Hinterkopf-Klassifikation in Tabelle 3.1 hervorgeht, erfolgt die bestmögliche Erkennung von Vorderkopfansichten bei einer Topologie von 70 versteckten Neuronen und einer Eingabebildgröße von 15×20 Pixel. Die Auswertung der Drehschätzung soll daher im folgenden unter Zuhilfenahme dieser Klassifikationsleistung der Vorderkopferkennung erfolgen.

Tabelle 3.2 verdeutlicht die Schätzleistung des Systems hinsichtlich der Kopfdrehung unter Berücksichtigung der Anzahl hierzu eingesetzter Vorderkopfansichten bzw. Einzelhypothesen. Besonders deutlich zu erkennen ist die Verbesserung der Drehschätzung mit zunehmender Anzahl eingesetzter Kameras bzw. zunehmender Anzahl erkannter Vorderkopfansichten. Der mittlere Fehler bezüglich Hypothesen die aus zwei Kameraansichten erstellt wurden liegt dabei zwischen $18,4°$ und $21,9°$. Bei Hypothesen die aus drei Kameraansichten erstellt wurden sogar nur zwischen $15,0°$ und $17,4°$.

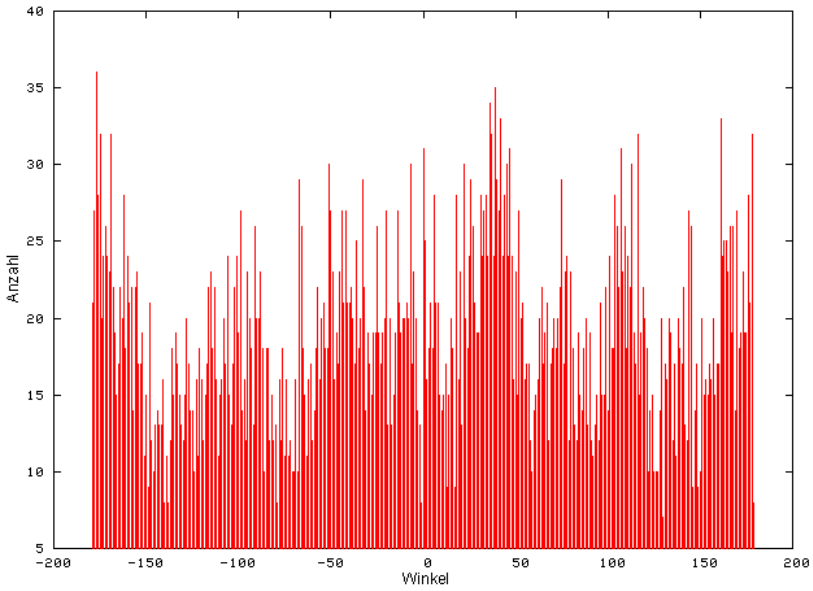

Abbildung 3.2. Die Verteilung der im Fall bekannter Personen in der Testmenge enthaltenen Kopfdrehungen ist unsymmetrisch: die Kopfbilder wurden nicht gespiegelt um die originalen Aufnahmen in ihrer Anzahl und zeitlichen Abfolge nicht zu verändern.

Der hohe Schätzfehler von bis zu $30,7°$ bei Hypothesen aus nur einer Kameraansicht setzt sich aus mehreren Gründen zusammen:
Aufgrund des Mangels an optischer Vergrößerung der Kopfregionen während der Aufnahme und der weit von den aufgenommenen Köpfen entfernten Anbringung der Kameras unterhalb der Decke, wirkt sich die geringe Auflösung der Kopfbilder durch den Verlust feingranularer Gesichtsmerkmale, die zur Unterscheidung verschiedener Kopfstellungen notwendig sind, in einer ungenauen Schätzung der Drehung aus.

	Anzahl versteckter Neuronen				
Bildgröße (Pixel)	50	60	70	80	100
15×20	87,3%	86,5%	87,9%	86,8%	86,0%
20×26	87,2%	87,3%	86,9%	85,6%	85,9%
25×33	86,2%	84,3%	86,5%	87,5%	87,8%

Tabelle 3.1. Die Elemente der Tabelle geben die korrekte Klassifikation zu Vorder- bzw. Hinterköpfen im Fall bekannter Personen wieder unter Einsatz unterschiedlicher Bildgrößen und Topologien des zugrundeliegenden neuronalen Netzes.

Desweiteren zieht die Anbringung der Kameras über den zu beobachtenden

Köpfen mit sich, dass schon durch geringe vertikale Neigungen[2] ein Großteil des Gesichts unerfasst bleibt, währenddessen der Haaransatz an Stirn und Schläfen den Großteil der Aufnahme einnimmt. Die Vermutung, dass die in der Frisur der Person enthaltenen, zufälligen Muster eine negative Einwirkung auf das Einlernen des Systems nehmen[3], würde eine mögliche Begründung liefern.
Ferner stellen die durch Kalibrierung und Triangulation entstandenen Messfehler, sowie minimale, während der Transkribierung der tatsächlichen Kopfdrehung auftretende Abweichungen weitere Ungenauigkeiten dar, die sich in Form von Rauschen auf das Einlernen der Netze sowie auf die eigentliche Hypothesenbildung auswirken.

Bildgröße	# N_D	# Einzelhypothesen 1	2	3	Gesamtfehler
15×20	50	$27,2°$	$19,8°$	$16,6°$	$22,9°$
	60	$26,7°$	$18,4°$	$15,5°$	$21,8°$
	70	$28,0°$	$20,3°$	$16,1°$	$23,4°$
	80	$29,5°$	$20,1°$	$15,0°$	$23,9°$
	100	$28,9°$	$20,7°$	$15,3°$	$24,0°$
20×26	50	$27,5°$	$21,9°$	$15,7°$	$23,9°$
	60	$30,7°$	$20,8°$	$16,0°$	$24,8°$
	70	$28,0°$	$18,9°$	$15,4°$	$22,7°$
	80	$28,5°$	$19,8°$	$15,9°$	$23,4°$
	100	$25,8°$	$19,1°$	$17,4°$	$21,9°$
25×33	50	$28,5°$	$19,6°$	$15,0°$	$23,2°$
	60	$29,0°$	$19,7°$	$16,0°$	$23,6°$
	70	$27,8°$	$20,0°$	$16,0°$	$23,2°$
	80	$27,5°$	$19,8°$	$15,2°$	$22,9°$
	100	$27,8°$	$18,8°$	$15,1°$	$22,5°$

Tabelle 3.2. Mittlerer Schätzfehler für den Fall bekannter Personen. Für jeweils drei verschiedene Bildgrößen wurden neuronale Netze mit unterschiedlicher Anzahl an Neuronen der mittleren Schicht (hier durch $\#N_D$ aufgeführt) implementiert und evaluiert. Die Anzahl an Einzelhypothesen gibt an wieviele Vorderkopfansichten jeweils zur Fusion eingesetzt wurden. Man erkennt dass bei Einsatz mehrerer Einzelhypothesen ein Sinken des Fehlers zu beobachten ist.

Wird in einer Kameraansicht ein Hinterkopf erkannt, so bedeutet dies für die Bildung einer Drehhypothese, dass Kopfstellungen, die zu dieser Kamera

[2]Die vertikale Neigung eines Kopfes entspricht der in Bild 1.2 beschriebenen *tilt* Drehung

[3]Unter anderem wären hiervon die dem neuronalen Netz zusätzlich angelegten Kantenbilder des Kopfes betroffen

gerichtet sind, ausgeschlossen werden können. Ein weiterer Faktor der somit einen unverhältnismäßig starken Einfluss auf die Bildung einer Gesamthypothese der Drehung nimmt stellt die Vorderkopf-Hinterkopf-Klassifikation dar, da sie Kopfansichten ggf. falsch klassifiziert und damit zu einem Ignorieren ganzer Winkelbereiche möglicher Endhypothesen führt. Die ungenügende Erkennungsrate korrekter Vorderkopfansichten von $87,9\%$ (siehe Tabelle 3.1) verlangt eine zusätzliche Verifizierung der klassifizierten Beobachtung: Anatomiemaße oder Historien werden wünschenswert um das durch die Vorderkopf-Hinterkopf-Klassifikation entstehende Rauschen zu minimieren.

Abhängig von der Anordnung der Kameras in einer Multikamera-Umgebung, scheint insbesondere eine Fusion über die Anzahl an Vorderkopfansichten interessant zu sein, die in der Regel von den Kameras erfasst wird. Im Fall dieser Arbeit betrifft das eine Fusion über zwei Vorderkopfansichten bzw. damit verbundenen Einzelhypothesen. Wie in Tabelle 3.2 ersichtlich ist, trat bei Einsatz von 60 versteckten Neuronen im Netz der Drehschätzung, 70 versteckten Neuronen im Netz der Vorderkopf-Hinterkopf-Klassifikation und einer Bildgröße von 15×20 Pixel der minimale Fehler von $18,4°$ auf. Gleichzeitig führt jene Architektur den minimalen Gesamtfehler des Systems auf, so dass im folgenden Fall unbekannter Personen eine Evaluation auf die hier entsprechenden Topologien bezogen sein soll.

3.3 Der Fall unbekannter Personen

Für eine Evaluation auf Kopfbildern von Personen die nicht im Training inbegriffen waren, wurde nach dem Leave-one-out Prinzip aus allen vorliegenden Aufnahmen jeweils eine Person dem Training vorenthalten, auf deren Aufnahme anschließend eine exklusive Evaluation geschah. Diese Vorgehensweise wurde für alle Personen eingehalten, so dass abschließend für jede aufgenommene Person eine Evaluation für den unbekannten Fall durchgeführt wurde. Die Topologie der hierfür eingesetzten neuronalen Netze richtete sich dabei nach der besten Drehschätzung bekannter Kopfbilder für den Fall einer Fusion mit zwei Kameraansichten (siehe Tabelle 3.2), da aufgrund der Anordnung der Kameras zueinander überwiegend paarweise Vorderkopfansichten erfasst werden. Die Evaluation geschah demnach mit einer Topologie von 60 versteckten Neuronen der mittleren Schicht für das Netz der Drehschätzung und 70 versteckten Neuronen für das Netz des Vorderkopf-Hinterkopf-Klassifikators sowie einer Eingabebildgröße von 15×20 Pixel.

Die Trainingsdaten wurden auch in diesem Fall gespiegelt um eine Verdoppelung der zur Verfügung stehenden Kopfbilder während der Einlernphase zu erreichen. Die Daten der Person, die dem Training vorenthalten wurde, blieben unverändert, da sie ausschließlich der Evaluation zugeschrieben wurden.

Person	# Einzelhypothesen 1	2	3	Gesamtfehler
1	$45,8°$	$28,8°$	$23,5°$	$33,1°$
2	$39,9°$	$27,6°$		$36,9°$
3	$36,8°$	$22,1°$	$12,9°$	$25,3°$
4	$44,6°$	$28,6°$	$32,1°$	$38,0°$
5	$65,7°$	$38,2°$	$44,5°$	$53,7°$
6	$34,3°$	$19,7°$	$2,9°$	$28,2°$
7	$37,3°$	$30,3°$	$14,5°$	$31,8°$

Tabelle 3.3. Mittlerer Schätzfehler für den Fall unbekannter Personen. Die Anzahl an Einzelhypothesen verdeutlicht dabei die Menge an Vorderkopfansichten die jeweils zur Fusion beigetragen haben. Auch hier im Fall unbekannter Personen wird deutlich, dass bei Einsatz mehrerer Vorderkopfansichten ein Sinken des Fehlers zu beobachten ist.

3.3.1 Analyse

Wie aus Tabelle 3.3 hervorgeht, zeigt das System auch im Fall unbekannter Personen ein vergleichbares Verhalten zur Evaluation bekannter Kopfbilder. Wie erwartet zeigt auch hier der Einsatz mehrerer Kameras in der Regel seine Auswirkungen in einer Stabilitätserhöhung der Drehschätzung. Die schlechten Ergebnisse der Personen 4 und 5 liegen unter anderem in der bereits schlechten Vorderkopferkennung begründet (siehe Tabelle 3.4). Insgesamt wird jedoch deutlich, dass eine Fusion mehrerer Einzelhypothesen mangelhafte Umstände während der Drehschätzung auszugleichen vermag und die Vorlage mehrerer Vorderkopfansichten eine deutliche Leistungssteigerung des Systems nach sich zieht.

Person	1	2	3	4	5	6	7
Korrekt	87,5%	77,9%	89,7%	80,2%	68,5%	87,2%	84,1%

Tabelle 3.4. Zuverlässigkeit der Vorderkopf-Hinterkopf-Klassifikation im Fall unbekannter Personen

Die auch hier im Fall der unbekannten Personen anzutreffende, schlechte Schätzung bei einer Hypothesenbildung durch nur eine Vorderkopfansicht un-

	# Einzelhypothesen			
	1	2	3	Gesamtfehler
bekannte Personen	$26,7°$	$18,4°$	$15,5°$	$21,8°$
unbekannte Personen	$44,1°$	$27,8°$	$21,9°$	$35,1°$

Tabelle 3.5. Gesamtfehler des Systems sowohl im Fall bekannter als auch unbekannter Personen.

terstützt die Vermutung der zu geringen Auflösung der extrahierten Kopfbilder und eventueller Fehlklassifikation durch irritierende Kanteninformationen bei Neigung des Kopfes bzw. seitlicher Aufnahme der Schläfen. Das generelle Verschlechtern der Schätzleistung liegt jedoch mitunter in der Extraktion der Köpfe begründet: die unterschiedlichen Größen und Formen der vorgefundenen Köpfe, sowie die verschiedenen Positionen der jeweiligen Personen zu den einzelnen Kameras, resultieren in unterschiedlichen Größen der extrahierten Kopfbilder. Durch die in der Vorverarbeitung stattfindende Skalierung resultiert dies in einer Weichzeichnung und damit einer Reduktion der im Kopfbild gespeicherten Informationen bezüglich der Kopfdrehung. Dadurch gehen relevante Kantenstrukturen verloren, die ohnehin geringe Auflösung der Kopfbilder wird weiter verschlechtert. Darüberhinaus entsteht durch den Einsatz einer farbbasierten Kopffindung ein erhebliches Rauschen in Form einer schwankenden Einbeziehung der Haare und einer dahinterliegenden Hintergrundszene. Zusätzlich zu den extrahierten Köpfen werden den neuronalen Netzen dadurch bisher unberücksichtigte Bilddaten angelegt was zu einer falschen Hypothesenbildung bezüglich der Vorderkopfansicht und der einzelnen Kopfdrehungen führen kann.
Eine reine Beschränkung auf Gesichtspartien innerhalb der Extraktionen könnte durch die Kombination mit einem Hautfarbmodell geschehen, fand im Rahmen dieser Arbeit jedoch nicht statt.

Kapitel 4

Zusammenfassung und Ausblick

In dieser Arbeit wurde ein System entwickelt das die horizontale Kopfdrehung einer zu beobachtenden Person schätzt. Hierzu sind vier Kameras in den Ecken eines Raums angebracht, die dem System vier unterschiedliche Ansichten der Person liefern. Der Kopf wird aus allen Aufnahmen extrahiert und dient zur jeweiligen Bildung einer Einzelhypothese. Die so entstandenen, einzelnen Schätzungen werden anschließend zu einer Gesamthypothese fusioniert. Im Fall unbekannter Personen beträgt der dabei erreichte Fehler des Systems bei einer Schätzung über zwei Kameraansichten $27,8°$, bei einer Fusion über drei Kameraansichten $21,9°$.

Das System geht hierfür in fünf Schritten vor:

1. Kopfsuche
2. Vorverarbeitung extrahierter Köpfe
3. Vorderkopf-Hinterkopf-Klassifikation
4. Schätzen einzelner Drehhypothesen
5. Fusion der Einzelhypothesen

Das vorgestellte Verfahren zur Kopfsuche baut sowohl auf einem farbbasierten Modell als auch auf einer Vordergrund-Segmentierung auf um etwaige Pixelcluster zu bilden, die durch ein Ellipsen-Suchverfahren auf ihre Ähnlichkeit hinischtlich einer vereinfachten Kopfform untersucht werden. Dabei wird jeweils die beste Hypothese - das am besten passende Pixelcluster - zurückgeliefert und extrahiert. Dabei findet noch keine Zuordnung zu einzelnen Personen statt, sollten sich mehrere Anwesende im Raum befinden. Hier wäre jedoch ein Bezug auf die übrigen Kameraansichten möglich, indem

Tiefeninformationen herangezogen werden um ein dreidimensionales Modell der Pixelcluster bzw. vorhandener Kopfkandidaten zu bilden. Um die Erweiterbarkeit des Systems zu gewährleisten wäre eine entsprechende Korrespondenzfindung bei beliebiger Anzahl eingesetzter Kameras eine zu lösende Aufgabe.
Die Ellipsen-Suche prüft Bildbereiche auf ihre Ähnlichkeit zu einer vereinfachten ellipsoiden Kopfform. Dabei werden bislang keine anatomischen Heuristiken verwendet die eine unmögliche Positionierung hypothetischer Kopfstellungen verhindern könnten. Desweiteren wird davon ausgegangen, dass der vorzufindende Kopf stets aufgerichtet anzutreffen ist: Eine Neigung zur Seite wäre in der Lage das Auffinden instabil werden zu lassen.

Die Vorverarbeitung normalisiert den Kontrast der extrahierten Kopfbilder und erstellt darüberhinaus ein vertikales als auch horizontales Kantenbild um zusammen mit den Kopfbildern Eingabemuster für die folgenden Schätzungen zu bilden.
Die im Rahmen dieser Arbeit aufgekommene Vermutung, dass die durch die Haare zustandekommenden, zufälligen Kantenmuster negative Auswirkungen auf die Hypothesenbildung nehmen könnten, bietet einen Ansatzpunkt für zukünftige Arbeiten, in denen unterschiedliche Vorverarbeitungsschritte evaluiert werden. Darunter wäre unter anderem der komplette Verzicht auf Kantenbilder möglich.

Die Aufgabe der Vorderkopf-Hinterkopf-Klassifikation ist es etwaige Ansichten von Hinterköpfen vor der eigentlichen Drehschätzung zu entfernen um der Hypothesenbildung eine optimale Menge an Vorderkopfansichten zur Extraktion genügender Information bezüglich der Kopfstellung zu ermöglichen. Das dabei erzielte Ergebnis einer korrekten Klassifikation von maximal $87,9\%$ im Fall bekannter Personen und $89,7\%$ im personenunabhängigen Fall weist Verbesserungsmöglichkeiten auf um eine stabilere Grundlage für die Drehschätzung und Fusion zu ermöglichen. Die Implementierung einer Historie zur Bildung einer Aussage hinsichtlich dem Vorliegen eines Vorder- oder Hinterkopfes könnte in Zusammenhang mit der eigentlichen Drehschätzung erfolgen: eine vorhergehende Schätzung der Kopfdrehung könnte zur Bildung einer klassenbedingten Wahrscheinlichkeit der Vorderkopf-Hinterkopf-Klassifikation herangezogen werden. Der Einsatz einer Bewegungsanalyse oder einfachen Heuristik könnte zu einer Multihypothesenbildung einer darauffolgenden Drehschätzung führen. Ein interessanter, im Rahmen der hier gestellten Aufgabe denkbarer Ansatz zur Implementierung eines Multihypothesenverfahrens stellt dabei ein Partikelfilter nach [7] dar.

Durch den erbrachten Nachweis der Fehlerreduktion bei Einsatz mehrerer Kameras wird die erweiterbare Struktur des entwickelten Systems bestärkt und bietet somit die Möglichkeit durch weiteres Hinzufügen von Kameras die Stabilität der Drehschätzung zu erhöhen. Eine nähere, relative Positionierung mehrerer Kameras zueinander macht Sinn, um die Anzahl an erfassten Vorderkopfansichten zu erhöhen als dies durch den bisherigen Aufbau möglich ist.

Zur Klassifikation von Vorder- bzw. Hinterköpfen und Schätzung der Kopfdrehung wurden neuronale Netze eingesetzt und mit unterschiedlichen Topologien evaluiert. Die Zuverlässigkeit und Einsetzbarkeit von neuronalen Netzen wurde bereits in [20] und [16] für den Einsatz einer Kopfdrehungsschätzung nachgewiesen. Eine Fusion beliebig vieler, durch neuronale Netze bestimmter Einzelhypothesen unterschiedlicher Kameraansichten wurde in dieser Arbeit vorgestellt. Die dabei eingesetzte, schnelle Vorverarbeitung extrahierter Kopfbilder und einfache Implementierung der Hypothesenfusion nach einem Maximum-Likelihood-Ansatz unterstützen dabei den Vorteil der Echtzeitfähigkeit, der durch neuronale Netze zur Verfügung steht. Darüberhinaus bietet die Konfusionsmatrix als Konfidenzmaß die einfache Möglichkeit den Einfluss einzelner Drehhypothesen bei der Fusion zu reduzieren. Mangelhafte Kameraansichten können so in ihrem Einwirken auf die Bildung einer Aussage über die Kopfdrehung abgeschwächt werden; das System kann an veränderte Umstände angepasst werden.

Anhang A

Neuronale Netze

Viele Aufgaben lassen sich durch Algorithmen auf Computer schneller lösen als der Mensch allein dazu fähig wäre. Der Mensch dagegen besitzt die Fähigkeit zusammenhängende Informationen besser zu abstrahieren und zu verstehen. Während Computer in kurzer Zeit strikt numerische Berechnungen auf beliebige Genauigkeit lösen, ermöglicht das menschliche Gehirn selbst fehlerbehaftete oder gar nicht vorhandene Information zu ersetzen bzw. zu ergänzen um das Gesamtbild zu komplettieren.

Künstliche neuronale Netze imitieren den Prozess der Informationsverarbeitung im Gehirn: Eine große Anzahl kleiner, miteinander verbundener Elemente, sog. *Neuronen* stellen das Äquivalent zu biologischen Nervenzellen dar. Information wird verarbeitet indem sich einzelne Neuronen gegenseitig, unter zuhilfenahme bestehender Verbindungen untereinander, aktivieren. Dieser Prozess geschieht analog zu den Vorgängen im Gehirn.

Ein großer Vorteil künstlicher neuronaler Netze ist ihre Lernfähigkeit gegenüber Trainingsmustern. Iterative Lernverfahren reinitialisieren Neuronen-Verbindungen ohne dass diese vorher nach einem bestimmten Muster programmiert worden sind. Desweiteren überzeugt der hohe Grad an Parallelität der Informationsverarbeitung gegenüber rein algorithmischen Verfahren:
Durch den Einsatz vieler Neuronen und ihrem untereinander gegenseitigen Zusammenspiel bei der Informationsspeicherung und -verarbeitung, verteilt sich die Aufgabenlast gleichmäßig über das komplette Netzwerk neuronaler Elemente. Bei Ausfall eines oder nur weniger Elemente in diesem komplexen Zusammenschluss bedeutet dies einen nur relativ kleinen Wissensausfall - die Gesamtfunktionsweise bleibt äquivalent zum menschlichen Gehirn erhalten.

Der vorliegende Anhang soll eine kurze Einführung in das komplexe Gebiet

der neuronalen Netze bieten. Dabei sollen insbesondere die Hintergründe dieser Arbeit näher erläutert werden, indem auf das Einlernen und Evaluieren neuronaler Netze eingegangen wird. Die Zusammenfassung stellt dabei ein Resümee unterschiedlicher Quellen dar - für weitergehende Informationen bezüglich möglicher Einsatzgebiete und Hintergründe sei der interessierte Leser daher auf [9], [2] und [14] verwiesen.

A.1 Topologie Neuronaler Netze

A.1.1 Neuronen

Neuronale Netze bestehen, wie in Abbildung A.1 schematisch dargestellt, aus miteinander verbundenen, Nervenzellen-ähnlichen Elementen, sogenannten *Neuronen*. Über gerichtete Verbindungen erhält jedes Neuron die Ausgaben aller ihm vorgeschalteten Elemente und reagiert entsprechend darauf, indem es selbst eine Ausgabe auf das anliegende Signal erzeugt und an alle ihm nachgeschalteten Elemente weiterleitet. Die Verbindungen unter den Neuronen sind dabei gewichtet und nehmen damit einen direkten Einfluss auf das durchzuleitende Signal, indem es entsprechend verstärkt oder abgeschwächt wird.

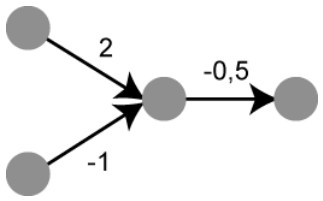

Abbildung A.1. Schematischer Aufbau eines neuronalen Netzes: gewichtete Verbindungen bestimmen wie die Information durch das Netz durchgereicht wird.

Die Eingabe eines Neurons j setzt sich zusammen als die gewichtete Summe der Ausgaben o_i aller vorgeschalteten Elemente i. Vereinfachend lässt sich die Teilmenge ausgebender Neuronen als eigenständiges Netzwerk auffassen, dessen Ausgabe diesem Neuron zur Eingabe dient. Im folgenden sei daher ein Eingangssignal eines Neurons j, zum Zeitpunkt t als $net_j(t)$ bezeichnet:

$$net_j(t) = \sum_i w_{ij} o_i(t) \tag{A.1}$$

w_{ij} gibt hierbei das Gewicht der Verbindung von Neuron i zu j und damit den Einfluss auf die nachfolgende Reaktion des Neurons j an: je ausgeprägter die Gewichtung desto stärker wirkt der Reiz den das Neuron j erfährt. Negative Gewichtungen reduzieren somit das übertragene Signal und schwächen

eine mögliche Reaktion nachfolgender Elemente entsprechend ab.

Durch ein anliegendes Signal wird ein Neuron zur Reaktion angeregt, man spricht von der *Aktivierung* des Neurons. Die Reaktion erfolgt dabei gemäß einer Aktivierungsfunktion $f(net)$, die, bezüglich eines Eingangssignals net ein entsprechendes Ausgabesignal $o = f(net)$ erstellt. Jenes Ausgabesignal wird schließlich an alle nachgeschalteten Elemente weitergeleitet, deren Reaktionen entsprechend verläuft. Verschiedene Reaktionen eines Neurons auf eine anliegende Eingabe sind denkbar, die bekanntesten drei Aktivierungen lassen sich jedoch wie folgt beschreiben:

- *Schwellwertfunktion*: Das Neuron reagiert mit einer Ausgabe von 0 sollte die Eingabe einen gewünschten Schwellwert θ nicht überschreiten. Im Gegensatz dazu erfolgt eine entsprechende Ausgabe von +1 bei Übertreten des jeweiligen Schwellwerts:

$$f(net) = \begin{cases} 0 : net \leq \theta \\ 1 : net > \theta \end{cases} \tag{A.2}$$

  Man spricht von einem Feuern der Zelle sobald sie einen hierfür notwendig großen Reiz erfährt.

- *Sigmoid-Funktion*: Die S-förmige Funktion approximiert die Ausgabe der Schwellwertfunktion, stellt jedoch eine differenzierbare Reaktion aufgrund ihrer Stetigkeit dar:

$$f(net) = \frac{1}{1 + e^{-net}} \tag{A.3}$$

- *Lineare Funktion*: Die Ausgabe eines Neurons steigt linear mit dem anliegenden Signal:

$$f(net) = net \tag{A.4}$$

A.1.2 Schichteneinteilung

Neuronale Netze dienen der Informationsverabeitung. Von angelegten Mustern wird eine entsprechende Reaktion in Form einer Ausgabe des Netzes erwartet. In der Regel wird hierzu eine *Schichteneinteilung* der Neuronen vorgenommen die zumindest eine Eingabe- als auch Ausgabeschicht des Netzes

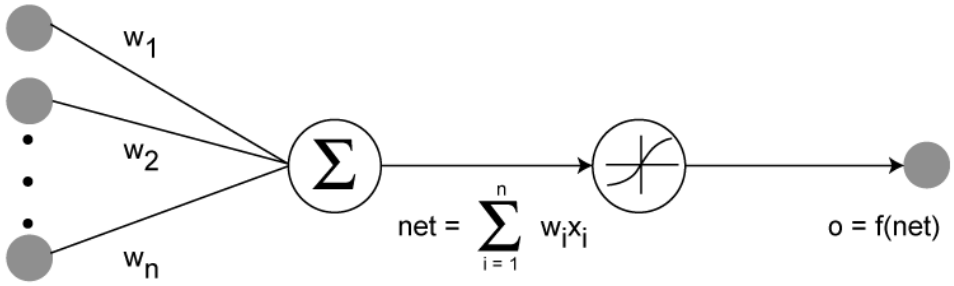

Abbildung A.2. Schematisches Zusammenspiel zwischen Neuronen: Die Eingabe eines Neurons entspricht der gewichteten Summe net aller vorgeschalteten Ausgaben. Das Neuron selbst reagiert darauf gemäß seiner Aktivierungsfunktion und erzeugt eine dementsprechende Ausgabe o.

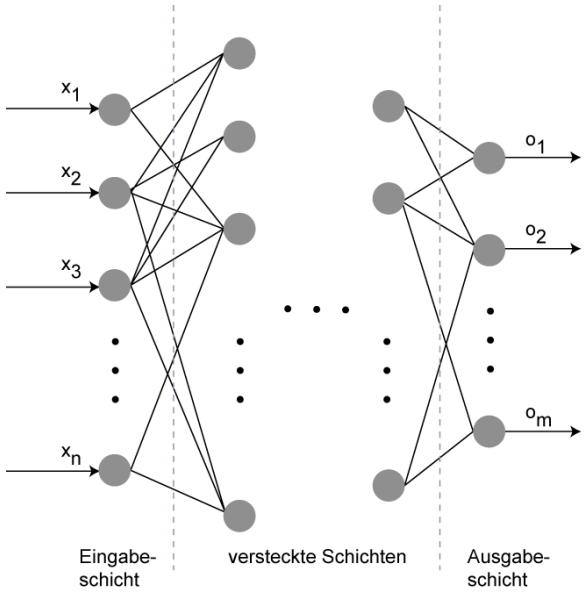

Abbildung A.3. Neuronale Netze werden oft in Schichten eingeteilt: Eine Eingabeschicht empfängt dabei angelegte Signale von außen und leitet sie nach innen an das Netz weiter. Nachdem sukzessive weitere Schichten durchlaufen werden, wird das Ergebnis der Informationsverarbeitung durch die Reaktion der Ausgabeneuronen nach außen zurückgegeben.

definiert und damit den Kontakt nach außen hin ermöglicht (siehe Abbildung A.3).
Die Eingabeschicht nimmt angelegte Muster in Form von Signalen auf und leitet sie in das neuronale Netz weiter. Anstatt einer Reaktion auf Ausgaben vorgeschalteter Elemente wird durch das angelegte Muster die Aktivierung der jeweiligen Eingabeneuronen vorgegeben. Die Menge der dadurch notwendigen Eingabeelemente korreliert somit mit der Dimension der angelegten Muster bzw. des jeweiligen Merkmalraums.
Die Aktivierungen der Neuronen in der Eingabeschicht werden durch Verbindungen zu den nachfolgenden Schichten in das Netz geleitet und verarbeitet. Eine sukzessive Aktivierung der Neuronen nachgeschalteter Schichten geschieht, die sich in der Ausgabeschicht durch eine Gesamtreaktion des Netzes in Form einer Ausgabe widerspiegelt. Soll eine diskrete Klassenzuteilung der angelegten Muster geschehen, so finden sich hier in der Regel so viele Ausgabeneuronen wie es an Klassen zu unterscheiden gibt. Durch Auswahl einer geeigneten Aktivierungsfunktion geschieht die Zuordnung zu einzelnen Klassen dabei entweder binär oder kontinuierlich in Form einer wahrscheinlichen Zugehörigkeit.

A.2 Perzeptron

Ein neuronales Netz mit einer Eingabeschicht und genau einer Ausgabeeinheit wird *Perzeptron* genannt. Man spricht von einem einschichtigen Netzwerk, dessen Architektur dabei Abbildung A.4 entspricht.

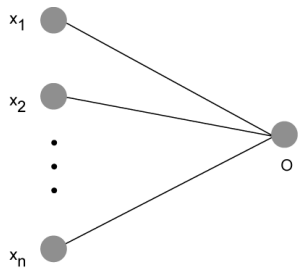

Abbildung A.4. Aufbau eines einfachen Perzeptrons: Die Eingabeneuronen erfassen ein angelegtes Signal und leiten ihre Reaktion darauf an das einzige Ausgabeneuron weiter.

Die Topologie des Perzeptrons ist *vollständig vorwärsgerichtet*: Jede der i Eingabeneuronen ist mit dem einzigen Ausgabeneuron verbunden. Die Ausgabe y des Netzwerks wird damit durch

$$y = f(net_j) = f\left(\sum_i w_i x_i\right) = f(\vec{w}^T \vec{x}) \tag{A.5}$$

berechenbar, wobei die einzelnen x_i den jeweiligen Aktivierungen der Eingabeneuronen und damit den dem Netz angelegten Signalen entsprechen.

Häufig versteht man unter einem Perzeptron ein binäres Modell, in dem Eingaben und Aktivierungen nur binäre Werte, also 0 oder 1 annehmen dürfen, während die Gewichte durchaus mit reellen Werte belegt werden können. Die Realisierung der binären Aktivierungen erfolgt hierbei durch den Einsatz einer Schwellwertfunktion gemäß Gleichung A.2. Durch die daraus entstehende Klassifikation bezüglich zweier Ausgabeklassen 0 und 1 hinsichtlich eines Schwellwerts θ, lässt sich das Perzeptron als Lösung eines *Zwei-Klassen-Problems* auffassen, deren Trennung in einem n-dimensionalen Merkmalraum durch eine *Hyperebene* geschieht:

$$\vec{w}^T \vec{x} = \theta \tag{A.6}$$

Die Klassifikation geschieht dabei durch die der Ebene zugehörigen Halbräume: Ein neu angelegtes, n-dimensionales Muster m wird demnach genau dann der Klasse 0 zugeordnet, sollte die Ausgabe des Netzes den Schwellwert θ nicht übertreffen. Eine Zuteilung zur übrigen Klasse geschieht dementsprechend durch Überschreiten des Schwellwerts.

Ein Bewertungskriterium für die Klassifikationsgüte eines Perzeptrons lässt sich durch eine möglichst optimale Ausrichtung der Hyperebene definieren, so dass die Anzahl an falsch klassifizierten Mustern minimal wird. Der dadurch beschriebene Fehler $E(\vec{w})$ bezüglich einer vorgegebenen Gewichtung $\vec{w}$ der Neuronenverbindungen, setzt sich demnach zusammen aus der Summe der einzelnen Fehler, die bei Anlegen jedes einzelnen Musters $m \in M$ beobachtet wurden:

$$E(\vec{w}) = \frac{1}{2} \sum_{m \in M} (t_m - o_m)^2 \tag{A.7}$$

t_m bezieht sich dabei auf ein gefordertes Soll-Ergebnis, o_m auf das tatsächlich durch das Netzwerk entstandene Resultat nach Anlegen eines Musters m. Abb. A.5 verdeutlicht beispielhaft den Zusammenhang zwischen der Fehlerfunktion $E(\vec{w})$ und Verteilung der Gewichte $\vec{w}$.

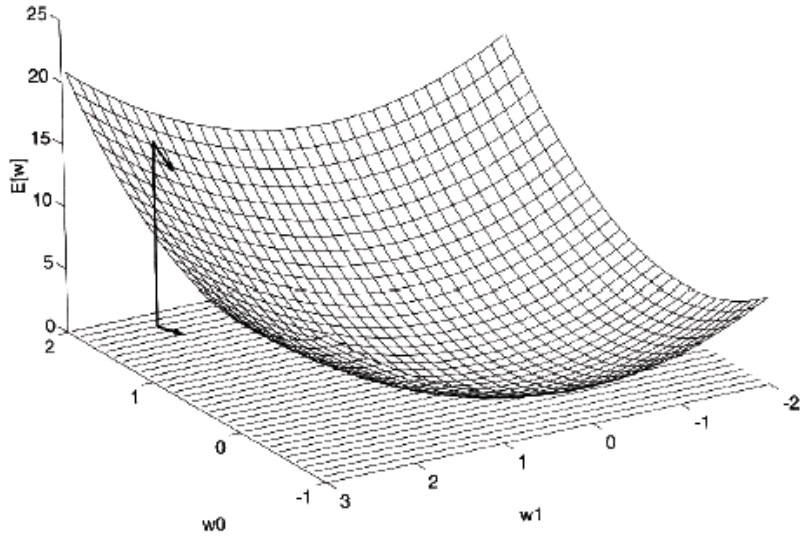

Abbildung A.5. Zusammenhang zwischen Fehlerfunktion $E(\vec{w})$ und Verteilung der Gewichte $\vec{w}$ am Beispiel eines Perzeptrons mit zwei Eingängen.

A.2.1 Delta-Lernregel

Eine Lösung zum Minimieren des Klassifikationsfehlers aus Gleichung A.7 wird durch ein Gradientenabstiegsverfahren erreicht, das, ausgehend von einer initialen Belegung, die Verbindungsgewichtungen $\vec{w}$ entgegengesetzt dem Gradienten der Fehlerfunktion verändert:

$$\forall w_i : \hat{w}_i = w_i + \eta \cdot \triangle w_i, \text{ mit } \triangle w_i = -\frac{\partial E}{\partial w_i} \tag{A.8}$$

η gibt dabei einen beliebigen Lernfaktor an, der die Schrittweite der Gewichtsveränderung und damit die Geschwindigkeit der Suche nach einem lokalen Minimum steuert.
Durch Einsetzen der Fehlerfunktion E aus Gleichung A.7 kann $\triangle w_i$ umgeschrieben werden zu

$$\triangle w_i = -\frac{\partial}{\partial w_i} \frac{1}{2} \sum_{m \in M} (t_m - o_m)^2 \tag{A.9}$$

Ein Verschieben der Differenzierung in die Summe sowie anschließende Anwendung der Kettenregel erlaubt schließlich die Berechnung von Gleichung A.9:

$$\begin{aligned}
\triangle w_i &= -\frac{\partial}{\partial w_i}\frac{1}{2}\sum_{m\in M}(t_m - o_m)^2 \\
&= -\frac{1}{2}\sum_{m\in M}\frac{\partial}{\partial w_i}(t_m - o_m)^2 \\
&= -\frac{1}{2}\sum_{m\in M}2(t_m - o_m)\frac{\partial}{\partial w_i}(t_m - o_m) \qquad \text{(A.10)}
\end{aligned}$$

Die Ableitung in obiger Gleichung verlangt den Einsatz einer differenzierbaren Aktivierungsfunktion. Wählt man eine lineare Aktivierung nach Gleichung A.4, lässt sich für die Ausgabe o_m des Ausgabeneurons daher auch

$$o_m = \sum_i w_i x_{i_m} = \vec{w}^T \cdot \vec{x}_m \qquad \text{(A.11)}$$

schreiben, womit sich Gleichung A.10 weiter auflösen lässt zu

$$\begin{aligned}
\triangle w_i &= -\frac{1}{2}\sum_{m\in M}2(t_m - o_m)\frac{\partial}{\partial w_i}(t_m - \vec{w}^T\vec{x}_m) \\
&= -\frac{1}{2}\sum_{m\in M}2(t_m - o_m)(-x_{i_m}) \\
&= \sum_{m\in M}(t_m - o_m)x_{i_m} \qquad \text{(A.12)}
\end{aligned}$$

Durch einen iterativen Prozess werden damit die Gewichte der Verbindungen verändert, bis ein gewünschter Schwellwert des Klassifikationsfehlers unterschritten wird.

A.3 Lineare Separierbarkeit

Der Einsatz eines Perzeptrons als Lösung für ein Zwei-Klassen-Problem impliziert eine *lineare Separierbarkeit* des Merkmalraums: die Beispielmuster, die an das Netz angelegt werden, müssen so verteilt sein, dass Muster der einen Klasse durch eine Hyperebene von den übrigen Mustern abgetrennt werden können. Das bekannte XOR-Problem verdeutlicht diesen Nachteil, indem es die Grenzen des einfachen Perzeptrons durch eine einfache Aufgabenstellung aufweist:

Implementiert werden soll eine XOR-Schaltung, eine Entweder-Oder-Schaltung aus der Bool'schen Algebra mit den folgenden Eigenschaften (Abb. A.6):

$$f(x, y) \mapsto \{0, 1\}, \text{wobei } x, y, \in \{0, 1\} \tag{A.13}$$

$$\text{mit } f(x, y) = \begin{cases} 0 : x = 0, y = 0 \text{ oder } x = 1, y = 1 \\ 1 : x = 1, y = 0 \text{ oder } x = 0, y = 1 \end{cases}$$

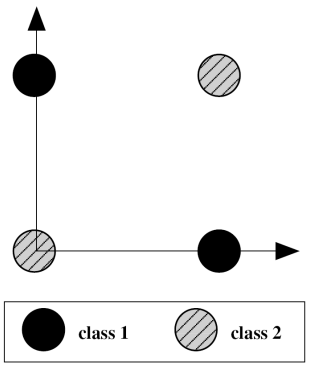

Abbildung A.6. Die beiden Klassen des XOR-Problems sind nicht linear durch eine Hyperebene separierbar.

Die beiden Klassen sind nicht durch eine Hyperebene voneinander trennbar, das einfache Perzeptron ist hier nicht einsetzbar. Mehr als eine Hyperebene ist vonnöten um die beiden Klassen zu separieren.

A.4 Das merschichtige Perzeptron

Die Hinzunahme weiterer klassifizierender Hyperebenen wird erreicht durch das Einfügen einer oder mehrerer Neuronenschichten in eine perzeptronische Topologie. Man spricht auch von einem *mehrschichtigen Perzeptron*; zusätzlich zu Eingabe- und Ausgabeschicht eingefügte Ebenen werden dabei als *mittlere* oder *versteckte* Schichten bezeichnet.
Ferner ist die Ausgabeschicht nicht mehr nur auf ein Neuron reduziert.
Die dadurch entstehende Konkatenation von Einzelnetzwerken verlangt die Verteilung des Fehlers über mehrere Schichten hinweg: das Anpassen der Gewichte geschieht nun in Abhängigkeit umliegender Neuronenschichten. Im folgenden soll daher auf die Erweiterung der allgemeinen Delta-Regel hinsichtlich mehrerer Neuronenebenen eingegangen werden.

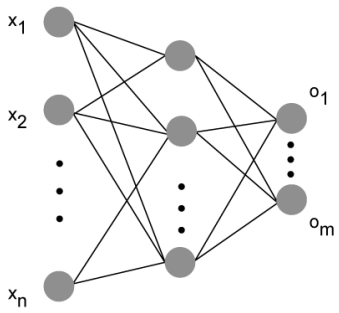

Abbildung A.7. Dreischichtiges Perzeptron mit einer versteckten Schicht.

A.4.1 Error Backpropagation

Der Klassifikationsfehler setzt sich auch in mehrschichtigen Perzeptronen durch die Summe der Einzelfehler pro angelegtem Muster $m \in M$ zusammen. Der Fehler $E_m(\vec{w})$ der dabei pro Muster entsteht lässt sich durch die Summe der einzelnen, in den jeweiligen Ausgabeneuronen entstandenen Fehler wie folgt beschreiben:

$$E_m(\vec{w}) = \frac{1}{2} \sum_{k \in Ausgabe} (t_{k_m} - o_{k_m})^2 \tag{A.14}$$

t_{k_m} gibt dabei die Sollausgabe für das Ausgabeneuron k, o_{k_m} die beobachtete Ausgabe des Ausgabeneurons bei angelegtem Muster $m \in M$ wieder.

Sei nun mit

$$\hat{w}_{ij} = w_{ij} + \eta \cdot \triangle w_{ij}, \text{ mit } \triangle w_{ij} = -\frac{\partial E}{\partial w_{ij}} \tag{A.15}$$

äquivalent zu Gleichung A.8 das Abbilden einer Verbindungsgewichtung w_{ij} von Neuron i zu Neuron j auf dessen neue Gewichtung $\hat{w}_{ij}$ anhand eines vorliegenden Fehlers E beschrieben. Durch Anwenden der Kettenregel ergibt sich eine Umformung gemäß:

$$-\frac{\partial E}{\partial w_{ij}} = -\frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial w_{ij}} \tag{A.16}$$

wobei net_j die akkumulierte Eingabe des Neurons j darstellt. Einsetzen von Gleichung A.1 ergibt dadurch:

$$
\begin{aligned}
-\frac{\partial E}{\partial w_{ij}} &= -\frac{\partial E}{\partial net_j}\frac{\partial net_j}{\partial w_{ij}} \\
&= -\frac{\partial E}{\partial net_j}\frac{\partial}{\partial w_{ij}}\sum_{i\in\text{Vorgänger}} w_{ij}o_i \\
&= -\frac{\partial E}{\partial net_j}o_i = \delta_j o_i
\end{aligned} \tag{A.17}
$$

o_i stellt die jeweilige Ausgabe des Neurons i dar, das durch die zu berechnende Verbindungsgewichtung w_{ij} mit Neuron j verbunden ist. δ_j lässt sich dagegen als Fehlersignal des Neurons j beschreiben und muss für Neuronen der Ausgabeschicht sowie der mittleren Schichten jeweils getrennt berechnet werden:

- Ausgabeschicht:

$$
\delta_j = -\frac{\partial E}{\partial net_j} = -\underbrace{\frac{\partial E}{\partial o_j}}_{1}\underbrace{\frac{\partial o_j}{\partial net_j}}_{2} \tag{A.18}
$$

Eine Umformung von Term 1 lässt sich dabei durch Einsetzen der Fehlerdefinition aus Gleichung A.14 erreichen:

$$
\begin{aligned}
\frac{\partial E}{\partial o_j} &= \frac{\partial}{\partial o_j}\frac{1}{2}\sum_{k\in\text{Ausgabe}}(t_k - o_k)^2 \\
&= \frac{\partial}{\partial o_j}\frac{1}{2}(t_j - o_j)^2 \\
&= (t_j - o_j)\frac{\partial}{\partial o_j}(t_j - o_j) \\
&= -(t_j - o_j)
\end{aligned} \tag{A.19}
$$

Für Term 2 aus Gleichung A.18 gilt entsprechend:

$$
\frac{\partial o_j}{\partial net_j} = \frac{\partial f(net_j)}{\partial net_j} \tag{A.20}
$$

Setzt man als Aktivierungsfunktion $f(net_j)$ bezüglich des Eingabesignals net_j die Sigmoid-Funktion aus Gleichung A.3, so gilt für deren Ableitung

$$
\frac{\partial f(net_j)}{\partial net_j} = \frac{\partial}{\partial net_j}\frac{1}{1+e^{-net_j}} = f(net_j)(1 - f(net_j)) = o_j(1 - o_j) \tag{A.21}
$$

Für Neuronen der Ausgabeschicht ergibt sich damit:

$$\delta_j = o_j(1 - o_j)(t_j - o_j) \tag{A.22}$$

- Mittelschicht:

$$\delta_j = -\frac{\partial E}{\partial net_j} = -\sum_{k \in \text{Nachfolger}} \frac{\partial E}{\partial net_k}\frac{\partial net_k}{\partial net_j} = -\sum_k -\delta_k \frac{\partial net_k}{\partial net_j} \tag{A.23}$$

Das Fehlersignal δ_j setzt sich also unter anderem aus den Fehlersignalen δ_k aller nachfolgenden Neuronen k zusammen die direkt mit Neuron j verbunden sind, dessen Fehlersignal δ_j gerade berechnet wird (siehe Abbildung A.8. Dasselbe gilt auch für die jeweiligen Fehlersignale δ_k, so dass hier eine Rückpropagierung des Gesamtfehlers, ausgehend von den Ausgabeneuronen, zurück bis zum Eingang des Netzes stattfindet, wodurch der Fehler gleichmäßig auf alle Neuronen verteilt wird.
Weiter auflösend ergibt sich somit für die Neuronen der mittleren Schichten:

$$\begin{aligned} -\sum_k -\delta_k \frac{\partial net_k}{\partial net_j} &= \sum_k \delta_k \frac{\partial net_k}{\partial o_j}\frac{\partial o_j}{\partial net_j} \\ &= \sum_k \delta_k w_{jk} \frac{\partial o_j}{\partial net_j} \\ &= \sum_k \delta_k w_{jk} o_j(1 - o_j) \\ &= o_j(1 - o_j)\sum_k \delta_k w_{jk} \end{aligned} \tag{A.24}$$

Die Aktualisierung des Gewichts w_{ij} von Neuron i zu Neuron j einer darauffolgenden Schicht kann also wie folgt zusammengefasst werden:

$$\hat{w}_{ij} = w_{ij} + \eta \cdot \triangle w_{ij} = w_{ij} + \eta \cdot \delta_j o_i \tag{A.25}$$

mit

$$\delta_j = \begin{cases} o_j(1 - o_j)(t_j - o_j) & \text{für Neuronen der Ausgabeschicht} \\ o_j(1 - o_j)\sum_{k \in Nachfolger} \delta_k w_{jk} & \text{für Neuronen der mittleren Schichten} \end{cases} \tag{A.26}$$

Das Vorgehen der *Fehler-Rückpropagierung* kann also wie folgt resümiert werden:

1. Initialisierung der Gewichte entlang der einzelnen Verbindungen
2. Anlegen eines Trainingsmusters m an die Eingabeneuronen und Berechnung des entsprechenden Fehlers E_m über alle Ausgabeneuronen
3. Bestimmung aller δs in Abhängigkeit des Fehlers E_m, beginnend bei den Ausgabeneuronen, rückwärtig durch alle Netzschichten
4. Berechnung der neuen Gewichtungen entlang aller Verbindungen anhand der ermittelten Fehlersignale δ und Gleichung A.25
5. Wiederholung ab Schritt 2, bis alle Trainingsmuster klassifiziert wurden und ein gewünschter Schwellwert des Klassifikationsfehlers unterschritten wurde

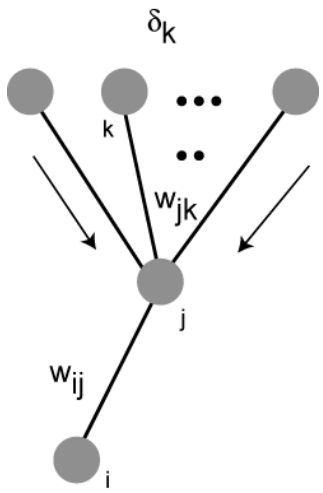

Abbildung A.8. Rückwärtspropagierung beim Backpropagation-Lernverfahren: Der Fehler des Netzes wird gleichmäßig durch die Fehlersignale δ über alle Schichten verteilt. Die Verteilung erfolgt dabei rückwärtig, indem zunächst alle Fehlersignale der Ausgabeneuronen berechnet werden, davon abhängig dann die jeweiligen δ aller Neuronen der vorhergehenden Schichten. Die Gewichte der Verbindungen werden daraufhin den jeweiligen Fehlersignalen entsprechend verändert.

A.5 Training, Testen und Evaluieren

Das Verhalten neuronaler Netze muss gegenüber einer ausgesuchten Menge an Trainingsbeispielen eingelernt werden. Entsprechend dem biologischen Vorbild ist ein untrainiertes Netz unfähig etwaige Muster zu erkennen und diese korrekt zu klassifizieren.

Der iterative Ansatz des Backpropagation-Lernverfahrens stellt ein mächtiges Lernmittel hinsichtlich der Gewichtung neuronaler Verbindungen dar, um den Klassifikationsfehler bezüglich einer Trainingsmenge an Mustern lokal zu minimieren.

Durch die gleichmäßige Verteilung des eintrainierten Wissens auf das gesamte Netz, findet eine ähnliche Generalisierung statt, wie sie im Gehirn anzutreffen ist: An das Netz angelegte Muster werden dabei in einer nur schwer nachvollziehbaren inneren Repräsentation verarbeitet und abstrahiert. Das Resultat ist eine hinreichend korrekte Verarbeitung bisher selbst unbekannter Daten.

Durch sich stetig wiederholende Trainingszyklen eines Netzes, stößt man mitunter jedoch auf den nachteiligen Effekt des sogenannten *Overfittings*: Darunter versteht man eine Überspezialisierung des Netzwerks hin zu den bisher benutzten Trainingsmustern und eine gleichzeitige Verschlechterung der Generalisierung hinsichtlich neuer, unbekannter Daten. Um den Effekt zu minimieren muss eine Überwachung des Fehlers durch eine, nicht im Training vorkommende Menge an Mustern stattfinden. Hierzu verwendet man in der Regel ein sogenanntes *cross evaluation set* an bisher unbekannten Daten, dessen Umfang in etwa 50% der Trainingsmenge ausmacht. Nach jeder Trainingsiteration wird der Gesamtfehler des Netzwerks auf dieser unbekannten Menge evaluiert und mit vorhergehenden Ergebnissen verglichen um so ein Übertrainieren zu verhindern. Abbildung A.9 verdeutlicht den Zusammenhang sich wiederholender Trainingszyklen in Bezug auf eine Überspezialisierung seitens der Trainingsmenge.

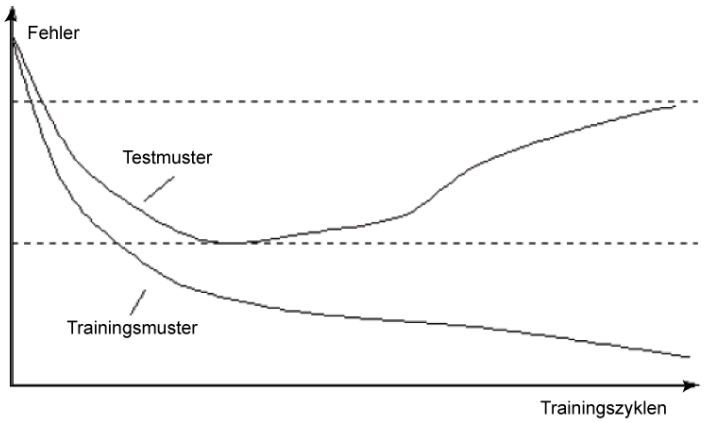

Abbildung A.9. Zusammenhang zwischen Fehlerrate und Anzahl der Trainingszyklen: Bei einer zu hohen Anzahl an Lerniterationen steigt der Fehler auf neuen, unbekannten Daten wieder an, obwohl bezüglich des Trainingssets eine weitere Abnahme zu verzeichnen ist.

Anhang B

Geometrische Modellierung

B.1 Kalibrierung

Durch die Kalibrierung werden projektive Kameraparameter bestimmt, die auf Aufnahmen dieser Kamera einwirken. Generell lassen sich dabei die folgenden Grundtypen an Einflussfaktoren unterscheiden:

1. Intrinsische Parameter
2. Extrinsische Parameter

Erstgenannte Faktoren beeinflussen das interne, optische Modell der Kamera und wirken sich damit direkt auf die Projektion einer dargestellten Szene aus. Extrinsische Parameter dagegen beschreiben die Ausrichtung und relative Anordnung mehrerer Kameras zueinander und begründen damit die unterschiedlichen Ansichten eines Objekts.

Im Rahmen dieser Arbeit wurde zur Kalibrierung die *Camera Calibration Toolbox for Matlab* [3] benutzt, deren Implementierung auch in der Open Source Computer Vision Library (OpenCV) von Intel inbegriffen ist. Die durch die Kalibrierung dabei neben den allgemeinen, externen Einflussfaktoren erhaltenen inneren Parameter, umfassen die Brennweite, radiale sowie tangentiale Verzerrung, Linsendezentrierung sowie das Seitenverhältnis der Pixelaufteilung des CCD-Sensors und beziehen sich damit auf intrinsische Parameter von Linsensystemen. Die Berechnung derselbigen erfolgt unter Zuhilfenahme mehrfacher Aufnahmen eines schachbrettartigen Musters mit allen zu kalibrierenden Kameras. Manuell vorzugebende Eckpunkte auf den verschiedenen Ansichten ermöglichen anschließend die Extraktion der beschriebenen Werte.

B.2 Triangulation

Eine durch Kalibrierung gewonnene Modellbildung eingesetzter Kameras ermöglicht es Rückschlüsse auf die dreidimensionale Struktur einer projizierten Szene zu bilden.

Sind die extrinsischen Parameter der Kameras bekannt und können in den verschiedenen Kameraansichten jeweils korrespondierende Pixel demselben Ursprungspunkt im Raum zugeordnet werden, so kann durch *Triangulation* die genaue Position jenes Punktes dreidimensional rekonstruiert werden. Die Lagebestimmung bzw. dreidimensionale Modellbildung von Objekten wird dadurch auf die Korrespondenzfindung entsprechender Merkmalspunkte über alle Ansichten hinweg reduziert.

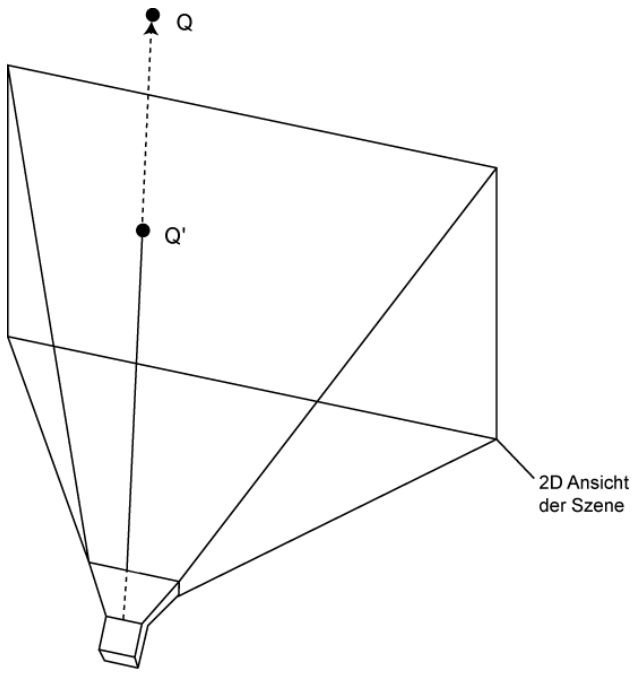

Abbildung B.1. Eine Sichtgerade vom Zentrum einer Kamera i durch einen projizierten Bildpunkt Q'_i eines dreidimensionalen Punktes Q.

Hierzu wird für jede Kamera i eine Sichtgerade konstruiert, die vom Zentrum der Kamera durch den auf die Bildfläche der Kamera projizierten Bildpunkt Q'_i von Q führt (siehe Abbildung B.1). Die Ermittlung der dreidimensionalen Position des Punktes Q besteht anschließend darin, den Schnittpunkt aller Sichtgeraden zu ermitteln (Abbildung B.2). Da durch minimale Kalibrierungs- und Korrespondenzfehler kein wirkliches Schneiden der Sichtgeraden im eigentlichen, mathematischen Sinn gefunden werden kann, wird

das durch den Schnitt der Geraden aufgestellte, überbestimmte, lineare Gleichungssystem durch Minimierung des quadratischen Fehlers gelöst. Für weitergehende Informationen bezüglich einer Implementierung sei hier auf [4] verwiesen.

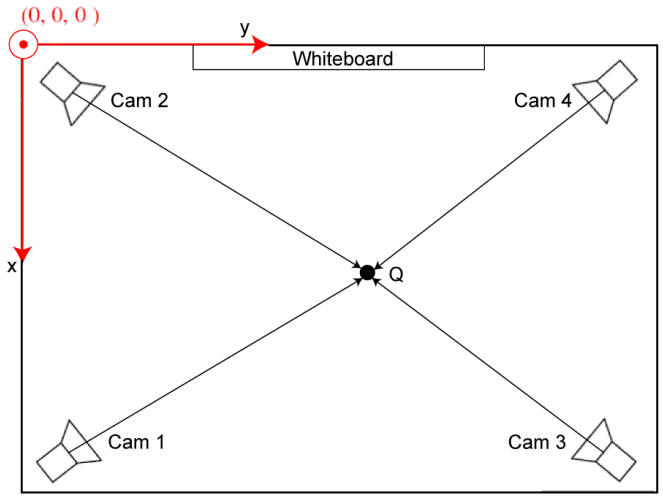

Abbildung B.2. Bei der Triangulation werden alle Sichtgeraden die von den Zentren der Kameras durch korrespondierende Bildpunkte führen geschnitten. Der Schnittpunkt beschreibt die dreidimensionale Position der jeweiligen, zweidimensionalen Bildpunkte.

B.3 Winkelberechnung

Die Angabe eines Drehwinkels geschieht immer in Abhangigkeit zu einem Bezugsystem. Im vorgestellten Ansatz wird die von einem neuronalen Netz erzeugte Hypothese der Kopfdrehung im Bezugsystem der Kamera angegeben, deren extrahiertes Kopfbild dem Netz als Eingabe diente. Um ein neuronales Netz zur Schätzung der Kopfdrehung hinsichtlich der unterschiedlichen Kameraansichten des Kopfes zu trainieren, muss deswegen der Sollwinkel, der bezüglich des Raumkoordinatensystems (Abbildung 2.1) gemessen wurde, in das Bezugsystem der jeweiligen Kamera umgerechnet werden (Abbildung B.3). Im folgenden soll daher die in dieser Arbeit implementierte Umrechnung des horizontalen Drehwinkels in eine kamerarelative Darstellung aufgeführt werden.

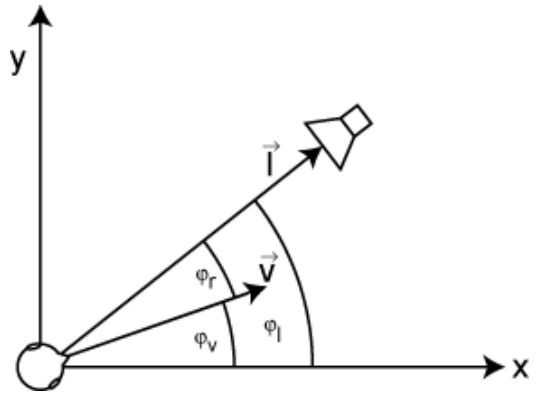

Abbildung B.3. Die Angabe eines Drehwinkels erfolgt immer in Abhängigkeit zu einem Bezugsystem. Der horizontale Drehwinkel des Kopfes kann hier sowohl durch φ_v in Bezug auf das Raumkoordinatensystem als auch durch φ_r, relativ zur Sichtlinie der Kamera, beschrieben werden.

B.3.1 Transformation eines Sichtvektors in eine kamerarelative Winkeldarstellung

Sei im folgenden $\vec{v}$ der Richtungsvektor der Kopfdrehung im Raum sowie $\vec{l_i}$ der Vektor der Sichtlinie vom Zentrum der Kamera i zum Kopfzentroiden. Ferner sei mit $\vec{\hat{v}} = (v_1, v_2, 0)^T$ bzw. $\vec{\hat{l}_i} = (l_{i1}, l_{i2}, 0)^T$ die orthogonale Projektion des Vektors $\vec{v}$ bzw. $\vec{l_i}$ auf die x/y-Ebene des in Abbildung 2.1 dargestellten Raumkoordinatensystems gegeben. Die horizontalen Drehwinkel[1] der beiden Vektoren $\vec{v}$ und $\vec{l_i}$ lassen sich jeweils durch das Winkel-Argument der Polarkoordinatendarstellung $(r_{\hat{v}}, \varphi_{\hat{v}})$ und $(r_{\hat{l}_i}, \varphi_{\hat{l}_i})$ ihrer Projektionen ausdrücken. Die Berechnung des Winkel-Arguments für den Blickrichtungsvektor $\vec{\hat{v}}$ erfolgt dabei gemäß:

$$\varphi_{\hat{v}} = \begin{cases} arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right) & \text{für } \hat{v}_1 > 0 \text{ und } \hat{v}_2 > 0 \\ arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right) + 2\pi & \text{für } \hat{v}_1 > 0 \text{ und } \hat{v}_2 < 0 \\ arctan\left(\frac{\hat{v}_2}{\hat{v}_1}\right) + \pi & \text{für } \hat{v}_1 < 0 \\ \frac{\pi}{2} & \text{für } \hat{v}_1 = 0 \text{ und } \hat{v}_2 > 0 \\ \frac{3\pi}{2} & \text{für } \hat{v}_1 = 0 \text{ und } \hat{v}_2 < 0 \end{cases} \tag{B.1}$$

Äquivalent ist für den Vektor $\vec{\hat{l}_i}$ vorzugehen.
Eine relative Angabe des horizontalen Drehwinkels des Vektors $\vec{v}$ zur Sichtlinie der Kamera i lässt sich nun durch

[1] Der horizontale Drehwinkel entspricht dabei dem *pan* Drehwinkel der in Abbildung 1.2 dargestellten Kopfdrehungen

$$\varphi_{r_i} = \begin{cases} \varphi_{\hat{v}} - \varphi_{\hat{l}_i} & \text{für } \mid \varphi_{\hat{v}} - \varphi_{\hat{l}_i} \mid \leq \pi \\ 2\pi + (\varphi_{\hat{v}} - \varphi_{\hat{l}_i}) & \text{für } \varphi_{\hat{v}} - \varphi_{\hat{l}_i} < -\pi \\ -(2\pi - (\varphi_{\hat{v}} - \varphi_{\hat{l}_i})) & \text{für } \varphi_{\hat{v}} - \varphi_{\hat{l}_i} > \pi \end{cases} \tag{B.2}$$

mit $\varphi_{r_i} \in [-\pi, +\pi]$, angeben.

B.3.2 Transformation einer kamerarelativen Winkeldarstellung in einen Sichtvektor

Sei gemäß Abschnitt B.3.1 φ_{r_i} die zu Kamera i relative Angabe des horizontalen Drehwinkels der Blickrichtung des beobachteten Kopfes. Ferner sei $\vec{l_i}$ der Vektor der Sichtlinie vom Zentrum der Kamera i zum Kopfzentroiden sowie $\vec{\hat{l}_i} = (l_{i1}, l_{i2}, 0)^T$ seine orthogonale Projektion auf die x/y-Ebene des gegebenen Raumkoordinatensystems und $\varphi_{\hat{l}_i}$ der gemäß Abschnitt B.3.1 berechnete Drehwinkel der Sichtlinie der Kamera. Für den auf die x/y-Ebene projizierten Richtungsvektor $\vec{\hat{v}}$ der Kopfdrehung gilt somit:

$$\vec{\hat{v}} = \begin{pmatrix} cos(\varphi_{\hat{v}}) \\ sin(\varphi_{\hat{v}}) \\ 0 \end{pmatrix} \tag{B.3}$$

wobei $\varphi_{\hat{v}}$ dem horizontalen Drehwinkel gemäß dem beschriebenen Raumkoordinatensystem entspricht und durch

$$\varphi_{\hat{v}} = \begin{cases} \varphi_{\hat{l}_i} + \varphi_{r_i} - 2\pi & \text{für } \varphi_{\hat{l}_i} + \varphi_{r_i} > 2\pi \\ \varphi_{\hat{l}_i} + \varphi_{r_i} + 2\pi & \text{für } \varphi_{\hat{l}_i} + \varphi_{r_i} < 0 \\ \varphi_{\hat{l}_i} + \varphi_{r_i} & \text{sonst} \end{cases} \tag{B.4}$$

berechnet werden kann.

Literaturverzeichnis

[1] Birchfield, S.: *Elliptical head tracking using intensity gradients and color histograms*, 1998.

[2] Bishop, C.: *Neural networks for Pattern Recognition.* Oxford University Press, 1995.

[3] Bouguet, Jean-Yves: *Camera Calibration Toolbox for Matlab.* `http://www.vision.caltech.edu/bouguetj/calib_doc/`.

[4] Focken, Dirk: *Vision-based 3-D Tracking of People in a Smart Room Environment.* Diplomarbeit, Universität Karlsruhe, 2002.

[5] Heinzmann, Jochen und Alexander Zelinsky: *3-D Facial Pose and Gaze Point Estimation using a Robust Real-Time Tracking Paradigm.* In: *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, Seiten 142–147, 1998.

[6] Horprasert, Thanarat: *Computing 3-D Head Orientation from a Monocular Image Sequence*, 1996.

[7] Isard, M. und A. Blake: *Condensation – conditional density propagation for visual tracking*, 1998.

[8] McKenna, S. J. und S. Gong: *Face Recognition from Sequences Using Models of Identity.* Lecture Notes in Computer Science, 1351, 1997.

[9] Mitchell, Tom M.: *Machine Learning.* McGraw Hill, 1997.

[10] Park, Sangho und J. K. Aggarwal: *Head Segmentation and Head Orientation in 3D space for Pose Estimation of Multiple People.* In: *IEEE proc. Southwest Symposium on Image Analysis and Interpretation, 2000, Austin, Texas, USA*, 2000.

[11] Peer, Peter, Jure Kovac und Franc Solina: *Human Skin Colour Clustering for Face Detection.*

[12] Qian, Richard J., M. Ibrahim Sezan und Kristine E. Matthews: *Face Tracking Using Robust Statistical Estimation.*

[13] Rae, Robert und Helge Ritter: *Recognition of Human Head Orientation Based on Artificial Neural Networks.* IEEE Transactions on Neural Networks, 9(2):257–265, 1998.

[14] Rojas, P.: *Theorie der Neuronalen Netze - Eine systematische Einführung.* Springer Verlag, 1993.

[15] Ruusuvuori, Johanna: *Looking means listening: coordinating displays of engagement in doctor-patient interaction.* 2001.

[16] Seemann, Edgar: *Estimating Head Orientation with Stereo Vision.* Diplomarbeit, Universität Karlsruhe, 2003.

[17] Srinivasan, S. und K. L. Boyer: *Head Pose Estimation Using View Based Eigenspaces.* In: *Intl. Conf. on Pattern Recognition, Quebec*, 2002.

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

[19] Stiefelhagen, R., J. Yang und A. Waibel: *Tracking Focus of Attention for Human-Robot Communication.* In: *IEEE-RAS International Conference on Humanoid Robots - Humanoids 2001*, 2001.

[20] Stiefelhagen, Rainer: *Tracking and Modeling Focus of Attention in Meetings.* Doktorarbeit, Universität Karlsruhe, 2002.

[21] Tian, Ying-Li, Lisa Brown, Jonathan Connell, Sharat Pankanti, Arun Hampapur, Andrew Senior und Ruud Bolle: *Absolute Head Pose Estimation From Overhead Wide-Angle Cameras.* 2003.

[22] Yang, Jie, Weier Lu und Alex Waibel: *Skin-Color Modeling and Adaptation.* In: *ACCV (2)*, Seiten 687–694, 1998.

[23] Yang, R. und Z. Zhang: *Model-based Head Pose Tracking With Stereovision*, 2002.