

Tiefen-basierte Bestimmung der Kopfposition und -orientierung im Fahrzeuginnenraum

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)
genehmigte

Dissertation

von

Anke Schwarz

aus Stuttgart

Tag der mündlichen Prüfung: 05. Februar 2018

Hauptreferent: Prof. Dr.-Ing. Rainer Stiefelhagen

Korreferent: Prof. Dr.-Ing. J. Marius Zöllner

Kurzzusammenfassung

Die Orientierung und Position des Kopfes sind wichtige Informationen für die Sicherheit und die Aufmerksamkeit des Fahrers. Zum einen leitet sich daraus die Kopfposition des Fahrers für Sicherheitsaspekte ab und zum anderen ist eine Abschätzung der Blickrichtung für die Aufmerksamkeit des Fahrers möglich. Für die Bestimmung der Kopfpose während des Autofahrens ist eine nicht-invasive Erfassung ohne störende Objekte, die am Kopf befestigt werden müssen, notwendig. Deshalb eröffnen visuelle Systeme mit Bildverarbeitungsalgorithmen für die Erfassung der Kopfpose eine vielversprechende Lösung. Diese Arbeit analysiert die herausfordernde Bestimmung der tiefen-basierten Kopfpose im Fahrzeuginnenraum unter realen Bedingungen, indem ein Datensatz und neue Algorithmen als Lösungsansätze vorgestellt werden.

Algorithmen zur Bestimmung der Kopfpose im Fahrzeuginnenraum müssen im Fahrzeugkontext vorhandenen Bedingungen und Einflüssen gerecht werden. Einerseits muss das Verfahren mit den limitierten Rechenkapazitäten auskommen und andererseits robuste und akkurate Werte für die Drehung und Position des Kopfes in Echtzeit liefern. Akkurate und robuste Werte für die Kopfpose müssen für die im Fahrzeugkontext möglichen Einflüsse in einem weiten Winkelbereich sichergestellt sein. Während die Robustheit gegenüber Fremdlicht durch geeignete Sensoren oder zusätzliche Beleuchtungsmodule verbessert werden kann, muss der Bildverarbeitungs-Algorithmus eine akkurate und robuste Bestimmung der Kopfpose auch bei Verdeckungen im Gesichtsbereich liefern. Neben Verdeckungen durch die Hand oder das Lenkrad ist der Einfluss von Brillen und Sonnenbrillen auf die Genauigkeit der Algorithmen von großem Interesse, da hier Abschätzungen der Blickrichtung anhand der Pupillen aufgrund von Reflexionen in den Augenbereichen häufig fehlerbehaftet sind.

Zur Analyse der Algorithmen für die Bestimmung der Kopfpose im Fahrzeuginnenraum sind Datensätze notwendig, die Referenzwerte der Orientierung und Position des Kopfes beinhalten. In der Wissenschaft sind zahlreiche Kopfposen-Datensätze vorhanden zur Evaluation von Algorithmen, basierend auf Grauwert- und Tiefenbildern in Laborszenarien sowie auf Grauwertbildern während Fahrszenarien. Allerdings fehlt für die ausführliche Auswertung im Fahrzeuginnenraum ein Datensatz, der sowohl Grau- als auch Tiefendaten während realer Autofahrten unter den oben

genannten Einflüssen beinhaltet. Im Rahmen dieser Arbeit wird ein Verfahren zur Referenzmessung der Kopfpose vorgestellt, welches eine akkurate Bestimmung der Orientierung und Position des Kopfes während realer Autofahrten ermöglicht. Mit diesem Referenzsystem wird ein Datensatz erzeugt, der Tiefen- und Infrarotdaten beinhaltet und neben den Referenzwerten der Kopfpose für jedes Bild manuelle Notationen für unterschiedliche Verdeckungen beinhaltet. Dabei wird zwischen Brillen, Sonnenbrillen und anderen Verdeckungen im Gesichtsbereich unterschieden. Der selbst aufgenommene *DriveAHead*-Datensatz beinhaltet 20 unterschiedliche Probanden sowie insgesamt über eine Million annotierte Infrarot- und Tiefendaten und steht der wissenschaftlichen Gemeinschaft zur Verfügung.

Die limitierten Rechenkapazitäten im Fahrzeug erfordern effiziente Algorithmen. Daher wird im Rahmen dieser Arbeit ein regressions-basiertes Verfahren zur Bestimmung der Orientierung und Position des Kopfes vorgestellt. Der Algorithmus bestimmt stufenweise mithilfe von Entscheidungsbäumen und einer linearen Regression die Kopfposition und -orientierung aus Tiefendaten. Im Vergleich zu vorhandenen Verfahren in der Literatur erreicht die regressions-basierte Methode im Labor vergleichbare Ergebnisse zu den Verfahren mit dem Stand der Forschung bei einer geringeren Rechenzeit pro Bild.

Im Fahrzeugkontext ist von großem Interesse, welche Art von Eingangsdaten (z.B. Tiefendaten, Infrarotbilder) zu einer hohen Genauigkeit der Bestimmung der Kopfpose führen. Deshalb wird im Rahmen dieser Arbeit eine tiefe neuronale Netzarchitektur zur Bestimmung der Orientierung des Kopfes vorgestellt, die sowohl auf den einzelnen Modalitäten, Infrarot- oder Tiefenbildern, angewendet wird als auch um Fusionsverfahren zur Kombination beider Modalitäten erweitert wird. Dabei wird eine *frühe Fusion* vorgestellt, bei der die Modalitäten bereits zu Beginn, und eine *späte Fusion*, bei der die Ausgänge zweier Sub-Netzwerke am Ende kombiniert werden. Im Gegensatz dazu findet bei der zusätzlich präsentierten *stitch-basierten Fusion* ein Austausch innerhalb der Sub-Netzwerke statt. Durch die Verwendung derselben Basis-Architektur sowohl für die tiefen neuronalen Netze auf einer Modalität als auch bei den Fusionen ist eine direkte Vergleichbarkeit der Verfahren und verwendeten Modalitäten möglich.

Der *DriveAHead*-Datensatz ermöglicht eine ausführliche Analyse der Genauigkeit von Kopfposen-Algorithmen im Fahrzeugkontext. Für diese Analyse wird eine neue Metrik, der *Balanced Mean Angular Error (BMAE)*, vorgestellt. Diese erlaubt erstmalig die aussagekräftige Bewertung über den kompletten Orientierungsbereich des Kopfes auch bei nicht gleichverteilten Daten, indem die häufige frontale Kopfausrichtung durch eine Gewichtung der Daten berücksichtigt wird. Mit dieser Metrik und der euklidischen Distanz für die Translation des Kopfes werden die in dieser Arbeit vorgestellten Methoden und Verfahren nach dem Stand der Forschung auf dem *DriveAHead*-Datensatz ausgewertet. Dabei wird neben der Genauigkeit der Algorithmen die Robustheit gegenüber Einflüssen von Brillen, Sonnenbrillen und weiteren Verdeckungen im Gesichtsbereich analysiert. Mit dem vorgestellten regressions-basierten Verfahren kann die Position des Kopfes im Fahrzeuginnenraum durch wenig Rechenaufwand mit einem Tiefenfehler von durchschnittlich 5 Millimetern

bestimmt werden. Damit liefert das Verfahren eine für die Anwendung im Fahrzeuginnenraum ausreichend hohe Genauigkeit für die Translation, während die tiefe neuronale Netzstruktur bei der Orientierung des Kopfes für starke Kopfdrehungen besser generalisieren kann. Beim Vergleich der Modalitäten, Grau- und Tiefenbilder, liefert das präsentierte tiefe neuronale Netz auf Tiefenbildern geringfügig bessere Ergebnisse als auf Infrarotdaten. Die Kombination beider Modalitäten mit der *späten Fusion* zeigen eine noch genauere Bestimmung der Orientierung des Kopfes. Das regressions-basierte Verfahren und das tiefe neuronale Netz sind robust gegenüber Verdeckungen im Gesichtsbereich sowie Brillen und Sonnenbrillen. Bei den tiefen neuronalen Netzen zeigt sich sogar eine Verbesserung der Genauigkeit bei Brillen, woraus folgt, dass die Rahmen der Brillen als zusätzliche Eigenschaften zur Erkennung der Kopfpose dienen. Mit der hohen Genauigkeit und der Robustheit zeigen die mit dieser Arbeit vorgestellten Kopfposen-Algorithmen Potenzial für den Einsatz im Fahrzeuginnenraum.

Danksagung

Auf Seiten des Karlsruher Instituts für Technologie möchte ich mich besonders bei Prof. Dr.-Ing. Rainer Stiefelhagen des Lehrstuhls für Computer Vision for Human-Computer Interaction (CV:HCI) bedanken. Die sehr gute wissenschaftliche Betreuung durch zahlreiche Diskussionen, hilfreiche Ideen und Gespräche haben die Dissertation maßgeblich geprägt. Prof. Dr.-Ing. Johann Marius Zöllner danke ich für die Übernahme der Zweitbegutachtung als Korreferent.

Den Mitarbeitern der Forschungsgruppe CV:HCI des Karlsruher Instituts für Technologie danke ich für die offene Aufnahme in Diskussionen und die Hilfsbereitschaft. Insbesondere danke ich Monica-Laura Haurilet für die unvergessliche Zusammenarbeit und die zahlreichen Diskussionen.

Meinen Kolleginnen und Kollegen bei der Robert Bosch GmbH danke ich für die außerordentliche Unterstützung, die zahlreichen Diskussionen und die motivierende Arbeitsatmosphäre. Insbesondere möchte ich mich bei Ernst Schermann, Rüdiger Henn, Torsten Maka, Zhuang Lin, Alexander Müller, Roland Wolman, Esther-Sabrina Wacker, Johannes Pallauf und Berthold Käferstein für die organisatorische und inhaltliche Unterstützung bedanken. Mein Dank geht auch an die studentischen Mitarbeiter, die mich bei Teilaspekten der Arbeit unterstützt haben.

Abschließend bedanke ich mich von Herzen bei meinem Freund Peter, meinen Eltern, meiner Schwester Maren, Isabel und meinen Freunden, die mich immer unterstützt haben in meinen Vorhaben und mir den nötigen Halt gegeben haben.

Stuttgart, im Mai 2018

Anke Schwarz

Inhaltsverzeichnis

1. Einleitung	1
1.1. Ziel der Arbeit	2
1.2. Hauptbeiträge	2
1.3. Gliederung	4
1.4. Bereits veröffentlichte Beiträge	6
2. Stand der Forschung	7
2.1. Kopfposen-Datensätze	8
2.1.1. Datensätze aus dem Labor	8
2.1.2. Datensätze aus dem Fahrzeuginnenraum	10
2.2. Koordinatensystem des Kopfes	11
2.3. Anforderungen an die Bestimmung der Kopfpose im Fahrzeuginnenraum	12
2.4. Algorithmen zur Bestimmung der Kopfpose	13
2.4.1. 2D-basierte Verfahren	13
2.4.2. Tiefen-basierte Verfahren	17
2.4.3. Fusionsverfahren von 2D- und Tiefendaten	20
2.5. Quaternionen zur Beschreibung von Rotationen	22
3. Kopfposen-Datensatz im Fahrzeuginnenraum	23
3.1. Ziel	24
3.2. Methodik der Datenaufnahme	25
3.2.1. Messaufbau	25
3.2.2. Spezifikation der Datenaufnahme	26
3.3. Auswahl eines Tiefen-Sensors	30
3.3.1. Anforderungen	30
3.3.2. Vergleich verschiedener TOF-Sensoren	31
3.3.3. Robustheit gegen Umgebungslicht	35
3.3.4. Ergebnisse	40
3.4. Methodik der Datennotation	41
3.4.1. Definition des Kopfkoordinatensystems	42
3.4.2. Referenzmessung der Kopfpose	44
3.4.3. Manuelle Annotation	45

3.5.	Charakterisierung des Referenzsystems	46
3.5.1.	Experimenteller Aufbau	47
3.5.2.	Statistische Analyse der Streuung des Referenzsystems	49
3.5.3.	Fazit	54
3.6.	Beschreibung der Stichprobe	55
3.7.	DriveAHead-Datensatz	57
4.	Regressions-basierte Bestimmung der Orientierung und Position des Kopfes	59
4.1.	Ziel	60
4.2.	Trainingsphase	61
4.2.1.	Binäre Eigenschaftsvektoren	61
4.2.2.	Lineare Regressionsmatrix	62
4.2.3.	Stufenweise Aktualisierung der Kopfpose	63
4.3.	Testphase: Bestimmung der Kopfpose	63
4.4.	Evaluation auf Labordaten	63
4.4.1.	BIWI Kinect-Datensatz	64
4.4.2.	Referenzpositionen innerhalb des Gesichtes	64
4.4.3.	Optimierung der Parameter	64
4.4.4.	Ergebnisse der 4-fachen Kreuzvalidierung	67
4.5.	Fazit	69
5.	Tiefes neuronales Netz zur Bestimmung der Orientierung des Kopfes	71
5.1.	Ziel	72
5.2.	Architektur des Head Pose Networks (HPN)	72
5.3.	Fusionsverfahren	74
5.3.1.	Frühe Fusion	74
5.3.2.	Späte Fusion	75
5.3.3.	Stitch-basierte Fusion	76
5.4.	Evaluation	77
5.4.1.	Heatmaps zur Visualisierung des Einflusses von Pixel-Regionen.	77
5.5.	Fazit	79
6.	Evaluation auf Kopfposendatensatz im Fahrzeuginnenraum	81
6.1.	Evaluationsmetrik	82
6.1.1.	Balanced Mean Angular Error (BMAE)	82
6.1.2.	Translation	83
6.2.	Vorverarbeitung	83
6.2.1.	Entfernung der sichtbaren Marker	84
6.2.2.	Detektion der Gesichtsregion	86
6.3.	Angewendete Methoden	87
6.3.1.	Prior	88
6.3.2.	Verfahren mit dem Stand der Forschung	88
6.3.3.	Verfahren aus Kapitel 4 und 5	89
6.4.	Orientierung des Kopfes	89
6.4.1.	Evaluation der Methoden	89

6.4.2. Einfluss der Modalitäten	92
6.4.3. Einfluss von Verdeckungen	94
6.4.4. Einfluss von Brillen und Sonnenbrillen	95
6.5. Translation des Kopfes	97
6.5.1. Einfluss der Modalitäten	98
6.5.2. Einfluss von Verdeckungen	98
6.5.3. Einfluss von Brillen und Sonnenbrillen	99
6.6. Fazit	100
7. Zusammenfassung	103
7.1. Einschränkungen und Ausblick.	105
A. DriveAHead Probanden	107
B. DriveAHead Beispielbilder	109
Eigene Veröffentlichungen	111
Literaturverzeichnis	113

Kapitel 1

Einleitung

Unaufmerksamkeit stellt eine der häufigsten Ursachen für Verkehrsunfälle dar. In der Studie von [Rueda-Domingo et al. \(2004\)](#) konnte gezeigt werden, dass die Anzahl der Unfälle durch die Anwesenheit eines Beifahrers drastisch verringert werden kann. Allerdings folgt aus der Statistik des [U. S. Department of Transportation \(2003\)](#), dass häufig Fahrzeuge ohne weitere Insassen unterwegs sind. Daher wird intensiv an einer Ersetzung des Beifahrers durch ein intelligentes Fahrerassistenzsystem gearbeitet. Hierfür werden aktuell innovative Ansätze untersucht, die den Fahrer in kritischen Situationen warnen sollen.

Neben der Erfassung von Geschehnissen auf der Straße soll das intelligente Fahrerassistenzsystem auch Bewegungen im Fahrzeuginnenraum erfassen. Durch die Erfassung im Innenraum können Warnungen und Sicherheitsvorkehrungen den Fahrer in kritischen Situationen unterstützen. Um diese Systeme an die aktuellen Bedürfnisse des jeweiligen Fahrers zu adaptieren, ist es essentiell, die Sitzhaltung und das Verhalten zu analysieren.

Im Fahrzeuginnenraum ermöglicht es eine optische Erfassung des Fahrers mithilfe einer Kamera, den Fahrerzustand und die Fahrerintention zu erkennen. Sowohl für die Erkennung des Zustands und der Intention des Fahrers, stellen die Aufmerksamkeit und die Blickrichtung elementare Komponenten dar. [Trefflich \(2010\)](#) konnte in seiner umfangreichen Analyse darstellen, dass die Aufmerksamkeit des Fahrers mit seiner Kopforientierung korreliert. Von der Kopforientierung wird eine Aufmerksamkeitsregion abgeleitet und der Fahrer als abgelenkt eingestuft, sobald sich diese Region für eine bestimmte Dauer außerhalb einer Grenzregion befindet. Eine weitere Untersuchung des Zusammenhangs der Kopfpose und der Blickrichtung von [Fridman et al. \(2016\)](#) ergibt, dass im Falle von starken Kopfdrehungen die Kopfpose der ausschlaggebende Faktor für die Blickrichtung ist. Aus diesen und anderen wissenschaftlichen Beiträgen folgt, dass die Orientierung und Position des Kopfes des Fahrers eine wichtige Rolle spielen.

Eine robuste und akkurate Bestimmung der Kopfpose ist für die Fahrerbeobachtung essentiell. Allerdings stellt die Bestimmung der Kopforientierung und -position

mittels bildgebender Verfahren, insbesondere deren Einsatz im Fahrzeug, eine große Herausforderung dar. In der Literatur sind zahlreiche Ansätze zur Bestimmung der Kopforientierung und -position in Laborumgebungen zu finden. Vor allem Algorithmen, die Tiefendaten verwenden, zeigen vielversprechende Ergebnisse. Um deren Robustheit und Genauigkeit im sich bewegenden Fahrzeug zu untersuchen, ist ein repräsentativer Datensatz mit realen Fahrten erforderlich. Hierbei ist von großem Interesse, ob tiefen-basierte Algorithmen im Vergleich zu 2D-basierten Verfahren besser geeignet sind. Zusätzlich kann eine Fusion beider Modalitäten die Resultate weiter verbessern. Da im Automobilkontext nur eine limitierte Rechenkapazität zur Verfügung steht, ist, neben der Exaktheit der Algorithmen, deren Effizienz von enormer Wichtigkeit.

1.1. Ziel der Arbeit

Das Ziel dieser Arbeit ist die robuste und effiziente Bestimmung der Kopfpose im Fahrzeug. Zur Bestimmung der Kopfpose aus Tiefendaten wird im Rahmen der Arbeit ein Algorithmus vorgestellt, der speziell auf die Bedürfnisse im Fahrzeug mit wenigen Rechenressourcen auskommt. Für die Validierung von Kopfposenalgorithmen unter realen Bedingungen wird im Rahmen dieser Arbeit ein neuer Datensatz im Fahrzeug entwickelt. Dieser Datensatz beinhaltet als erster Kopfposendatensatz Infrarot- und Tiefendaten mit einer akkuraten Referenzmessung der Kopforientierung und -position, aufgenommen während realer Autofahrten. Der neue Datensatz ermöglicht eine ausführliche Evaluation von Algorithmen auf dem neuesten Stand der Technik im Fahrzeug sowie die Anpassung von Deep Learning Modellen zur Kopfposenschätzung. Zusätzlich zur Evaluierung von Algorithmen auf 2D- oder Tiefendaten werden Fusionsverfahren der beiden Modalitäten für die Bestimmung der Kopfpose im Fahrzeug entwickelt. Im Fahrzeug ist die Effizienz der Algorithmen ausschlaggebend, deshalb wird zusätzlich ein effizienter Algorithmus zur Kopfposenschätzung aus Tiefendaten vorgestellt.

1.2. Hauptbeiträge

Die Hauptbeiträge der Arbeit sind:

1. Entwicklung eines Datensatzes mit Referenzmessungen der Kopfposition und -orientierung zur Evaluation von Algorithmen basierend auf Infrarot- und Tiefendaten.
2. Entwicklung eines Algorithmus zur echtzeitfähigen Bestimmung der Kopfposition und -orientierung aus Tiefendaten.
3. Ausführliche Validierung von Algorithmen auf Infrarot- und Tiefendaten sowie die Entwicklung und Bewertung von Fusionsverfahren der beiden Modalitäten.

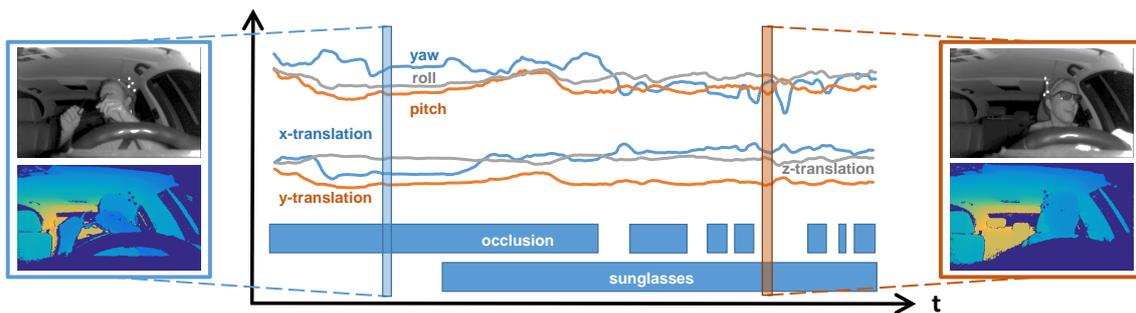


Abbildung 1.1.: Übersicht des Datensatzes. Für jedes Bild ist eine Referenzmessung der Orientierung und Position vorhanden sowie die manuellen Annotationen für Verdeckungen, Brillen und Sonnenbrillen. Aus Schwarz et al. (2017) © 2017 IEEE.

Entwicklung eines Kopfposendatensatzes im Fahrzeug

Für eine sinnvolle Evaluation von Algorithmen zur Erkennung der Kopfpose im Fahrzeuginnenraum ist ein realistischer Datensatz mit Referenzmessungen notwendig, der schwierige Szenarien für Methoden der Bildverarbeitung während Fahr Szenarien berücksichtigt. Zum einen stellt die akkurate Bestimmung der Orientierung und Position eine anspruchsvolle Aufgabe dar und zum anderen muss eine Robustheit für Fahrer mit Brillen, Sonnenbrillen und bei weiteren Verdeckungen im Gesichtsbereich sichergestellt werden. In dieser Arbeit wird ein Datensatz vorgestellt und verwendet, der während realer Fahrten im Fahrzeuginnenraum aufgenommen wird und diese Herausforderungen adressiert. Zusätzlich zu Infrarot- und Tiefendaten wird die Kopforientierung und -position mit einem Referenzsystem gemessen. Die Arbeit umfasst zunächst die Anpassung eines exakten Motion Capture-Systems an die Messung der Kopforientierung und -position im Fahrzeug. Des Weiteren werden manuell für jedes Bild binäre Zustandswerte notiert, die beschreiben, ob der Fahrer eine Brille oder eine Sonnenbrille trägt und ob Verdeckungen vorhanden sind. Der Datensatz beinhaltet die Daten aus Realfahrten von 20 Probanden, insgesamt sind es mehr als eine Million Infrarot- und Tiefenbilder. Abbildung 1.1 zeigt exemplarisch zwei Datenpaare, bestehend aus Tiefen- und Infrarotdaten, zusammen mit den Referenzmessungen und den manuellen Notationen für Verdeckungen, Brillen und Sonnenbrillen. Um die Forschung im Bereich Kopfposenschätzung im Fahrzeug voranzutreiben, wird dieser Datensatz der Wissenschaft frei zur Verfügung gestellt.

Echtzeitfähiger Algorithmus zur Bestimmung der Kopfpose

Für die Anwendung der Kopfposenschätzung im Fahrzeuginnenraum ist ein effizienter Algorithmus entscheidend. Im Rahmen der Arbeit wird ein solcher entwickelt. Der Algorithmus bestimmt iterativ aus Tiefendaten die Kopfpose, beginnend mit einer frontalen Rotation und dem Mittelwert der Punktwolke in der Gesichtregion.

Dabei entspricht die frontale Rotation der Einheitsmatrix mit allen Rotationswinkeln auf null. Mithilfe von Entscheidungsbäumen und einer linearen Regression wird die Kopfposition und -orientierung in jedem Schritt aktualisiert. Die Entscheidungsbäume und die lineare Regression stellen eine effiziente Herangehensweise dar. Im Vergleich zu Verfahren, die den Stand der Forschung widerspiegeln, erreicht dieser Algorithmus bei gleichbleibender Genauigkeit eine dreimal schnellere Rechenzeit.

Evaluation und Entwicklung von Fusionsverfahren von Infrarot- und Tiefendaten

Auf dem neu erstellten Kopfposen-Datensatz im Fahrzeug erfolgt die Evaluation von Algorithmen mit dem aktuellen Stand der Forschung. Zusätzlich werden aktuelle tiefe neuronale Netze zur Schätzung der Kopforientierung angepasst. Um eine aussagekräftige Bewertung der Algorithmen zu erhalten, wird eine neue Metrik definiert. Diese neue Metrik berücksichtigt, dass während Autofahrten die frontale Orientierung am häufigsten vorkommt. Bei der Fehlerbestimmung wird ein Gleichgewicht hergestellt, indem die Winkel in Segmente eingeteilt werden und anschließend der Durchschnittsfehler der Segmente berechnet wird. Mit dieser Metrik werden Algorithmen, basierend auf Infrarotdaten, gegenüber Algorithmen, die auf Tiefendaten basieren, evaluiert. Die Validierung ergibt, dass Verfahren, die Tiefendaten verwenden, bessere Ergebnisse liefern. Um die Performance weiter zu erhöhen, werden Verfahren entwickelt, die Infrarot- und Tiefendaten fusionieren. Mithilfe dieser Fusionsverfahren kann die Kopforientierung weitaus genauer bestimmt werden. Bei der ausführlichen Evaluierung werden die unterschiedlichen Verdeckungen der Fahrer genauer betrachtet, welche unterteilt sind in Verdeckungen durch Brillen, Sonnenbrillen und weitere Verdeckungen (z.B. durch die Hand des Fahrers oder das Lenkrad).

1.3. Gliederung

Der folgende Abschnitt beschreibt die Gliederung der vorliegenden Arbeit.

Kapitel 2 gibt einen Überblick über die vorhandenen Arbeiten in Bezug auf die im vorherigen Abschnitt genannten Ziele. Vorhandene Kopfposendatensätze werden vorgestellt, die in der Literatur zur Evaluation von Algorithmen verwendet wurden. Es werden sowohl in Laborumgebung als auch im Fahrzeuginnenraum aufgenommene Datensätze genannt und in Relation zu dem im Rahmen dieser Arbeit entwickelten Datensatz gesetzt. Die Bestimmung der Winkel und Position des Kopfes basieren auf einem im Kopf verankerten Koordinatensystem. In der Literatur vorhandene Definitionen werden vorgestellt und die Vorteile des neu definierten Koordinatensystems aufgezeigt. Anschließend werden die Anforderungen an ein System zur Bestimmung der Kopfpose im Fahrzeug genannt. Basierend auf diesen Anforderungen werden vorhandene Algorithmen zur Bestimmung der Kopfpose aus 2D-Daten

und Tiefendaten sowie Fusionsverfahren erläutert. Für die jeweiligen Modalitäten wird auf Verfahren eingegangen, die für den Fahrzeuginnenraum konzipiert wurden. Dabei wird der Unterschied zu den im Rahmen der vorliegenden Arbeit entwickelten Verfahren herausgearbeitet sowie Algorithmen mit dem Stand der Forschung gewählt zur Evaluation auf dem neu entwickelten Datensatz im Fahrzeuginnenraum.

In **Kapitel 3** wird die Entwicklung des neu entstandenen Datensatzes zur Evaluation von Algorithmen für die Bestimmung der Kopfpose im Fahrzeuginnenraum vorgestellt. Dabei wird die Methodik für die Datenaufnahme mit dem verwendeten Messaufbau und der Spezifikation beschrieben sowie die Methodik zur Annotation der Daten. Zur Annotation der Daten wird ein neues Kopfkoordinatensystem definiert, die Referenzmessung der Kopfpose hergeleitet und die zusätzlichen manuellen Annotationen beschrieben. Anschließend wird die Auswahl des Sensors zur Datenaufnahme erläutert und die Streuung des Referenzsystems charakterisiert. Zum Schluss werden die Eigenschaften der Stichprobe, die den Datensatz bildet, genannt. Dieser Datensatz wird der wissenschaftlichen Gemeinschaft öffentlich zur Verfügung gestellt.

Inhalt von **Kapitel 4** ist das im Rahmen dieser Arbeit entwickelte regressionsbasierte Verfahren zur Bestimmung der Orientierung und Position des Kopfes. Zum Einen wird die Methodik beschrieben und zum Anderen der Algorithmus auf dem öffentlich verfügbaren BIWI-Datensatz [Fanelli et al. \(2013\)](#) angewendet. Bei der Evaluation werden die Parameter des Algorithmus optimiert, die Rechenzeit ausgewertet und das Ergebnis der 4-fachen Kreuzvalidierung mit vorhandenen Verfahren verglichen.

Kapitel 5 stellt die Architektur des neu entwickelten tiefen neuronalen Netzes für die Berechnung der Orientierung des Kopfes vor. Zusätzlich werden die angewendeten Fusionsverfahren zur Kombination von 2D- und Tiefendaten erläutert. Dabei wird eine *frühe*, eine *späte* und eine *stitch-basierte Fusion* vorgestellt.

Kapitel 6 evaluiert die Bestimmung der Orientierung und Position des Kopfes im Fahrzeuginnenraum, abhängig von unterschiedlichen Einflüssen. Die in Kapitel 4 und 5 vorgestellten Verfahren werden auf dem neu entwickelten Datensatz aus Kapitel 3 mit Algorithmen verglichen, die den aktuellen Stand der Forschung repräsentieren. Dabei wird analysiert, ob 2D- oder Tiefenbilder sowie die Fusion beider Modalitäten bessere Ergebnisse liefern. Zusätzlich wird die Robustheit der Verfahren auf Einflüsse wie Verdeckungen im Gesicht, Brillen und Sonnenbrillen untersucht.

In **Kapitel 7** werden die Ergebnisse und Beiträge dieser Arbeit zusammengefasst. Zudem wird ein Ausblick auf zukünftige Forschungsfelder gegeben.

1.4. Bereits veröffentlichte Beiträge

Die Hauptbeiträge dieser Arbeit wurden bereits in Konferenzen veröffentlicht. Der Datensatz in Kapitel 3 wurde bereits in der Veröffentlichung [Schwarz et al. \(2017\)](#) mit der Methodik für die Annotation der Daten vorgestellt. Das regressions-basierte Verfahren in Kapitel 4 zur Bestimmung der Kopfpose wurde bereits in [Schwarz et al. \(2016\)](#) präsentiert. Das tiefe neuronale Netz zur Bestimmung der Kopforientierung aus Infrarot- oder Tiefendaten sowie die Fusionsverfahren zur Kombination beider Modalitäten in Kapitel 5 wurden in [Schwarz et al. \(2017\)](#) veröffentlicht. Die in Kapitel 6 präsentierten Ergebnisse auf dem neuen Kopfposendatensatz im Fahrzeuginnenraum wurden teilweise bereits in der Veröffentlichung von [Schwarz et al. \(2017\)](#) diskutiert.

Eine komplette Liste der im Laufe dieser Arbeit entstandenen Veröffentlichungen, inklusive einer Veröffentlichung, deren Ergebnisse nicht in diese Arbeit eingeflossen sind, ist unter *Eigene Veröffentlichungen* nach dem Anhang zur Arbeit zu finden.

Kapitel 2

Stand der Forschung

Die Kopforientierung stellt eine Schlüsselkomponente zur Erfassung der Blickrichtung und Aufmerksamkeit von Personen dar. Für zahlreiche Systeme mit einer Mensch-Maschine Interaktion, wie zum Beispiel im Anwendungsbereich Robotik, sind diese Eigenschaften ausschlaggebend, um das Verhalten von Personen zu bestimmen. Im Fahrzeugkontext kann damit das Verhalten des Fahrers analysiert werden. Kamera-basierte Algorithmen zur Bestimmung der Kopfpose haben den Vorteil, dass die Person in ihrem Verhalten nicht eingeschränkt wird, im Vergleich zu Systemen bei denen zusätzliche am Kopf befestigte Marker notwendig sind. Dadurch ist der Forschungsbedarf zur visuellen Bestimmung der Orientierung und Position des Kopfes in den letzten Jahren enorm angestiegen. Im Rahmen dieser Arbeit wird die Bestimmung der Orientierung und Position des Kopfes aus Tiefenbildern im Fahrzeuginnenraum betrachtet. Dieses Kapitel gibt einen Überblick über verwandte Arbeiten und diskutiert im Hinblick darauf die Hauptbeiträge der vorliegenden Arbeit.

Ein Hauptbeitrag dieser Arbeit ist die Erstellung eines neuen Datensatzes im Fahrzeuginnenraum mit Referenzmessungen der Position und Orientierung des Kopfes. In Bezug auf diesen Datensatz werden in Kapitel 2.1 Datensätze mit Referenzmessungen der Position und Orientierung des Kopfes diskutiert. Hierfür werden sowohl im Labor als auch im Fahrzeuginnenraum aufgenommene Datensätze analysiert. Zusätzlich beinhaltet Kapitel 2.2 einen Überblick der verwendeten Kopfkoordinatensysteme.

Die beiden anderen Hauptbeiträge der Arbeit beschäftigen sich mit den Algorithmen zur Bestimmung der Orientierung und Position des Kopfes im Fahrzeuginnenraum. Einer der Beiträge ist die Konzeption eines echtzeitfähigen Algorithmus zur Bestimmung der Kopfpose aus Tiefendaten. Der andere Beitrag beschäftigt sich mit der Evaluation von 2D und tiefen-basierten Algorithmen im Fahrzeuginnenraum sowie der Entwicklung von Fusionsverfahren zur Bestimmung der Kopfpose. Hierfür beschreibt Kapitel 2.3 die Anforderungen an Algorithmen zur Anwendung im Fahrzeuginnenraum. Für die Konzeption des Algorithmus und die Evaluation von 2D und tiefen-basierten Algorithmen werden in Kapitel 2.4 in der Literatur vorhandene

relevante Algorithmen vorgestellt. Die Evaluation beinhaltet neben konventionellen Verfahren auch die Anpassung von neuronalen Netzen zur Bestimmung der Kopfpose und die Entwicklung von Fusionsverfahren zur Verwendung von 2d und Tiefendaten. Relevante Fusionsverfahren werden in Kapitel 2.4.3 diskutiert.

2.1. Kopfposen-Datensätze

In dieser Arbeit wird ein neuer Datensatz mit Referenzmessungen der Orientierung und Position des Kopfes vorgestellt. Hierbei handelt es sich um einen im Fahrzeuginnenraum aufgenommenen Datensatz. Im Folgenden werden vorhandene Datensätze diskutiert, die entweder im Labor oder im Fahrzeuginnenraum entstanden sind. Die wichtigsten Systemfaktoren bei der Aufnahme eines Datensatzes sind das Kamerasystem, das Referenzsystem, die Probanden und die Umgebung. Das Kamerasystem legt fest, welche Art von Bildern aufgenommen wird. Während der Aufnahme misst das Referenzsystem die Orientierung und Position des Kopfes als Referenzwerte. Die Wahl der Probanden sowie die Umgebung in der die Daten aufgenommen werden legen fest, was auf den Bildern zu sehen ist. Im Folgenden werden in Abschnitt 2.1.1 und 2.1.2 die Datensätze mit Blick auf diese Faktoren untersucht, siehe Tabelle 2.1 für einen Überblick.

2.1.1. Datensätze aus dem Labor

Die folgenden Datensätze wurden alle in einer Laborumgebung aufgenommen.

Der *BU - Boston University head tracking dataset* (La Cascia et al., 2000) ist ein Videodatensatz mit 72 Sequenzen, der fünf unterschiedliche Probanden beinhaltet. Ein "Flock of birds"-System bestimmt die Referenzdaten mit einem magnetischen Sensor. Das Besondere dieses Datensatzes ist, dass bei 27 der Sequenzen die Beleuchtung zeitlich variiert.

Im Gegensatz zum *BU - Boston University head tracking dataset* handelt es sich bei *Pointing '04* (Gourier et al., 2004b) um einen Blickrichtungsdatensatz. In der Forschung wurde dieser Datensatz häufig zur Evaluation von Algorithmen zur Bestimmung der Orientierung des Kopfes verwendet. Während der Aufnahme geben Marker an bestimmten Positionen vor, wohin die Probanden blicken. Anhand der Positionen der Marker wird die vertikale und horizontale Orientierung des Kopfes abgeleitet. Der Datensatz beinhaltet 30 Sequenzen von 15 unterschiedlichen Probanden. Die Teilnehmer tragen teilweise eine Brille oder einen Bart und haben unterschiedliche Hautfarben.

Ein weiterer Datensatz, der für eine andere Anwendung erstellt wurde, ist der *Bosphorus* (Savran et al., 2008). Bei diesem Datensatz handelt es sich um einen Datensatz zur Gesichtsanalyse, der mit einem 3D-Sensor aufgenommen ist. Die sieben

Year	2000	2004	2008	2011	2012	2012	2012	2014	2016	2017
Driving	-	-	-	-	-	-	-	-	-	-
Publicly available	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RGB/grayscale	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Depth	-	-	✓	✓	✓	✓	✓	✓	✓	✓
IR	-	-	-	-	-	-	-	-	-	-
Video	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Resolution	320×240	384×288	1600×1200	640×480	640×480	640×480	640×480	640×360	1280×720	512×424
Pixel aligned	N/A ^b	N/A ^b	✓	✓ ^a	✓ ^d	✓ ^d	N/A ^b	N/A ^b	N/A ^b	✓
N° subjects	5	15	105	20	10	14	14	N/A ^c	10	20
N° images	15k	3k	5k	15k	14k	200k	200k	90k	36k	1M
Female / male	0 / 5	1 / 14	45 / 60	6 / 14	6 / 4	N/A ^c	N/A ^c	N/A ^c	4 / 6	4 / 16
N° video sequences	72	30	N/A ^a	24	10	14	14	N/A ^c	120	21
Glasses labels	-	-	-	-	-	-	-	-	-	-
Sunglasses labels	-	-	-	-	-	-	-	-	-	-
Occlusion labels	-	-	-	-	-	-	-	-	-	-
Reference system	magnetic	marker	guided	Facelift Facelift AG	magnetic	magnetic	no-cap	inertial	magnetic	no-cap
Continuous labels	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Head orientation labels	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Head position labels	✓	-	✓	✓	✓	✓	-	-	✓	✓

^a not applicable since we do not have video

^b not applicable since we only have a single image modality

^c information not provided by the authors

^d transformation between depth and RGB modality available

Table 2.1.: Übersicht der vorhandenen Kopfposen-Datensätze, die entweder im Fahrzeug oder im Labor aufgenommen wurden. Die Tabelle stellt für jeden Datensatz die Aufnahme modalität, den Inhalt des Datensatz und die Eigenschaften der Referenzmessungen dar. Aus Schwarz et al. (2017) © 2017 IEEE.

unterschiedlichen Gierrotationen und sechs Nickrotationen der Daten beschreiben diskrete Referenzwerte. Zur Messung der Gierrotation positionieren die Probanden ihren Stuhl entsprechend markierter Linien. Für die Nickrotation betrachten die Probanden an der Wand befestigte Marker ohne die Augen zu bewegen. Der Datensatz beinhaltet insgesamt 4652 Daten von 105 unterschiedlichen Probanden. Neben den unterschiedlichen Kopforientierungen sind zahlreiche Emotionen vorhanden. Zusätzlich beinhaltet der Datensatz natürliche Verdeckungen durch die Hand, eine Brille oder die Haare. Diese Verdeckungen sind manuell markiert.

Der *BIWI kinect head pose dataset* (Fanelli et al., 2011a, 2013) ist ebenfalls ein 3D-Kopfposedatensatz, der im Gegensatz zum *Bosphorus* Datensatz mit der Kinect v1 aufgenommen ist. Dadurch enthält der Datensatz sowohl RGB Videosequenzen als auch Tiefendaten. Die Referenzwerte der Orientierung und Position des Kopfes bestimmt der 3D-basierte Algorithmus Faceshift (Faceshift AG). Die Sequenzen beinhalten insgesamt mehr als 15.000 Einzelbilder von 20 unterschiedlichen Probanden. Sowohl Brillen als auch Bärte sind teilweise bei den Teilnehmern vorhanden.

Bei dem *ICT 3dHP* (Baltrušaitis et al., 2012) Kopfposedatensatz ist das Kamerasystem ebenfalls eine Kinect v1. Die Referenzwerte misst ein magnetischer “Flock of birds”-Sensor. Hierbei ist der Sensor an einer Schirmmütze befestigt. Diese Mütze tragen die Teilnehmer mit dem Schirm nach hinten, damit das Gesicht nicht verdeckt ist. Die 10 unterschiedlichen Probanden bewegen ihren Kopf frei, um alle drei Achsen der Kopforientierung.

GI4E (Ariz et al., 2016) ist ein 2D Videodatensatz, aufgenommen mit einer Webcam. Als Referenzsystem wurde ebenfalls ein “Flock of birds”-System verwendet. Der Sensor ist am Kopf der Probanden befestigt. Zusätzlich zur Orientierung und Position des Kopfes werden die Gesichtslanmarken anfangs markiert und sind so für alle Daten vorhanden. Der Datensatz beinhaltet Aufnahmen von zehn unterschiedlichen Teilnehmern bestehend aus jeweils zwölf Videosequenzen, das heißt insgesamt 120 Videosequenzen. Die Probanden bewegen ihren Kopf in einem Teil der Sequenzen anhand eines von einer Tonspur vorgegebenen Musters und in dem anderen Teil bewegen die Teilnehmer ihren Kopf frei.

2.1.2. Datensätze aus dem Fahrzeuginnenraum

Die folgenden Datensätze wurden während realer Autofahrten im Fahrzeuginnenraum aufgenommen. Dadurch ergeben sich unterschiedliche Lichtverhältnisse.

Der Videodatensatz *Lisa-P* (Martin et al., 2012) ist mit einer Farbkamera im Fahrzeuginnenraum aufgenommen. Die Referenzwerte der Kopforientierung misst ein Motion Capture System. Dieses Motion Capture System verfolgt während der kompletten Fahrt Marker, die am Hinterkopf des Fahrers befestigt sind. Damit diese Marker sichtbar sind, ist die Nackenstütze des Fahrers ausgebaut. Insgesamt beinhaltet der Datensatz 14 Videosequenzen von unterschiedlichen Fahrern.

Ein weiterer im Fahrzeuginnenraum aufgenommener Kopfposendatensatz, ist der *CoHMEt* Datensatz (Tawari et al., 2014). Hierbei ist die Referenzorientierung des Kopfes mit einem Inertialsensor gemessen. Da es bei Inertialsensoren zu einem starken Drift kommt, ist die Kopforientierung alle zehn Sekunden manuell gelabelt. Die Aufnahmen stammen von unterschiedlichen Fahrern während realen Autofahrten. Insgesamt besteht der Datensatz aus mehr als 60 Minuten Videosequenzen mit einer Aufnahmezeit von 30 Bildern pro Sekunde.

Beitrag dieser Arbeit. In Kapitel 3 wird der im Rahmen dieser Arbeit neu erstellte Datensatz *DriveAhead* beschrieben. Die Daten sind mithilfe einer Kinect v2 (Microsoft, 2017) aufgenommen, dadurch beinhaltet der Datensatz als erster Datensatz zusätzlich zu Tiefendaten auch Infrarotdaten (IR). Für jedes Bild sind Referenzwerte vorhanden, die mithilfe eines Motion Capture Systems erzeugt wurden. Im Gegensatz zu den vorhandenen Datensätzen im Fahrzeuginnenraum wird zusätzlich zu Referenzwerten der Orientierung des Kopfes die Translation bestimmt. Der Datensatz beinhaltet 20 unterschiedliche Fahrer mit insgesamt mehr als einer Million Tiefen- und Infrarotdaten. Die Datenaufnahme findet während realen Autofahrten im Fahrzeuginnenraum statt. Dieser Datensatz ist der erste im Fahrzeuginnenraum aufgenommene veröffentlichte Datensatz, der neben 2D-Infrarotdaten Tiefendaten beinhaltet. Dadurch ermöglicht er eine Evaluation (siehe Kapitel 6) von Algorithmen abhängig der Eingangswerte, Grauwerte, Tiefenwerte oder einer Fusion beider Modalitäten.

2.2. Koordinatensystem des Kopfes

Das Koordinatensystem des Kopfes ist fest im Kopf verankert und bewegt sich bei Rotationen und Translationen des Kopfes mit. Bei allen Datensätzen wird die Rotation und Translation dieses Systems in Bezug zum Kamerakoordinatensystem gemessen und als Referenzmessung bereitgestellt. Die exakte Definition dieses Kopfkoordinatensystems ist bei der Messung der Orientierung und Position des Kopfes essentiell, um absolute Werte für die Position und Orientierung des Kopfes zu erhalten. Im folgenden Abschnitt wird genauer erläutert, wie das Koordinatensystem des Kopfes bei vorhandenen Arbeiten definiert ist und der Vorteil der im Rahmen dieser Arbeit vorgestellten Definition aufgezeigt.

Eine Möglichkeit ist die Orientierung der frontalen Pose des Kopfes zu verwenden (Gourier et al., 2004b; Savran et al., 2008) und anzunehmen, dass bei dieser Rotation das Koordinatensystem des Kopfes exakt mit dem des Kamerasystems übereinstimmt, d.h. die Rotation null beträgt. Allerdings muss berücksichtigt werden, dass diese Orientierung von derselben Person häufig nicht exakt reproduziert wird.

Eine weitere Möglichkeit ist die Definition des Koordinatensystems anhand der Positionen der Gesichtslandmarken (Baltrušaitis et al., 2016; Tawari et al., 2014). Hierbei wird ein allgemeines Modell bestehend aus den 3D-Punkten der Gesichtslandmarken auf die aktuell bestimmten Landmarken projiziert. Mit dem generellen Modell ist

festgelegt, wo der Ursprung des Koordinatensystems des Kopfes liegt und wie die Orientierungsachsen liegen. Die Berechnung der Projektion findet häufig mithilfe von POSIT (Dementhon und Davis, 1995) statt, indem die Abstände der individuellen 2D Gesichtslanmarken zu dem generellen Modell minimiert werden.

Des Weiteren kann das Koordinatensystem anhand eines vorhandenen 3D Modells des Kopfes definiert werden, wie zum Beispiel bei dem von der Firma [Faceshift AG](#) verwendeten Verfahren zur Verfolgung der Gesichtsbewegung. Dieser Algorithmus legt den Ursprung des Kopfkoordinatensystems in die Mitte des Kopfes. Dieses Verfahren wurde bei dem Datensatz von [Fanelli et al. \(2013\)](#) angewendet.

Beitrag dieser Arbeit. In der folgenden Arbeit wird eine Definition des Koordinatensystems des Kopfes basierend auf den individuellen Gesichtslanmarken vorgestellt, siehe Kapitel 3.4.1. Hierbei wird im Gegensatz zu ([Faceshift AG](#)) kein komplettes 3D-Modell des Kopfes benötigt. Außerdem ist kein generelles Modell der Gesichtslanmarken wie in ([Baltrušaitis et al., 2016](#)) notwendig, da die individuellen Gesichtslanmarken das Koordinatensystem definieren.

2.3. Anforderungen an die Bestimmung der Kopfpose im Fahrzeuginnenraum

Systeme, die im Fahrzeuginnenraum die Orientierung und Position des Kopfes bestimmen, müssen bestimmte Kriterien erfüllen. Diese werden im folgenden Abschnitt diskutiert.

[Tawari et al. \(2014\)](#) definiert die folgenden Kriterien, die ein System zur Anwendung im Fahrzeuginnenraum erfüllen muss:

- *Automatisiert:* Ein zeitaufwändiger Initialisierungsschritt ist nicht erforderlich.
- *Effizient:* Das System soll die Orientierung und Position des Kopfes in Echtzeit erfassen.
- *Umfangreiche Abdeckung:* Das System muss große Rotationen sowie schnelle Bewegungen abdecken.
- *Personeninvariant:* Das System muss unabhängig vom Fahrer funktionieren.
- *Verdeckungsvariant:* Das System muss mit Verdeckungen des Fahrers umgehen können.
- *Lichtinvariant:* Das System muss bei unterschiedlichen Lichtverhältnissen funktionieren.

Die genannten Kriterien sind ausschlaggebend zur Bestimmung der Kopfpose im Fahrzeuginnenraum, daher werden sie in der vorliegenden Arbeit berücksichtigt. Zum Einen können diese Kriterien von dem gewählten Algorithmus erfüllt werden und zum Anderen von dem Kamerasystem.

In dieser Arbeit werden folgende Kriterien von den Algorithmen berücksichtigt. Um ein *automatisiertes* System zu erhalten, werden *frame-to-frame* Verfahren konzipiert und untersucht. Bei diesen Verfahren wird für jedes Bild die Kopfpose unabhängig bestimmt. In Kapitel 2.4 wird ein *effizientes* Verfahren zur Bestimmung der Kopfpose aus Tiefendaten vorgestellt. Die restlichen Kriterien werden durch die Evaluation auf dem neu vorgestellten Datensatz untersucht. Der in Kapitel 2.1 vorgestellte Datensatz beinhaltet Daten von 20 unterschiedlichen Personen während realen Autofahrten. Indem das Training auf einem Teil der Personen durchgeführt wird und beim Testen die Daten der restlichen Personen betrachtet werden, ist eine *Personeninvarianz* der Algorithmen sichergestellt. Zusätzlich beinhaltet der Datensatz für jedes Einzelbild eine Notation, ob die Personen eine Brille oder eine Sonnenbrille tragen und ob zusätzliche Verdeckungen (z.B. durch das Lenkrad oder die Hand) im Gesicht vorhanden sind. Mit dieser Information werden in Kapitel 6 Algorithmen zur Bestimmung der Kopfpose auf *Verdeckungsinvarianz* untersucht.

Die letzte Anforderung wird berücksichtigt, indem ein Kamerasystem gewählt wird, welches robust gegenüber unterschiedlichen Lichtverhältnissen ist. Das Kamerasystem zur Aufnahme der Daten im Fahrzeug wird hierfür in Kapitel 3.3.3 auf *Lichtinvarianz* untersucht.

2.4. Algorithmen zur Bestimmung der Kopfpose

Gegenstand dieser Arbeit ist die Schätzung der Kopfpose im Fahrzeuginnenraum. Der folgende Abschnitt diskutiert bestehende 2D-basierte Verfahren 2.4.1, 3D-basierte Verfahren 2.4.2 sowie Fusionsverfahren 2.4.3. Für die unterschiedlichen Modalitäten werden Verfahren mit dem Stand der Forschung vorgestellt.

Im Rahmen dieser Arbeit werden *frame-to-frame* Ansätze untersucht, da diese für jedes Bild unabhängig die Orientierung und Position bestimmen. Anschließend können diese Verfahren mit einem *tracking* Verfahren ergänzt werden. Der folgende Abschnitt fokussiert sich deshalb auf *frame-to-frame* Ansätze und gibt für 3D-basierte Verfahren einen kurzen Überblick von vorhandenen *tracking* Ansätzen, da diese im Fahrzeuginnenraum bereits robuste Ergebnisse geliefert haben.

2.4.1. 2D-basierte Verfahren

In der Literatur sind zahlreiche Verfahren dokumentiert, die aus Farb- oder Grauwertbildern die Orientierung und Position des Kopfes bestimmen. Der folgende Ab-

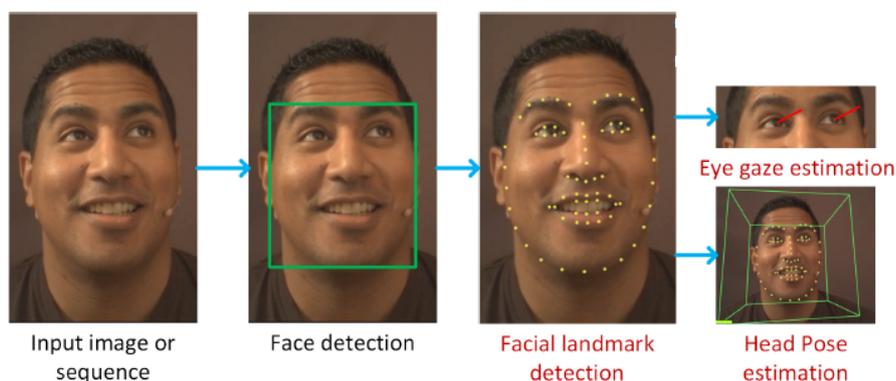


Abbildung 2.1.: Im ersten Schritt lokalisieren Conditional Local Neural Fields die Positionen der Gesichtslandmarken, anschließend wird die Kopfpose bestimmt. Aus Baltrušaitis et al. (2016) © 2016 IEEE.

schnitt gibt zuerst einen Überblick über allgemein vorhandene Verfahren und erläutert anschließend welche Algorithmen für die Anwendung im Fahrzeug präsentiert wurden.

Die Veröffentlichung von Murphy-Chutorian und Trivedi *et al.* (Murphy-Chutorian und Trivedi, 2009) fasst vorhandene Bild-basierte Verfahren zur Bestimmung der Orientierung des Kopfes zusammen. Im Folgenden werden 2D-basierte *frame-to-frame* Methoden diskutiert, welche die Kopfpose unabhängig vom vorherigen Bild für die Orientierung des Kopfes ausgeben. Murphy-Chutorian und Trivedi (2009) teilen diese Art von Verfahren in *Geometrische Methoden*, *Flexible Modelle* und *Nichtlineare Regressionsverfahren* ein.

Geometrische Methoden (Gee und Cipolla, 1994; Horprasert et al., 1996; Wang und Sung, 2007) bestimmen die Kopfpose anhand der relativen Positionen der Gesichtslandmarken, wie der Nasenspitze, den Augeneckpunkten oder den Mundwinkeln. Xiong und De la Torre (2013) präsentieren ein Verfahren zur akkuraten Bestimmung der Gesichtslandmarken und bestimmt daraus die Position und Orientierung des Kopfes. Bei diesem Algorithmus minimiert ein Abstiegsverfahren eine nichtlineare Funktion und liefert damit die optimale Position der Gesichtslandmarken. Anschließend wird anhand von Punkt-zu-Punkt Übereinstimmungen der geschätzten Gesichtslandmarken und eines rigiden 3D-Modells, die Orientierung und Position des Kopfes bestimmt. Ein ähnliches Verfahren präsentieren Baltrušaitis et al. (2016) (siehe Abbildung 2.1). Im Gegensatz zu der Methode von Xiong und De la Torre (2013) bestimmen Conditional Local Neural Fields die Positionen der Gesichtslandmarken. Da dieses Verfahren momentan den Stand der Technik bei der Landmarkendetektion widerspiegelt, wird es in der folgenden Arbeit als Baseline Methode verwendet und auf dem neu erstellten Datensatz in Kapitel 6 evaluiert.

Bei den *flexiblen Modellen* (Cootes et al., 2001; Krüger et al., 1997; Lanitis et al., 1997) handelt es sich um nicht rigide Modelle, die sich an die Gesichtsform anpassen.

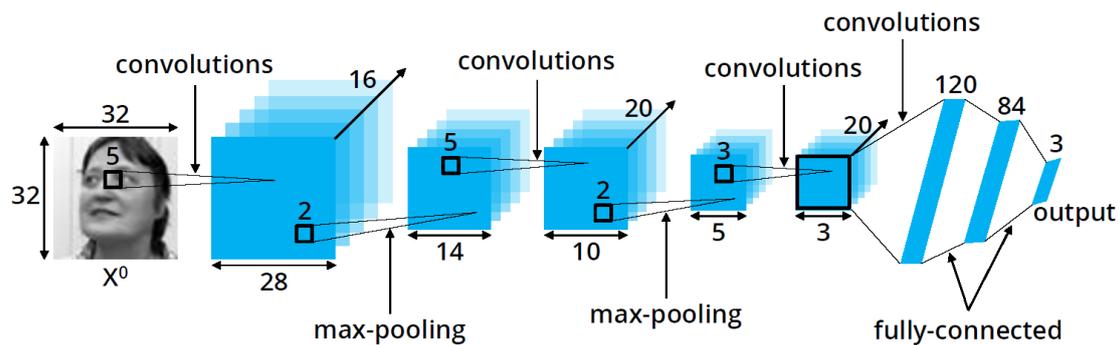


Abbildung 2.2.: Von Ahn et al. (2014) vorgestelltes tiefes Neuronales Netz zur Bestimmung der Orientierung und Position des Kopfes aus Grauwertbildern. Reprinted by permission from Springer Ahn et al. (2014) © 2014.

Nachdem die Positionen der Gesichtslanmarken bekannt sind werden statistische Verfahren, wie das *Active Appearance Model* (Cootes et al., 2001) oder das *multi-view Active Appearance Model* (Ramnath et al., 2008) angewendet. Für diese Verfahren sind ungleichmäßige Beleuchtungen eine große Herausforderung. Um robust gegenüber Beleuchtungsvarianzen zu werden, präsentieren Kingma und Ba (2014a) *asymmetric Appearance Models*.

Nichtlineare Regressionsmethoden bestimmen die Orientierung des Kopfes indem eine nichtlineare Projektion von dem Bild zu einer Orientierung gelernt wird. Diese Methode ist die in der Literatur am häufigsten verwendete Art um die Kopforientierung aus Bildern zu bestimmen. Vorhandene Verfahren verwenden hierfür *Support Vector Regressors* (SVR) (Li et al., 2004; Murphy-Chutorian et al., 2007) sowie SVRs an den Gesichtslanmarken *Feature-SVRs* (Ma et al., 2006; Moon und Miller, 2004). Zahlreiche Verfahren existieren, die auf neuronalen Netzwerken basieren. Beginnend mit *locally linear map* (LLM) neuronalen Netzen (Rae und Ritter, 1998), erzielten Verfahren mit *Multi-Layer Perceptron* (MLP) neuronalen Netzwerken (Seemann et al., 2004; Stiefelhagen et al., 2002; Voit et al., 2008) und *Convolutional Neural Networks* (CNN) (Osadchy et al., 2007) sehr gute Ergebnisse. Eine weitere Methode verwendet direkt an den Gesichtslanmarken ein *Feature-MLP* (Gourier et al., 2004b). Eine Möglichkeit, um die Rechenzeit zu minimieren, sind Entscheidungsbaumme (Dantone et al., 2012; Zhu und Ramanan, 2012). In den letzten Jahren erzielten tiefe neuronale Netze in zahlreichen Gebieten der Bildverarbeitung herausragende Ergebnisse. Für die Bestimmung der Orientierung des Kopfes aus Grauwerten erzielten Ahn et al. (2014) bereits mit wenigen Neuronen (siehe Abbildung 2.2) gute Ergebnisse auf dem BIWI Datensatz (Fanelli et al., 2013). Dieses Verfahren wird in Kapitel 6 zur Bestimmung der Kopfpose im Fahrzeug evaluiert.

Anwendungsgebiet Fahrzeuginnenraum

Für die Bestimmung der Kopfpose aus Grau- oder Farbwerten im Fahrzeuginnenraum, sind mehrere Verfahren in der Literatur vorhanden (Alioua et al., 2016; Martin

et al., 2012; Murphy-Chutorian und Trivedi, 2008; Murphy-Chutorian et al., 2007; Tawari et al., 2014). Murphy-Chutorian et al. (2007) bestimmen auf der Gesichtsregion Histogramme von *Localized Gradient Orientation* (LGO). Mithilfe dieser Histogramme wird ein *Support Vektor Regressor* (SVR) trainiert, um daraus die Gier- und Nick-Richtung des Kopfes zu bestimmen. Dieses Verfahren wurde in Murphy-Chutorian und Trivedi (2008) mit einer zeitlichen Verfolgung erweitert. Beide Algorithmen wurden auf Daten aus der Testumgebung von Lisa-P evaluiert, ähnlich zu dem veröffentlichten Lisa-P Datensatz Martin et al. (2012). In der Methode von Martin et al. (2012) werden Gesichtslanmarken mit dem modelbasierten *Constrained Local Models* (CLM) Ansatz detektiert. Anschließend bestimmt POSIT (Dementhon und Davis, 1995), durch die perspektivische Projektion eines 3D-Modells auf die 2D-Gesichtslanmarken im Bild, die Orientierung des Kopfes. In der Veröffentlichung von Tawari et al. (2014) wird die Bestimmung der Landmarkpositionen durch CLM mit dem Ansatz *Mixture of Pictorial Structures* (MPS) verglichen. Bei der Evaluation auf realen Autofahrten von mehreren Fahrern zeigt MPS mit POS bessere Ergebnisse in der Genauigkeit und eine geringere Fehlerrate. Der neu aufgenommene Datensatz CoHMEt beinhaltet Videos aus mehreren Kameras und ermöglicht dadurch den Vergleich von unterschiedlichen Positionen der Kamera.

Rezaei und Klette (2014) stellen ein Fahrerassistenzsystem basierend auf der Kopfpose vor, zur Warnung des Fahrers vor Unfällen. Mithilfe von *Asymmetric-Appearance-Modeling* (AAM) wird die 2D-Gesichtsstruktur auf ein 3D-Modell projiziert und damit die Kopfpose bestimmt. Das Verfahren wird auf zwei 60 Sekunden Videos von realen Fahrten qualitativ zur Anwendung für ein Warnsystem evaluiert. Ein weiteres Verfahren wurde von Alioua et al. (2016) vorgestellt, welches die Kopforientierung in fünf Orientierungen (frontal, nach oben, nach unten, nach links, nach rechts) klassifiziert. Das zweidimensionale Bild wird mit Hilfe von Deskriptoren (HOG, SF, Haar, SURF) beschrieben. Anschließend klassifizieren zwei gelernte SVMs die Kopforientierung entsprechend den Gier- und Nick-Winkeln. Das Verfahren wird qualitativ im Fahrzeug ausgewertet, für die quantitative Evaluation wird der *Pointing '04* Blickrichtungsdatensatz (Gourier et al., 2004b) verwendet.

Beitrag dieser Arbeit. In den letzten Jahren sind eine Vielzahl an Algorithmen zur Bestimmung der Kopfpose aus 2D-Daten entstanden. Sogar im Fahrzeuginnenraum wurden einige Algorithmen getestet, jedoch fehlt eine genaue Analyse, ob die Hinzunahme von Tiefenwerten die Ergebnisse verbessert. Diese Untersuchung wird durch den im Rahmen dieser Arbeit vorgestellte Datensatz ermöglicht. In Kapitel 6 wird eine Evaluation von vorhandenen 2D-basierten Verfahren zu der im Rahmen der Dissertation entwickelten echtzeitfähigen tiefen-basierten Methode (siehe Kapitel 2.4) und den Fusionsverfahren von tiefen neuronalen Netzen (siehe Kapitel 2.4.3) durchgeführt. Da das nichtlineare Verfahren von Ahn et al. (2014) den aktuellen Stand der Technik auf dem BIWI-Datensatz (Fanelli et al., 2013) definiert, wird es als Baseline Methode auf dem neu erstellten Datensatz im Fahrzeuginnenraum trainiert und in Kapitel 6 evaluiert. Allerdings bestimmt dieses Verfahren nicht die Position des Kopfes. Um zusätzlich zur Orientierung des Kopfes die Position des Kopfes bei der Verwendung unterschiedlicher Eingangsbilder zu evaluieren, wird das

geometrische Verfahren von [Baltrušaitis et al. \(2016\)](#) ausgewertet.

2.4.2. Tiefen-basierte Verfahren

In den letzten Jahren sind einige kostengünstige Tiefensensoren auf den Markt gekommen und dadurch ist das Interesse der Industrie an tiefen-basierten maschinellen Lernverfahren gestiegen. Die Algorithmen profitieren von der vorhandenen Punktwolke der Umgebung, die diese Sensoren liefern. Anhand der dreidimensionalen Form extrahieren maschinelle Lernverfahren ausschlaggebende Merkmale zur Bestimmung der Orientierung und Position des Kopfes. Im Folgenden wird unterschieden zwischen *frame-to-frame*-Algorithmen, die zu jedem Zeitpunkt die Kopfpose unabhängig von der vorherigen Kopfpose bestimmen, und *tracking*-Algorithmen, die eine zeitliche Verfolgung der Kopfpose anwenden. In dieser Arbeit werden *frame-to-frame* Algorithmen zur robusten Bestimmung der Orientierung und Position des Kopfes untersucht. Die Ergebnisse können anschließend als Initialisierungen für *tracking*-Verfahren verwendet werden.

Im Folgenden werden relevante *frame-to-frame*-Algorithmen vorgestellt. [Seemann et al. \(2004\)](#) vergleichen die Bestimmung der Kopforientierung aus Tiefendaten von Stereokameras, Grauwerten sowie der Kombination beider Modalitäten bei unterschiedlichen Lichtverhältnissen. Mithilfe einer Variation des Verfahrens von [Nickel und Stiefelhagen \(2003\)](#) wird die Gesichtsregion anhand eines Histogramms der Hautfarben bestimmt. Aus der Gesichtsregion bestimmt ein neuronales Netz die Orientierung des Kopfes, das Netz besteht aus drei komplett verbundenen Schichten. Das Ergebnis der Untersuchung verschiedener Modalitäten ergibt, dass Tiefendaten die Schätzung verbessern. Insbesondere im Fall von wechselnden Lichtverhältnissen ist die Bestimmung der Orientierung des Kopfes robuster bei Hinzunahme der Tiefendaten.

Im Vergleich zu der Methode von [Seemann et al. \(2004\)](#), bei der zuerst die Gesichtsregion extrahiert wird, bestimmt der Ansatz von [Breitenstein et al. \(2008\)](#) direkt aus der Punktwolke die Orientierung des Kopfes. Aus der Punktwolke ergeben die Werte von 3d shape signatures Hypothesen für die Position der Nase. Diese Hypothesen werden mit denen von Referenzmodellen verglichen, um die Orientierung des Kopfes zu erhalten. Da sowohl die Berechnung der shape signatures als auch der Vergleich mit vorhanden Modellen rechenintensiv sind, läuft das Verfahren auf einer GPU.

In dem Algorithmus, vorgestellt von [Fanelli et al. \(2011b\)](#), bestimmen Entscheidungsbäume auf eine effiziente Weise die Position und Orientierung des Kopfes. Innerhalb der Bäume wird der Unterschied zwischen Regionen aus Tiefendaten berechnet und mit einem Grenzwert verglichen. Auf diese Art wird ein Blatt des Baumes erreicht, in dem Orientierungswinkel und ein Vektor zur Position des Kopfes ausgegeben werden. Dieses Verfahren wurde von [Fanelli et al. \(2011a\)](#) erweitert. Es liefert zusätzlich zur Orientierung und Position des Kopfes eine Wahrscheinlichkeit,

ob die verwendete Region einen Kopf enthält. Damit kann das Verfahren auf beliebige Regionen des Bildes angewendet werden. Die effiziente Berechnung mithilfe von Entscheidungsbäumen ermöglicht es dem Verfahren, dass es auf einer CPU in Echtzeit laufen kann.

Die Anwendung von Entscheidungsbäumen für die Bestimmung der Orientierung und Position des Kopfes wurde in den Arbeiten von [Riegler et al. \(2014\)](#), [Redondo-Cabrera et al. \(2014\)](#) und [Schulter et al. \(2013\)](#) weiterentwickelt. Die Methode von [Riegler et al. \(2014\)](#) besteht aus einer Kombination von *Hough Forests* und *Convolutional Neural Networks*. Im Gegensatz zu Regressionsbäumen wird innerhalb der Knoten auf den Bildregionen ein *Convolutional Neural Network* trainiert, um die besten Eigenschaften dieser Bildregion zu extrahieren. Dieser Algorithmus verbessert die Ergebnisse der Kopfposenschätzung auf dem BIWI-Datensatz ([Fanelli et al., 2013](#)), allerdings verwendet die Methode eine GPU, um echtzeitfähig zu sein. [Schulter et al. \(2013\)](#) präsentieren *Alternated Regression Forests (ARFs)*, ein neues Optimierungsverfahren der Regressionsbäume. Hierbei wird eine globale Funktion für die Knoten minimiert, die im Gegensatz zu gewöhnlichen Entscheidungsbäumen die Funktionen innerhalb der Knoten für den kompletten Baum global minimiert. Die Methode von [Redondo-Cabrera et al. \(2014\)](#) bestimmt, zusätzlich zur Bestimmung der Position und Orientierung des Kopfes, die Gesichtsregion. Dadurch kann das Verfahren direkt auf ein beliebiges Bild angewendet werden, da die Gesichtsregion automatisch segmentiert wird.

Im Gegensatz zu den bislang beschriebenen Methoden, verwenden *tracking*-Verfahren die Schätzung der Kopfpose des aktuellen Bildes in zeitlich folgenden Bildern als Initialisierung. Diese Verfahren verwenden häufig die dreidimensionale Punktwolke und optimieren den Abstand eines Modells mit der geschätzten Rotation und Translation zu der aktuellen Punktwolke. Als Optimierungsfunktionen werden in der Literatur häufig die *Iterative-Closest-Points (ICP)*-Methode ([Bär et al., 2012](#); [Cai et al., 2010](#); [Martin et al., 2014](#); [Papazov et al., 2015](#)) und das *Particle-Swarm-Optimization (PSO)* - Verfahren ([Padeleris et al., 2012](#)) verwendet. Während für ICP eine gute Initialisierung benötigt wird, konvergiert PSO nur langsam zu dem finalen Optimum. Die Methode von [Meyer et al. \(2015\)](#) kombiniert die beiden Verfahren. Das Verfahren vermeidet so das Erreichen von lokalen Minimas aufgrund von schlechten Initialisierungen im Gegensatz zur ausschließlichen Verwendung von ICP und beschleunigt die Optimierung der PSO. Bei den *tracking*-Verfahren ist eine initiale Kopfposition notwendig. Im Gegensatz zu Verfahren bei welchen diese Anfangsorientierung näherungsweise frontal sein muss, kombiniert [Tulyakov et al. \(2014\)](#) einen *frame-to-frame*-Ansatz basierend auf Entscheidungsbäumen mit einem ICP *tracking*-Verfahren. Eine weitaus effizientere Methode zur zeitlichen Verfolgung der Orientierung und Position des Kopfes ist das Verfahren von [Tan et al. \(2017\)](#). Hier basiert die *tracking* Methode auf Entscheidungsbäumen. Das Verfahren besteht aus einer generischen offline-Lernphase und einer personenspezifischen online-Optimierung.

Anwendungsgebiet Fahrzeuginnenraum

Im Fahrzeuginnenraum wurden zahlreiche *tracking*-Verfahren (Bär et al., 2012; Breidt et al., 2016; Peláez C. et al., 2014) getestet. Hierbei wird das *Iterative-Closest-Point*-Verfahren zur Optimierung verwendet. Damit die Rechenzeit gering bleibt und trotzdem eine gute Schätzung erreicht wird, wechselt Bär et al. (2012) zwischen einem iterativen Punkt-zu-Punkt-Abgleich zu einem Punkt-zu-Ebene-Abgleich. Zusätzlich zur Kopfpose wird die Augenposition bestimmt, um die Blickrichtung des Fahrers zu erhalten. Die finale Blickrichtung wird auf einem selbst annotierten Datensatz im Fahrzeug evaluiert. Die Genauigkeit der geschätzten Kopfpose wird auf dem BIWI-Datensatz (Fanelli et al., 2013) evaluiert. Die Robustheit des Verfahrens gegenüber schnellen Kopfbewegungen wird zusätzlich auf realen Fahrten ausgewertet. Breidt et al. (2016) passt das 3D-Kopfmodell individuell für jede Person zur Bestimmung der Kopfpose an. Das individuelle Kopfmodell liefert bessere Ergebnisse. Der Algorithmus wird auf einer realen Autofahrt von einer Person evaluiert und mit einem 2D-Verfahren von Zhu und Ramanan (2012) verglichen. Zur Messung der Referenzkopfpose verfolgt ein Motion-Capture-System drei Marker, die auf einer Schirmmütze befestigt sind. Die Referenzmessung wird mit Hilfe eines 3D-Modells des Kopfes, welches auch die Marker enthält, auf das Kopfkoordinatensystem transformiert. Für das 2D-Verfahren wird die Referenzmessung mit der im ersten Bild vom Algorithmus bestimmten Kopfpose transformiert. Dadurch wird die Kopforientierung nur relativ gemessen und angenommen, dass die Schätzung im ersten Bild korrekt ist. Da es sich bei diesen Verfahren um ein *tracking*-Verfahren handelt, ist eine initiale Kopfpose notwendig, die hier als frontal angenommen wird.

Das Verfahren von Höffken et al. (2014) kombiniert einen *frame-to-frame*-Ansatz mit einem *tracking*-Algorithmus. Das System zur Bestimmung der Kopforientierung besteht aus drei wesentlichen Teilen: der Segmentierung des Kopfes, einer *frame-to-frame*-Bestimmung der Kopfpose und einem zusätzlichen *tracking* der Kopfpose. Als erstes wird anhand der 3D-Daten der Hintergrund entfernt und der Kopf des Fahrers segmentiert. *Synchronized Submanifold Embedding* klassifiziert basierend auf trainierten Kopfposen eine Schätzung für die Orientierung des Kopfes. Zusätzlich verbessert ein zeitliches tracking die Schätzung der Kopfpose. Die Arbeit von Tessema et al. (2016) untersucht den Schritt der *frame-to-frame*-Bestimmung der Kopfpose genauer, indem drei Verfahren hierfür verglichen werden: Support-Vector-Regression (SVR), Random-Regression Forest (RRF) und Extremely-Randomized-Trees (ERT). Bei der Evaluation auf realen Daten aus dem Fahrzeuginnenraum zeigen extremely randomized trees (ERT) die besten Ergebnisse. Im Vergleich zu den Systemen von Höffken et al. (2014) und Tessema et al. (2016) wird in der vorliegenden Arbeit zusätzlich zur Orientierung des Kopfes die Translation untersucht.

Mit dem enormen Erfolg von tiefen neuronalen Netzen in zahlreichen Gebieten der Bildverarbeitung präsentieren Borghi et al. (2017); Venturelli et al. (2017) Modelle für die Bestimmung der Kopfpose im Fahrzeuginnenraum aus Tiefendaten. Das Verfahren von Venturelli et al. (2017) ist ähnlich zu dem von Ahn et al. (2014) präsentierten Modell zur Bestimmung der Kopfpose aus Farbbildern. Borghi et al.

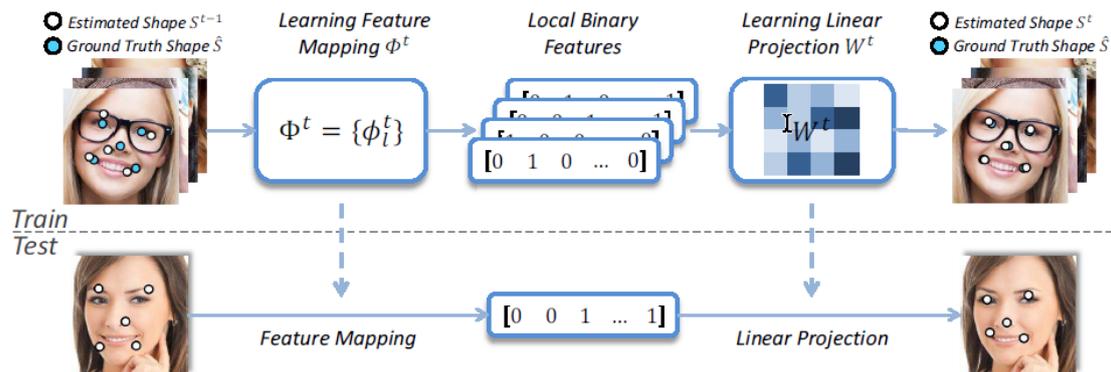


Abbildung 2.3.: Effiziente Methode zur Lokalisierung der Gesichtsländern in Farbbildern. Regressionsbäume erzeugen binäre Eigenschaftsvektoren, die anschließend multipliziert mit einer linearen Regression die Positionen der Landmarken schrittweise optimieren. Aus Ren et al. (2014) © 2014 IEEE.

(2017) präsentieren ein tiefes neuronales Netz POSEidon zur Bestimmung der Orientierung des Kopfes und des Oberkörpers aus Tiefendaten. Zur Evaluation des Verfahrens wird ein neuer Datensatz im Fahrsimulator vorgestellt.

Beitrag dieser Arbeit. Die Vielzahl der vorhandenen Veröffentlichungen zeigen das enorme Interesse der Wissenschaft im Bereich der akkuraten Schätzung der Orientierung und Position des Kopfes aus Tiefendaten. Die folgende Arbeit konzentriert sich auf die effiziente und akkurate Bestimmung der Kopfpose für jedes Bild unabhängig vom vorherigen Bild im Fahrzeugkontext. Anschließend können diese Verfahren mit einem *tracking*-Ansatz optimiert werden. Im Vergleich zu vorhandenen *frame-to-frame*-Ansätzen zur Kopfposenschätzung ist das in Kapitel 4 vorgestellte Verfahren durch die Verwendung einer Kombination von binären Werten, die aus Regressionsbäumen generiert werden, und einer linearen Matrixmultiplikation weitaus effizienter als die bisher aus der Literatur bekannten Ansätze. Dieses Methode basiert auf dem von Ren et al. (2014) vorgestellten Verfahren zur Bestimmung der Gesichtsländern in Farbbildern (siehe Abbildung 2.3). Im Gegensatz dazu bestimmt die in Kapitel 4 vorgestellte Methode direkt aus Tiefenbildern die Orientierung und Position des Kopfes. Die Evaluation dieses Verfahrens auf dem BIWI Datensatz (Fanelli et al., 2013) zeigt ähnliche Ergebnisse wie vorhandenen tiefen-basierten Verfahren (Fanelli et al., 2013; Papazov et al., 2015; Riegler et al., 2014), deshalb wird dieses Verfahren als Baseline Methode auf dem neu erstellten Datensatz bei der Evaluation in Kapitel 6 verwendet. Zusätzlich werden tiefe neuronale Netze mit dem Stand der Forschung (Simonyan und Zisserman, 2015) auf die Anwendung zur Bestimmung der Kopfpose angepasst.

2.4.3. Fusionsverfahren von 2D- und Tiefendaten

Im folgenden Kapitel werden unterschiedliche Fusionsverfahren diskutiert, die zur Bestimmung der Kopforientierung und Position in der Literatur bekannt sind.

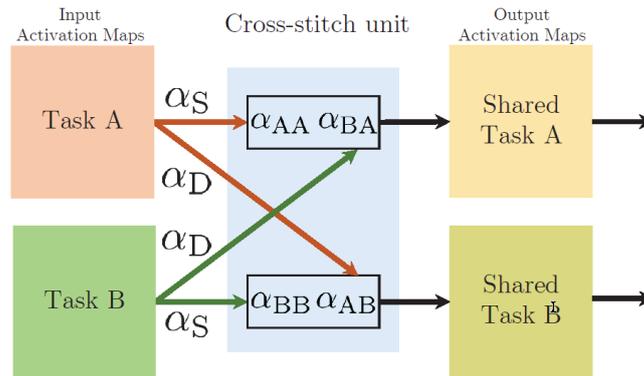


Abbildung 2.4.: Vorgestelltes *stitch* Verfahren, welches mit Hilfe von trainierten Gewichten zwei tiefe Neuronale Netze kombiniert. Damit werden zwei unterschiedliche Aufgaben gelöst. Aus Misra et al. (2016) © 2016 IEEE.

Seemann et al. (2004) vergleichen neuronale Netze auf Tiefen- und Grauwerten sowie die Fusion von beiden Modalitäten. Die Fusion liefert die besten Ergebnisse. Bei der Fusion bestimmt ein neuronales Netz direkt aus einer zweidimensionalen Matrix mit jeweils einer Schicht von Tiefen- und Grauwerten die Kopfpose. Im Folgenden wird diese Art als *frühe* Fusion bezeichnet.

Yun et al. (2014) fusionieren RGB und Tiefenwerte mit Hilfe von AdaBoost. Aus den Farb- und Tiefenwerten werden zuerst getrennt Hypothesen berechnet. Diese werden anschließend in einem Fusionschritt zusammengeführt und es berechnet sich daraus eine finale Schätzung der Kopfpose. Im Gegensatz dazu kombinieren Kaymak und Patras (2012) Farb- und Tiefenwerte innerhalb von Regressionsbäumen. Regressionsbäume, die zusätzlich zu Tiefenwerten, das Farbbild berücksichtigen zeigen geringfügig bessere Ergebnisse.

Redondo-Cabrera et al. (2014) vergleichen in ihrem Ansatz die Verwendung von Hough Networks zur Kopfposesbestimmung aus Farbbildern, Tiefendaten und Fusionsverfahren. Hierbei werden *frühe* und *späte* Fusionsverfahren auf dem BIWI Datensatz (Fanelli et al., 2013) evaluiert. Bei der *frühen* Fusion werden RGB und Tiefeninformation im selben Hough Netzwerk verarbeitet während bei der *späten* Fusion getrennte Hough Netzwerke verwendet werden. Die *späte* Fusion erreicht bessere Schätzungen für die Gier- und Rollbewegung, während für die Nickbewegung der Ansatz von Fanelli et al. (Fanelli et al., 2011a) basierend auf Tiefendaten die besten Ergebnisse liefert.

Beitrag dieser Arbeit. In dieser Arbeit werden in Kapitel 5, aufbauend auf den in der Literatur vorhandenen Fusionsverfahren, eine *frühe* und *späte* Kombination vorgestellt. Im Vergleich zu den oben genannten Verfahren wird hier ein tiefes neuronales Netz von Simonyan und Zisserman (2015) verwendet, da dieses in der Literatur auf zahlreichen Gebieten Ergebnisse mit dem Stand der Technik lieferte. Zusätzlich wird ein Verfahren präsentiert, das den *Cross-Stitch Net*-Ansatz von Misra et al. (2016) zur Fusion von zwei Eingangsmodalitäten erweitert. Das *Cross-Stitch*

Net optimiert die Zusammenführung von zwei neuronalen Netzen zur gemeinsamen Lösung von zwei Aufgaben (siehe Abbildung 2.4). Dieses Verfahren wird in Kapitel 5 zur Bestimmung der Kopfpose durch die Kombination zweier neuronaler Netze angepasst. Während das Eine auf Tiefenwerten agiert, verwendet das Andere Infrarotwerte.

2.5. Quaternionen zur Beschreibung von Rotationen

In der folgenden Arbeit werden Rotationen anhand von Quaternionen-Vektoren beschrieben. Dieser Abschnitt gibt einen Überblick über die Definition von Quaternionen und die Vorteile dieser Darstellung für Rotationen.

Ein Quaternion \mathbf{q} ist wie folgt definiert, siehe Dam et al. (1998):

$$\mathbf{q} = x_0 + x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k} \quad (2.1)$$

wobei x_0, x_1, x_2 und $x_3 \in \mathbb{R}$ reellen Zahlen entsprechen und $\mathbf{i}, \mathbf{j}, \mathbf{k}$ imaginären Zahlen entsprechen mit der Eigenschaft $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$. Die hintereinander Ausführung von zwei Drehungen entspricht der Multiplikation der zugehörigen Quaternionen. Bei der Multiplikation zweier Quaternionen werden die imaginären Zahlen mithilfe der *Hamilton-Regeln* (Hamilton, 1844) verknüpft. Quaternionen erfüllen in der Algebra die Eigenschaften eines Schiefkörpers. Die Eigenschaften eines Ringes sind nicht erfüllt, da wie auch bei Rotationsmatrizen die Multiplikation von Quaternionen nicht kommutativ ist. Es gelten allerdings die Assoziativ- und Distributivgesetze und es existiert ein inverses Element für jedes Quaternion.

In Anwendungen zeigt sich die Darstellung von Rotationen anhand von Quaternionen als vorteilhaft gegenüber der Verwendung von Matrizen und Eulerwinkeln, wie zum Beispiel in Collet et al. (2011) und Beyer et al. (2015). Der Vorteil von Quaternionen gegenüber Rotationsmatrizen ist die Möglichkeit der Interpolation zwischen zwei Quaternionen und die kompakte Darstellung von vier Werten, im Gegensatz zu neun Werten. Im Vergleich zu Eulerwinkeln bieten Quaternionen eine eindeutige Darstellung von Rotationen. Deshalb wird in der folgenden Arbeit die Orientierung des Kopfes mithilfe von Quaternionen-Vektoren beschrieben.

Kapitel 3

Kopfposen-Datensatz im Fahrzeuginnenraum

Dieses Kapitel beinhaltet die Beschreibung des im Rahmen der vorliegenden Arbeit aufgenommenen Datensatzes im Fahrzeuginnenraum. Die verwendete Methodik und der Datensatz wurden in [Schwarz et al. \(2017\)](#) bereits veröffentlicht. Die Experimente und Ergebnisse zur Auswahl des Tiefensensors (siehe Abschnitt [3.3](#)) sind während der Bachelorarbeit von [Wolman \(2014\)](#) entstanden, die im Rahmen dieser Arbeit betreut wurde.

Die Bestimmung der Kopfpose im Fahrzeuginnenraum unter realen Bedingungen ist eine herausfordernde Aufgabe. Die Schwierigkeit liegt zum Einen an der akkuraten Bestimmung der Kopfpose und zum Anderen an den Gegebenheiten im Fahrzeuginnenraum. Im Fahrzeugkontext muss eine Robustheit gegenüber verschiedenen Lichteinflüssen, für Fahrer mit Brillen oder Sonnenbrillen und bei weiteren Verdeckungen im Gesichtsbereich sichergestellt werden. Zur Wahl eines geeigneten Algorithmus ist ein Datensatz mit realistischen Gegebenheiten essentiell. Während existierende Datensätze entweder keine Tiefenbilder beinhalten ([Martin et al., 2012](#); [Tawari et al., 2014](#)) oder im Fahrsimulator aufgenommen wurden ([Borghi et al., 2017](#)), wird in diesem Kapitel ein neuer Datensatz mit Infrarot- und Tiefendaten aus dem Fahrzeuginnenraum präsentiert. Die Datensatzaufnahme wurde während realer Autofahrten durchgeführt.

Das folgende Kapitel beinhaltet eine ausführliche Beschreibung des neu entstandenen Datensatzes zur Evaluation der Kopfpose im Fahrzeuginnenraum. Im ersten Abschnitt werden das Ziel der Datenaufnahme und die damit verbundenen Kriterien an den Datensatz definiert [3.1](#). Anschließend wird in Kapitel [3.2](#) die verwendete Methodik für die Datenaufnahme und in Kapitel [3.3](#) die Auswahl des dabei verwendeten Tiefen-Sensors erläutert. Die Methodik der Notation der Daten wird in Kapitel [3.4](#) beschrieben. Kapitel [3.5](#) charakterisiert das Referenzsystem zur Messung der Kopfpose. In Kapitel [3.6](#) wird die aufgenommene Stichprobe beschrieben. Die erzeugte

Stichprobe wird als Datensatz mit dem Namen *DriveAhead* der Öffentlichkeit zur Verfügung gestellt, siehe Kapitel 3.7

3.1. Ziel

Das Ziel der Entwicklung des neuen Datensatzes ist die Analyse von Verfahren zur Bestimmung der Kopfpose. Im Rahmen der vorliegenden Arbeit wird untersucht, von welchen Einflüssen und auf welche Weise die Genauigkeit der Verfahren im Fahrzeuginnenraum beeinflusst werden. Für die Verwendung der Algorithmen zur Bestimmung der Kopfpose im Fahrzeuginnenraum ist ein Datensatz notwendig, der folgende Kriterien der Verfahren adressiert:

- Genauigkeit der Bestimmung der Kopfpose
- Anwendbarkeit im Fahrzeuginnenraum während Autofahrten
- Personeninvarianz
- Robustheit gegenüber Brillen und Sonnenbrillen
- Robustheit gegenüber Verdeckungen im Gesichtsbereich
- Vergleich von 2D-Grauwert- und Tiefenbildern als Eingangswerte

Aus diesen Faktoren wird im folgenden Abschnitt die Teststrategie zur Aufnahme des Datensatzes hergeleitet.

Teststrategie. Für die Datensatzaufnahme wird die folgende Teststrategie verwendet, welche die Anforderungen an die Datenaufnahme bildet.

- Messung der Referenzorientierung und -position des Kopfes
- Aufnahme während realer Fahrten im Fahrzeuginnenraum
- Aufnahme von mehreren Personen
- Aufnahme von Personen mit Sonnenbrillen und Brillen
- Realistische Verdeckungen während echter Autofahrten
- Aufnahme von Grauwert- und Tiefenbildern

Die einzelnen Punkte der Teststrategie leiten sich aus den Zielen her. Sie dienen als Grundlage zur Entwicklung der Methodik.

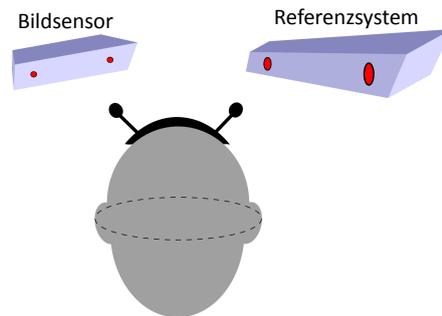


Abbildung 3.1.: Überblick des Systems zur Aufnahme des Kopfposen-Datensatzes, bestehend aus einem Bildsensor und einem Referenzsystem.

3.2. Methodik der Datenaufnahme

Im folgenden Kapitel wird die verwendete Methodik zur Aufnahme des Datensatzes beschrieben. Hierbei werden die aufgestellten Ziele der Datensammlung und die daraus abgeleitete Teststrategie aus dem vorherigen Abschnitt als Basis verwendet.

Beginnend mit dem Messaufbau in Abschnitt 3.2.1 wird der allgemeine Aufbau des Systems zur Datenaufnahme im Fahrzeuginnenraum beschrieben. Anschließend wird der Ablauf der Datenaufnahme in Abschnitt 3.2.2 spezifiziert.

3.2.1. Messaufbau

Im Folgenden wird der Aufbau des Systems zur Aufnahme des Kopfposendatensatzes im Fahrzeug beschrieben. Das System besteht aus einem Bildsensor und einem Referenzsystem zur Messung der Kopforientierung und -position, wie in Abbildung 3.1 dargestellt.

Die Auswahl des 3D-Kamerasystems wird in Kapitel 3.3 beschrieben und das System auf Umgebungslicht untersucht. Als Bildsensor wird die Microsoft Kinect V2, auch bekannt als Microsoft Kinect One, verwendet. Dieser Sensor bietet sowohl Tiefen-, Infrarot- und RGB-Bilder. Bei der Aufnahme werden aufgrund der limitierten Aufnahmezeit ausschließlich Tiefen- und Infrarotdaten aufgezeichnet. Durch die Aufnahme von zusätzlichen RGB-Bildern käme es zu erhöhten *Frame-Drops*. Außerdem sind die beiden Datenstränge von Infrarot- und Tiefenbildern für die Verwendung im Fahrzeug geeignet, da sie ausreichend robust gegenüber Fremdlicht sind.

Für die Referenzmessung der Kopfposition und -orientierung wird das SMART-TRACK Motion Capture-System von der Firma [Advanced Realtime Tracking \(ART\)](#) angepasst. Dieses System ermöglicht eine exakte Positions- und Orientierungsmessung von einem Objekt, das mit speziellen Kugeln versehen ist. Die Anpassung des

Systems zur Messung der Kopfposition und -orientierung wird in Kapitel 3.4.2 genauer erläutert. Das entwickelte Referenzsystem wird in Kapitel 3.5 charakterisiert, indem die Streuung des Systems untersucht wird.

Während der Versuchsdurchführung werden die beiden Systeme mit dem Network Time Protocol (NTP) zeitlich synchronisiert. Die Daten der beiden Systeme wurden zusammen mit den synchronisierten Zeitstempeln mithilfe der Software *ADTF* (Automotive Data and Time-Triggered Framework) aufgezeichnet. ADTF ist ein im Automotive-Bereich häufig verwendetes Framework zur Aufnahme von Daten aus mehreren Quellen.

3.2.2. Spezifikation der Datenaufnahme

Nachfolgend wird die Auswahl der Fahrstrecke und der Ablauf der Datenaufnahme spezifiziert. Die Fahrstrecke ist so gewählt, dass sie unterschiedliche Fahrscenarien ermöglicht. Der Versuchsablauf beinhaltet die einzelnen Schritte, welche zur Aufnahme der Referenz- und Bilddaten notwendig sind. Die Schritte beinhalten eine Kalibrierungsphase im Stillstand vor und nach der Datenaufnahme sowie die Versuchsfahrt zur Datenaufnahme.

Fahrstrecke

Der in dieser Arbeit beschriebene Datensatz wurde unter realen Bedingungen auf öffentlichen Straßen im Fahrzeuginnenraum aufgenommen. Die gewählte Strecke ist in Abbildung 3.2 dargestellt und befindet sich in der Nähe von Heilbronn in Deutschland. Die Route beinhaltet sowohl einen Abschnitt auf der Autobahn als auch ländliche Ortsdurchfahrten. Die Landstraßen befinden sich zum Teil in einem Wald mit zahlreichen Kurven, hierbei kommt es zu vielen Kopfdrehungen. Zusätzlich wurde während den meisten Versuchsfahrten ein Wechsel zu einer Sonnenbrille durchgeführt, hierfür haben die Fahrer angehalten und ein Parkmanöver durchgeführt. Insgesamt beträgt die Strecke 22 Kilometer mit einer Fahrzeit von ca. 30 Minuten.

Koordinatensysteme

Die folgenden Koordinatensysteme sind notwendig, um die Referenz der Kopfpose herzuleiten. Als Referenz wird die Rotation und Translation des Kopfkoordinatensystems h im Kamerakoordinatensystem k für jedes Bild angegeben. Abbildung 3.3 visualisiert die folgenden Koordinatensysteme:

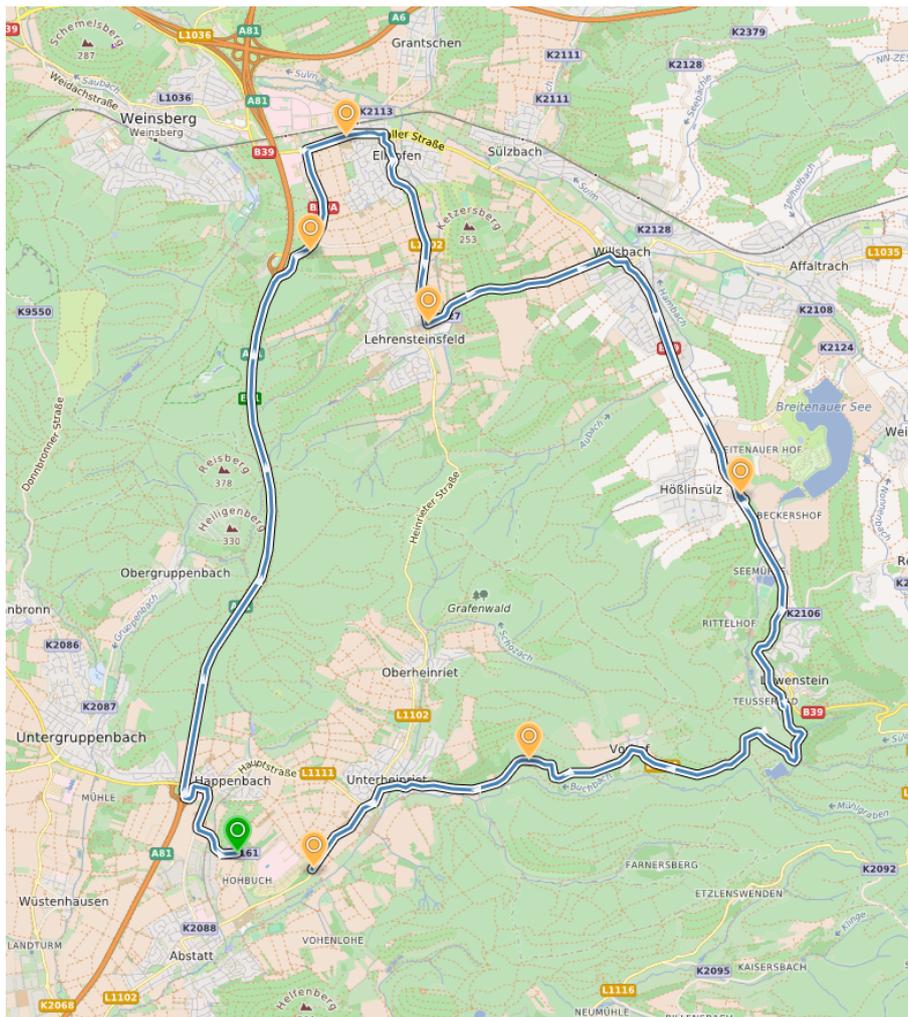


Abbildung 3.2.: Fahrstrecke der Versuchsfahrten. Es sind Autobahnabschnitte sowie Landstraßen und Ortsdurchfahrten enthalten. © OpenStreetMap-Mitwirkende (2017)

- Das **Kamerakoordinatensystem** k befindet sich an dem Kamerasystem und bleibt während der gesamten Aufnahme konstant. Zwischen den Aufnahmen kann die Position der Kamera theoretisch verändert werden. In der folgenden Datenaufnahme bleibt das System fixiert wie im Fall eines zukünftigen Fahrzeugs, bei dem die Kamera einmalig während der Produktion fest montiert ist.
- Das **Weltkoordinatensystem** w beschreibt ein im Fahrzeug definiertes Koordinatensystem, welches während einer Aufnahme konstant bleibt. Es wird zu Beginn der Messung in Relation zum Kamerakoordinatensystem k anhand eines Kalibrierungsschrittes definiert. Im Fall dieser Datenaufnahme kann das Weltkoordinatensystem w zwischen zwei Sequenzen an unterschiedlichen Positionen definiert werden, solange die Kalibrierungsmatrix neu erzeugt wird. Allerdings wird im Fahrzeugkontext häufig dieses Koordinatensystem statt dem Kamerakoordinatensystem als finales System verwendet, da dieses in Relation zum Fahrzeug steht. Deshalb wird für diesen Datensatz das Weltkoordinaten-

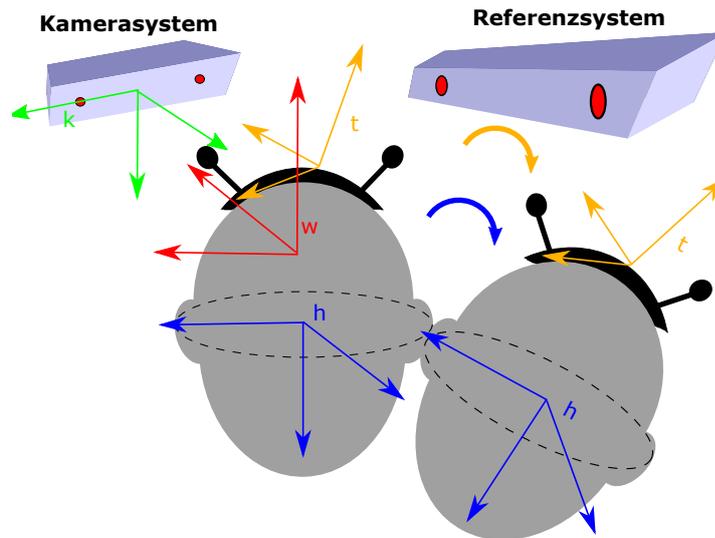


Abbildung 3.3.: Übersicht der Koordinatensysteme zur Aufnahme des Datensatzes: Das Weltkoordinatensystem w , das Target Koordinatensystem t , das Kopfkoordinatensystem h und das Kamerakoordinatensystem k .

system für jede Aufnahme an derselben Stelle definiert.

- Das **Targetkoordinatensystem** t ist auf dem Target des Motion Capture-Systems fixiert. Von dem Motion Capture System wird zu jedem Zeitpunkt die Transformation von diesem Koordinatensystem im Weltkoordinatensystem w gemessen. Das Target ist ein rigides Objekt, welches an die Kopfform angepasst ist. Auf diesem Objekt sind Marker montiert, die bereits bei der Produktion des Targets in Relation zu dem Targetkoordinatensystem t vermessen sind.
- Das **Kopfkoordinatensystem** h ist ein im Kopf fixiertes Koordinatensystem, welches sich bei Bewegungen des Kopfes mit bewegt. Es wird anhand der in Kapitel 3.4.2 beschriebenen Definition für jedes Individuum neu berechnet. Dabei wird die Orientierung und Position aus den 3D-Landmarken im Gesicht hergeleitet, die während der Datenaufnahme für jeden Probanden vermessen werden. Dadurch ist die Lage des Koordinatensystems in Relation zu diesen Landmarken eindeutig bestimmt. Diese Eindeutigkeit ist wichtig, um für die unterschiedlichen Probanden zueinander vergleichbare Referenzmessungen der Orientierung und Position des Kopfes zu erhalten.

Versuchsablauf

In Abbildung 3.4 (a)-(d) sind die einzelnen Schritte des Versuchsablaufs dargestellt. Diese Schritte sind erforderlich, um die Lage und Position der im vorherigen Abschnitt beschriebenen Koordinatensysteme zu definieren. Im Folgenden wird für jeden Schritt beschrieben, welches der Koordinatensysteme damit definiert wird.



(a) Kalibrierung



(b) Markierung der Gesichtslandmarken zur Definition des Kopfkoordinatensystems



(c) Versuchsfahrt



(d) Markierung der Gesichtslandmarken zur Validierung

Abbildung 3.4.: Übersicht des Versuchsablaufs. Für jeden Probanden werden die in Abbildung (a)-(d) dargestellten Schritte durchgeführt.

- (a) **Kalibrierung:** Das Kamerasystem und das Referenzsystem werden auf dasselbe gemeinsame Weltkoordinatensystem \mathbf{T}^w kalibriert. Hierfür wird ein Multifunktionsbrett für jeden Versuch an derselben exakt definierten Stelle im Fahrzeug befestigt. Das Multifunktionsbrett ist mit einem visuellen Schachbrettmuster und exakt montierten Markern ausgestattet. Das Schachbrettmuster ist zur Kalibrierung des Kamerasystems notwendig, während das Referenzsystem die Marker zur Kalibrierung benötigt.
- (b) **Erste Markierung der Gesichtslandmarken zur Definition des Kopfkoordinatensystems:** Der Proband markiert mit einem Zeiger, dessen 3D-Position an der Spitze von dem Referenzsystem erfasst wird, acht Landmarken im Gesicht. Bei den Landmarken handelt es sich um die in Abbildung 3.12 gezeigten Gesichtslandmarken. Diese definieren das Kopfkoordinatensystem h . Die Definition wird in Kapitel 3.4.2 detailliert erläutert.
- (c) **Versuchsfahrt:** Während der Versuchsfahrt wird zu jedem Zeitpunkt i die Position und Orientierung des Kopftargets $\mathbf{T}_i^{w \rightarrow t}$ mit dem Referenzsystem im

Weltkoordinatensystem gemessen. Die Transformation in das Kopfkoordinatensystem wird in Kapitel 3.4.2 erläutert. Die Versuchsfahrt beinhaltet sowohl eine Fahrt auf der Autobahn als auch Ortsdurchfahrten. Die Route wird im nächsten Abschnitt genauer beschrieben.

- (d) **Zweite Markierung der Gesichtslanmarken zur Validierung:** Nach der Versuchsfahrt werden die Gesichtslanmarken nochmals vermessen, um die Fahrt zu validieren, sowie die Messunsicherheit des Kopfkoordinatensystems zu bestimmen. Es ist notwendig die Fahrt zu validieren, da das Kopftarget während der Fahrt verrutschen kann, falls der Fahrer damit das Fahrzeugdach oder die Kopfstütze berührt.

3.3. Auswahl eines Tiefen-Sensors

Im Vorfeld der Datenaufnahme wird anhand definierter Kriterien ein geeigneter Sensor für die Fahrzeugmessungen ausgewählt und charakterisiert. Diese Vorarbeiten fanden im Rahmen einer Bachelorarbeit (Wolman, 2014), die während dieser Doktorarbeit betreut wurde, statt.

Zuerst wird ein geeigneter Sensor zur Aufnahme der Gesichtsregion im Fahrzeuginnenraum gewählt. Hierbei werden die Anforderungen (siehe Abschnitt 3.3.1) an einen Tiefensensor für dieses Anwendungsgebiet aufgestellt und experimentell für unterschiedliche Sensoren untersucht (siehe Abschnitt 3.3.2). Anschließend wird in Abschnitt 3.3.3 die Robustheit des Sensors gegenüber Fremdlicht untersucht.

3.3.1. Anforderungen

Für den Anwendungsfall im Fahrzeuginnenraum werden die Anforderungen an den Tiefensensor definiert, die anschließend untersucht werden:

- *Genauigkeit und Streuung:* Die Genauigkeit und Präzision des Sensors legen den Fehler der Tiefenbestimmung fest. Während die Genauigkeit den systematischen Fehler misst, beinhaltet die Präzision die Streuung der Werte. Die Genauigkeit spielt bei der Auswahl der Sensoren eine untergeordnete Rolle, da dieser systematische Fehler mit einer entsprechenden Kalibrierung reduziert werden kann (Foix et al., 2011).
- *Laterale Auflösung:* Mit der lateralen Auflösung wird angegeben, wie exakt der Sensor in die laterale Richtung auflösen kann. Damit wird festgelegt, wie fein Strukturen sein dürfen, um noch Unterschiede in der Tiefe zu erkennen. Das bedeutet, ein niedriger Wert entspricht einer besseren Auflösung. Für die Auflösung des Gesichts wird eine laterale Auflösung gefordert, die es ermöglicht, Strukturen wie die Augen, den Mund und die Nase zu unterscheiden.

Sensor	Kinect One (Microsoft, 2017)	pmd nano (Photonics, 2014)	DS325 (Softkinetic, 2014)
Auflösung	512×424	160 × 120	320 × 240
Field of View (FOV)	70° × 60°	90° × 68°	74° × 58°
Reichweite [in mm]	500-4000	0-2000	150-1000
Präzision	keine Angaben	5 mm bei 1,5 m	14 mm bei 1m

Tabelle 3.1.: Technische Eigenschaften der Sensoren Kinect One, pmd nano und DS325.

- *Reichweite und Robustheit gegenüber Umgebungslicht:* Der Anwendungsbereich ist die Kopfpose des Fahrers im Fahrzeuginnenraum. Für die Reichweite des Systems genügt deshalb ein Aufnahmebereich zwischen 500 und 1000 mm. Das System muss eine Robustheit gegenüber Umgebungslicht aufweisen, um im Fahrzeuginnenraum zu jedem Zeitpunkt verwertbare Daten zu liefern.

3.3.2. Vergleich verschiedener TOF-Sensoren

Zur Verwendung eines Tiefen-Sensors im Fahrzeuginnenraum wird ein geeigneter Time-Of-Flight-Sensor (TOF) gesucht. Alternativ wären auch Stereo-Vision-Systeme oder Structured-Light-Sensoren denkbar. Letztere übertreffen TOF-Sensoren bei der Tiefenauflösung und der lateralen Auflösung, allerdings ist ein großer Vorteil von TOF-Sensoren die geringe Rechenleistung gegenüber Stereo-Vision-Systemen, welche zur Berechnung der Tiefendaten benötigt wird. Während Stereo-Vision-Systeme die Tiefenwerte aus korrespondierenden Punkten berechnen, werden diese Informationen bei TOF-Sensoren aus der Zeit, die ein einzelner Lichtstrahl benötigt, mit weniger Rechenaufwand bestimmt. Zusätzlich bleiben TOF-Sensoren im Gegensatz zu Structured-Light-Sensoren auch bei Fremdlicht noch funktionsfähig (Li, 2014). Ein weiterer Vorteil der TOF-Sensoren ist, dass neben dem Tiefenbild ein pixel-weise übereinstimmendes Infrarotbild geliefert wird, welches Grauwerte zeigt. Insbesondere im Hinblick auf den Serieneinsatz in Fahrzeugsteuergeräten und die spezifischen Anforderungen im Kraftfahrzeug sind TOF-Sensoren vorteilhaft, so dass der Vergleich auf TOF-Sensoren beschränkt ist.

Im Folgenden werden drei Time-Of-Flight(TOF)-Sensoren vorgestellt und eine experimentelle Analyse durchgeführt. Anhand der aufgestellten Anforderungen im vorherigen Abschnitt wird der Sensor gewählt, welcher am besten geeignet ist.

Tabelle 3.1 gibt einen Überblick der technischen Eigenschaften der verglichenen Sensoren, die im Folgenden untersucht werden. Der Kinect One Sensor (Microsoft, 2017) wird zusammen mit der Spielekonsole Xbox One geliefert und kann damit als Eingabegerät im Wohnzimmer für Spiele mit Bewegungserkennung verwendet werden. Ähnlich zu dem Sensor der Spielekonsole sind die Sensoren PMD nano (Photonics,

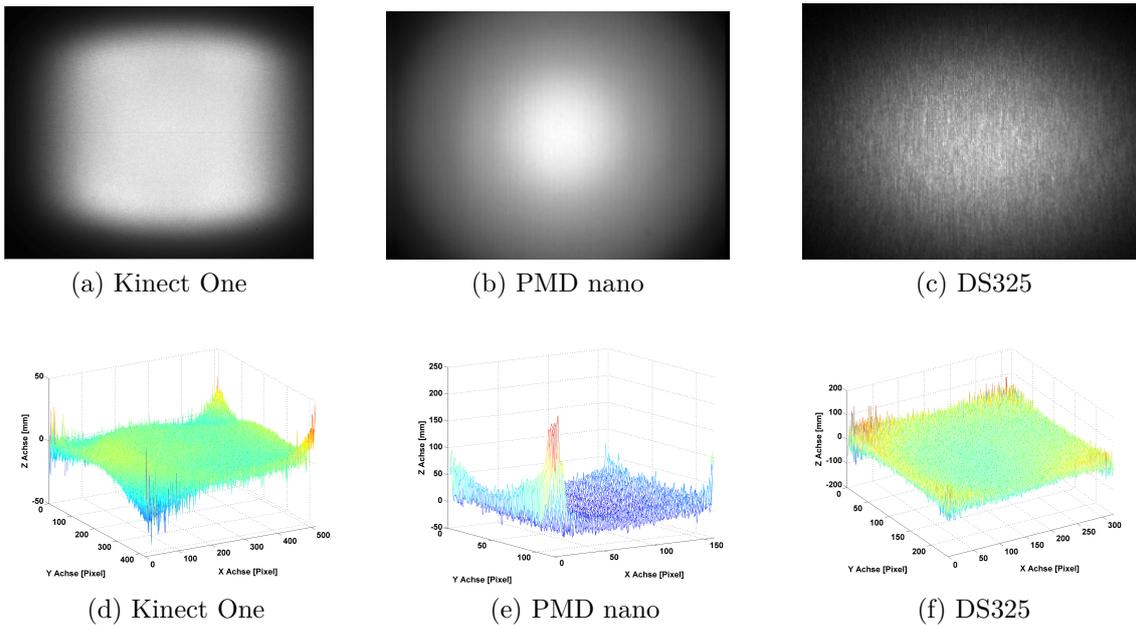


Abbildung 3.5.: Intensitäts- und Tiefenbilder einer Wand aufgenommen mit den verschiedenen TOF-Sensoren aus Wolman (2014). In den Spalten sind die Ergebnisse für die Sensoren Kinect One, PMD nano und DS325 visualisiert.

2014) und DS325 (Softkinetic, 2014) zur Gestensteuerung konzipiert. Die drei Sensoren sind alle für den Indoor Bereich hergestellt. Deshalb geben die Datenblätter der Sensoren keine Hinweise auf die Robustheit gegenüber Sonnenlicht.

Nachfolgend wird ein qualitativer Vergleich der Sensoren diskutiert und anschließend die Präzision und die laterale Auflösung bestimmt. Basierend auf den Ergebnissen der experimentellen Analyse wird der beste Sensor im Sinne der Anforderungen ausgewählt und zusätzlich auf Robustheit gegen Umgebungslicht untersucht. Bei dem PMD Nano (Photonics, 2014) Sensor muss die Integrationszeit manuell eingestellt werden. In den folgenden Experimenten wird der maximale Wert von $2000 \mu s$ eingestellt, um die bestmögliche Präzision zu erhalten. Bei der Standardeinstellung wird ein bilateraler Filter zur Glättung der Daten verwendet, welcher die gemessene laterale Auflösung verringert und die Präzision erhöht. Dieser wird für die Bestimmung der lateralen Auflösung ausgeschaltet, um hierbei den bestmöglichen Wert zu bestimmen.

Qualitativer Vergleich

Um die Sensoren qualitativ zu vergleichen, wird eine ebene Fläche aufgenommen und die einzelnen Tiefen- und Infrarotbilder verglichen. Hieraus kann die Qualität der Daten abgeleitet werden.

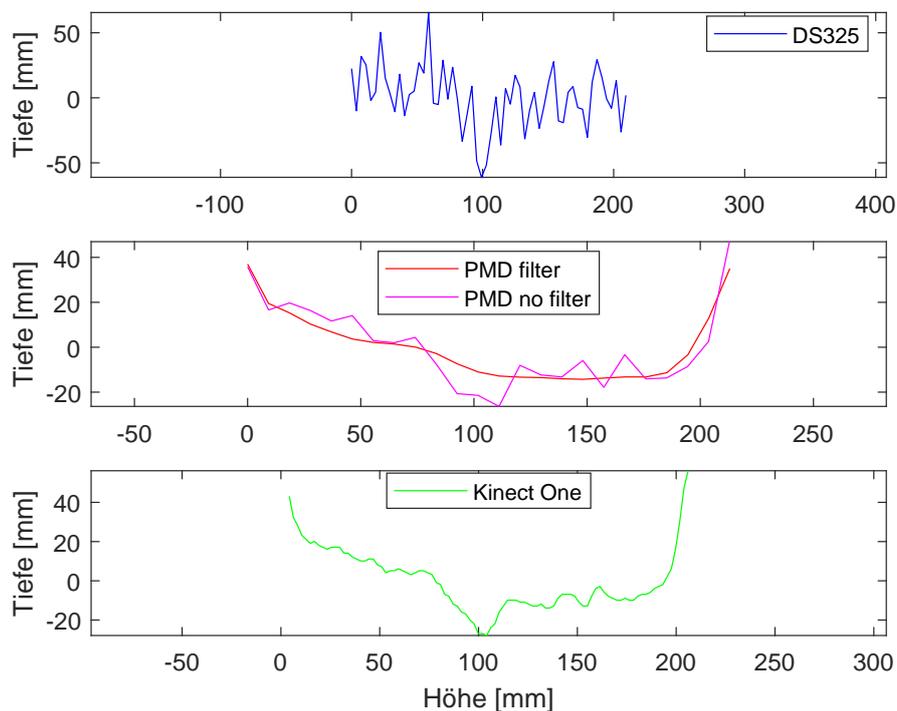


Abbildung 3.6.: Tiefenprofil eines seitlich aufgenommenen Gesichts mit den Sensoren DS325, PMD nano und Kinect One aus Wolman (2014).

Abbildung 3.5 zeigt die Intensitäts- und Tiefenbilder der unterschiedlichen Sensoren. Bei den Sensoren DS325 (Softkinetic, 2014) und PMD (Photonics, 2014) sind kreisförmige Flächen zu erkennen bei denen die Intensität radial nach außen abnimmt. Zusätzlich beinhaltet das Intensitätsbild des DS325 Sensors ein kontinuierliches Rauschen. Durch die kreisförmige Intensität sind die Ränder nicht ausreichend stark ausgeleuchtet. Dadurch zeigen die Tiefenbilder ein verstärktes Rauschen an den Rändern. Während bei dem DS325-Sensor nicht erkennbar ist, ob am Rand die Tiefenwerte über- oder unterschätzt werden, ist bei dem PMD-Sensor eine systematische Überbewertung der Tiefe vorhanden. Im Gegensatz dazu ist das Intensitätsbild des Kinect One-Sensors (Microsoft, 2017) eher rechteckig und zeigt innerhalb des Rechtecks eine homogene Intensität. Das Tiefenbild hat eine konkave Form, was auf eine Unterbewertung der Tiefe an den Bildrändern hindeutet.

Um die Qualität der Sensoren speziell für den im Rahmen dieser Arbeit untersuchten Anwendungsfall zur Bestimmung der Kopfpose qualitativ zu untersuchen, wird ein weiteres Experiment durchgeführt. Hierzu wird ein Kunstkopf jeweils im Abstand von 700 mm zu den verschiedenen Sensoren positioniert, um Aufnahmen des Tiefenprofils des Gesichts zu erhalten.

Abbildung 3.6 zeigt das seitliche Gesichtsprofil aufgenommen mit den drei ausgewählten Sensoren. Während mit dem DS325-Sensor (Softkinetic, 2014) die Struktur eines Gesichts durch die starke Streuung der Messwerte nicht erkennbar ist, kann

Sensor	Streuung (2σ) in mm	
	Ohne Beleuchtung	Halogen Beleuchtung
Kinect One	$\pm 2,03$	$\pm 3,04$
DS325	$\pm 17,27$	$\pm 78,68$
pmd nano	$\pm 10,34$	$\pm 48,83$

Tabelle 3.2.: Experimentelle Ergebnisse der Streuung unter idealen Bedingungen im abgedunkelten Raum.

mit dem PMD nano-Sensor (Photonics, 2014) die Gesichtsstruktur bereits ohne Anwendung des Filters grob erkannt werden. Ein weitaus besseres Gesichtsprofil zeigt der Kinect One Sensor (Microsoft, 2017), bei welchem sogar die Lippen erkennbar sind.

Streuung der Tiefendaten

Zur experimentellen Untersuchung der Streuung der Tiefendaten werden Tiefenbilder einer Wand im Abstand von 700 mm aufgenommen. Aus den Tiefenbildern wird eine Fläche die der Größe eines Kopfes entspricht aus der Mitte herausgeschnitten. Dadurch haben die Randflächen, welche ein starkes Rauschen beinhalten, keinen Einfluss auf die Messung. Die Aufnahme der Wand wird 100-Mal ohne Beleuchtung und 100-Mal unter Beleuchtung mit einem Halogenstrahler wiederholt. Der Halogenstrahler liefert annähernd die Intensität des Sonnenlichts im Infrarotbereich von 850 nm.

Tabelle 3.2 zeigt die statistisch berechnete Präzision über alle Bilder und Bildpunkte der einzelnen Sensoren mit und ohne Beleuchtung. Hierbei ist die Streuung mit $\pm 2\sigma$ angegeben. Insgesamt wird die Streuung durch die Beleuchtung mit dem Halogenstrahler stark beeinflusst. Die Streuung der Sensoren DS325 und pmd nano liegen ohne Beleuchtung bei $\pm 17,27$ mm und $\pm 10,34$ mm, während sie unter Beleuchtung auf $\pm 78,68$ mm und $\pm 48,83$ mm ansteigen. Die geringste Streuung liefert die Kinect One unter Beleuchtung mit $\pm 3,04$ mm. Ohne Beleuchtung liegt die Streuung der Kinect One nur bei $\pm 2,03$ mm.

Laterale Auflösung

Die laterale Auflösung wird sowohl theoretisch als auch experimentell für alle Sensoren bestimmt. Nach Langmann et al. (2012) wird zur theoretischen Untersuchung das angegebene effektive Messfeld, Field of View, durch die Anzahl der Pixel dividiert. Zur experimentellen Untersuchung wird ein Böhlerstern verwendet, der von Wolfgang Böhler 2003 zum Vergleich von Laserscannern verwendet wurde (Langmann et al., 2012). Abbildung 3.7 zeigt den verwendeten Nachbau des Böhlersterns.

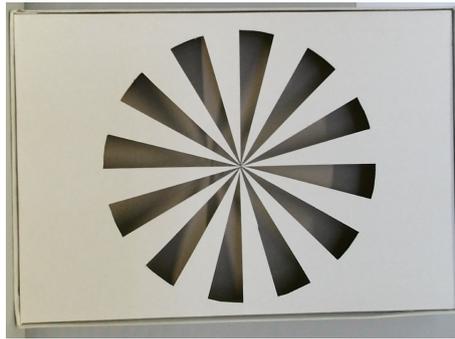


Abbildung 3.7.: Böhlerstern zur Bestimmung der lateralen Auflösung der Sensoren aus Wolman (2014).

Die genaue Analyse ist in Wolman (2014) beschrieben. Hierbei ergeben sich vergleichbare Ergebnisse bei der theoretischen und experimentellen Untersuchung. Die Kinect One und der DS325-Sensor liefern eine laterale Auflösung von 3 mm während der PMD nano eine laterale Auflösung von 8 mm erreicht.

3.3.3. Robustheit gegen Umgebungslicht

Aus dem vorherigen Abschnitt geht hervor, dass der Kinect One Sensor die besten Eigenschaften liefert. Im Folgenden wird dieser Sensor detailliert auf den Einfluss von Umgebungslicht untersucht. Hierbei wird sowohl im Labor ein Experiment mit einem Halogenstrahler durchgeführt, als auch im Fahrzeug unter realer Sonneneinstrahlung gemessen.

Versuchsaufbau

Abbildung 3.8 (a) zeigt den schematischen Aufbau des Experimentes im Labor. Ein Kunstkopf ist frontal auf den Kinect One-Sensor ausgerichtet. Zur Bestrahlung werden im Labor zwei Halogenstrahler verwendet, die sich über dem Sensor befinden. Zusätzlich befindet sich an dem Kinect One Sensor der Messkopf eines Radiometers. Für diesen Versuch wurde das Radiometer ILT 1400 der Firma International Light Technology verwendet. Dieses misst die Bestrahlungsstärke E in $\frac{\mu W}{cm^2}$ bei Veränderung der Bestrahlungsstärke der Halogenstrahler.

Im Fahrzeug ist der Versuchsaufbau vergleichbar, indem der Kunstkopf hinter dem Lenkrad fixiert ist (siehe Abbildung 3.8 (b)). Statt der künstlichen Beleuchtung mit Halogenstrahlern wird ein sonniger Tag gewählt und das Fahrzeug im Freien platziert. Dasselbe Radiometer misst die Bestrahlungsstärke E in $\frac{\mu W}{cm^2}$ für unterschiedliche Situationen, wie zum Beispiel Öffnen des Fensters, des Schiebedachs oder der Türe.

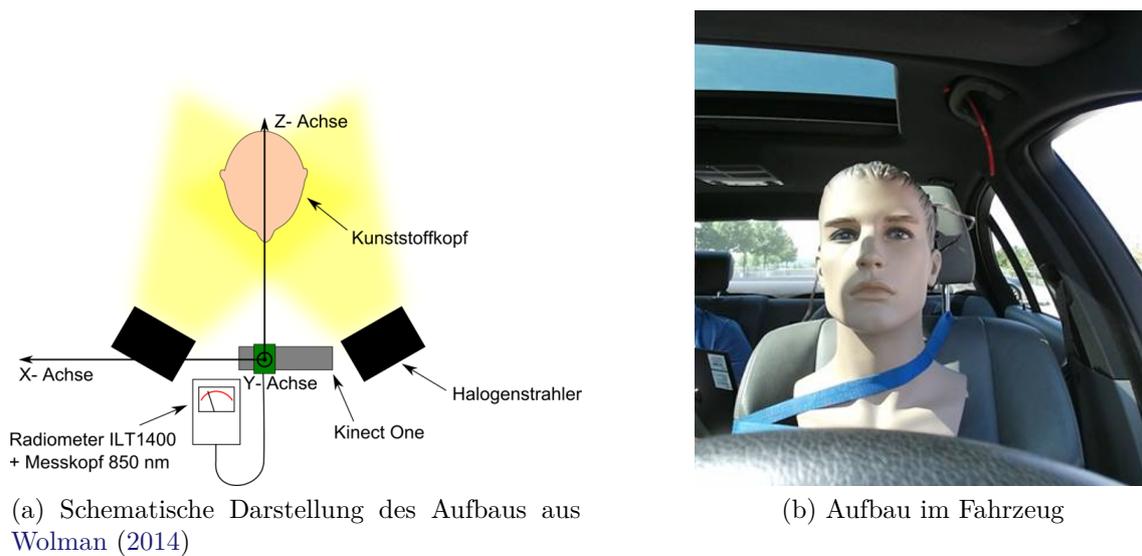


Abbildung 3.8.: Experimentaufbau zur Untersuchung der Robustheit gegen Umgebungslicht, im Labor (a) und im Fahrzeug (b).

Statistische Auswertung

Zur statistischen Analyse der Messdaten unter Sonneneinstrahlung wird im Labor und Fahrzeug die Streuung der Daten bei unterschiedlichen Lichtverhältnissen gemessen. Hierfür ist eine Referenzmessung unter idealen Bedingungen, ohne Lichteinstrahlung, notwendig. Anschließend wird in Bezug auf diese Referenzmessung das Rauschverhalten der Daten analysiert. Während im Labor die Szene zwischen der Messung ohne Lichteinstrahlung und mit Lichteinstrahlung unverändert bleibt, ist dies im Fahrzeug nicht möglich, da das Fahrzeug zwischendurch bewegt wird. Deshalb werden im Folgenden zwei Methoden zur Bestimmung der Streuung beschrieben. Im Labor werden beide Methoden angewendet, während im Fahrzeug nur die zweite Methode angewendet wird. Es wird zuerst die Referenzmessung beschrieben und anschließend werden die beiden Methoden vorgestellt.

Referenzmessung. Die Referenzmessung soll die Daten ohne Rauschen präsentieren. Da im Rahmen dieser Untersuchung kein weiteres hoch-genaues 3D-Messsystem, wie zum Beispiel ein 3D-Laserscanner, zur Verfügung stand, wird eine Referenzpunktewolke durch Mittelung mehrerer Aufnahmen erstellt. Die einzelnen Messungen wurden unter idealen Lichtbedingungen im abgedunkelten Raum ohne Veränderung der Szene aufgenommen. Hierfür wurden 200 Punktewolken von dem Kunststoffkopf aufgenommen und der Mittelwert daraus gebildet. Daraus ergibt sich eine glatte 3D-Punktewolke des Kunststoffes, siehe Abbildung 3.9 (a). Mit den folgenden Methoden werden Bilder des Kunststoffes bei unterschiedlichen Lichtverhältnissen mit dieser Referenzpunktewolke verglichen.

Methode 1 zur Bestimmung der Streuung: Standardabweichung des Differenzbildes. Falls die Szene bei der Aufnahme der Referenzmessung und unter

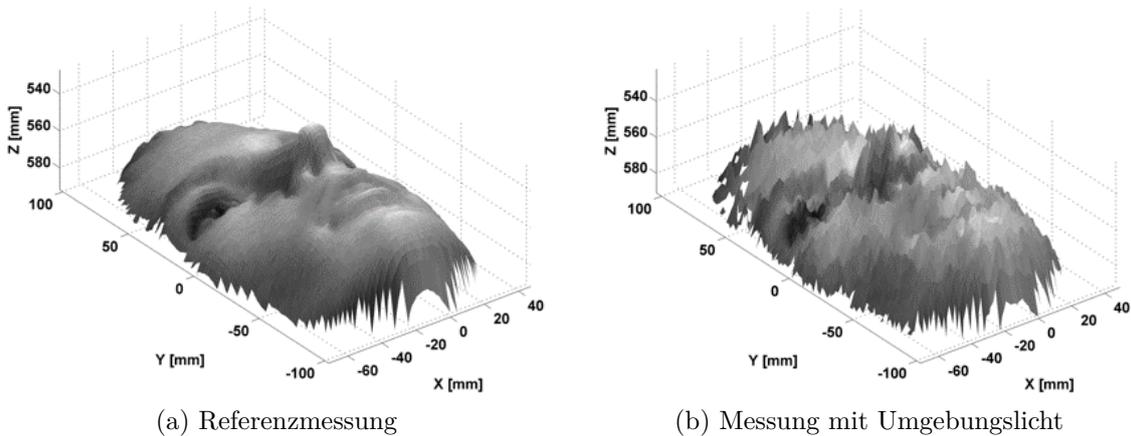


Abbildung 3.9.: (Wolman, 2014): Punktwolke der Referenzmessung (a) und einer Messung mit Umgebungslicht (b).

Einstrahlung des Lichtes unverändert bleibt, kann die Standardabweichung direkt bestimmt werden. Hierfür wird das Differenzbild der Referenzpunktwolke und der momentanen Aufnahme berechnet. Anschließend wird aus der Verteilung der Differenzen die Streuung bestimmt. Da das Histogramm der Differenzen normal-verteilte Häufigkeiten anzeigt (siehe Wolman (2014), Seite 58, Abb. 4.9), kann die Streuung wie folgt bestimmt werden:

$$\sigma = \sqrt{\frac{1}{K-1} \sum_{m=1}^K (d_m - \bar{d})^2} \quad (3.1)$$

wobei d_m die Tiefendifferenz des Pixels m ist und \bar{d} der Mittelwert der Tiefendifferenzen ist.

Methode 2 zur Bestimmung der Streuung: *Root Mean Squared-Fehler (RMS) nach ICP-Registrierung.* Bei Veränderung der Szene existiert keine pixel-basierte Übereinstimmung der Referenzpunktwolke und der aufgenommenen Szene. Um diese herzustellen wird eine *Iterative Closest Point*-Registrierung (ICP) (Besl und McKay, 1992) der aufgenommenen Szene zur Referenzmessung durchgeführt. Zu Beginn werden manuell die Gesichtslandmarken des Kunstkopfes der neuen Szene und der Referenzmessung markiert und eine Anfangsregistrierung hergestellt. Anschließend werden zehn Iterationen des *ICP*-Verfahrens mit dem Matlabcode von Kjer und Wilm (2010) durchgeführt. Bei dem *ICP*-Verfahren wird eine Punkt-zu-Punkt-Minimierung durchgeführt, hierbei wird folgender *Root Mean Squared-Fehler (RMS)* minimiert:

$$RMS = \arg \min_{\mathbf{p}_m, \mathbf{t}} \sqrt{\frac{1}{K} \sum_{m=1}^K \|\mathbf{R}\mathbf{p}_m + \mathbf{t} - \mathbf{q}_m\|^2} \quad (3.2)$$

wobei \mathbf{p}_m der m -te 3D-Voxel der Punktwolke der Quelle und \mathbf{q}_m der 3D-Voxel der Referenzpunktwolke ist. \mathbf{t} ist die Translation und \mathbf{R} die Rotation zur Registrierung der Punktwolken.

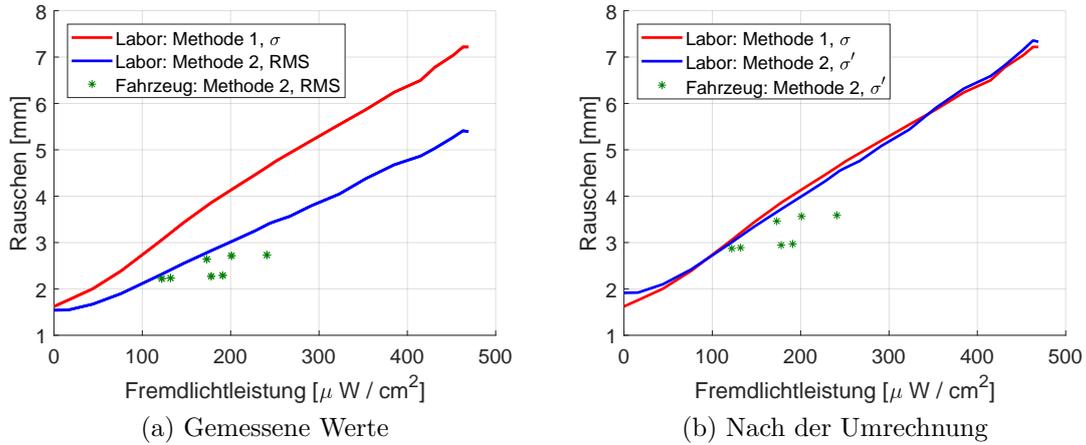


Abbildung 3.10.: Einfluss von Fremdlicht auf das Rauschen der Tiefenwerte. Die linke Grafik zeigt die gemessenen Werte des Rauschens mithilfe von Methode 1 und 2, die rechte Abbildung zeigt die umgerechneten Werte von Methode 2. Aus [Wolman \(2014\)](#).

Bestimmung des systematischen Fehlers. Neben dem Rauschen der einzelnen Datenpunkte ist es interessant zu sehen, ob es einen systematischen Fehler der Datenpunkte gibt. Der systematische Fehler führt in diesem Fall zu einer Verschiebung der Daten in z-Richtung. Dieser systematische Fehler wird anhand des Mittelwerts der wiederholten Messungen bestimmt, mit:

$$err_{sys} = \frac{1}{K} \sum_{m=1}^K d_m \quad (3.3)$$

wobei d_m der Differenzwert des Pixels m vom Referenzbild zum aufgenommenen Wert ist.

Resultate

Abbildung 3.10 (a) zeigt das Rauschen in Millimeter abhängig von der gemessenen Bestrahlungsstärke E in $\frac{\mu W}{cm^2}$. Für jede Beleuchtungsstärke wurden im Labor zehn Messungen aufgenommen und für jede Messung die Standardabweichung und der *RMS*, wie oben beschrieben, bestimmt. In der Grafik sind die Mittelwerte der Streuung aufgetragen und die Fehlerbalken geben die Spannweite der Ergebnisse an. Die Kurven zeigen jeweils die Wurzel des mittleren quadratischen Fehlers der euklidischen Distanz zwischen den Referenzpunkten und der gemessenen Szene, wobei die Bestimmung aus dem Differenzbild in rot dargestellt ist und die Berechnung des Fehlers nach einer *ICP*-Registrierung in blau gezeigt wird. Im Vergleich zur Bestimmung der Streuung aus dem Differenzbild, beinhaltet die zweite Methode eine zusätzliche Registrierung durch das *ICP*-Verfahren. Dadurch ist die ermittelte Streuung mit der zweiten Methode geringer als mit der ersten Methode. Beide Methoden zeigen jedoch einen näherungsweise linearen Zusammenhang zwischen der

Fenster	Türen	Schiebedach	Schwebehimmel	$E[\frac{\mu W}{mm^2}]$	σ' [mm]
				122	3
			x	132	3,2
		x	x	173	3,7
x			x	178	3,2
x				191	2,9
x		x	x	201	3,6
x	x	x	x	241	3,9

Tabelle 3.3.: Wolman (2014): Robustheit gegen Umgebungslicht verschiedener Szenarien im Fahrzeug. σ' gibt das Rauschen in mm an.

Zunahme der Fremdlichtleistung und der Zunahme des Rauschens. Mithilfe der Ausgleichsgeraden $y_1 = m_1 * x + b_1$ durch die Streuungen aus den Differenzbildern und $y_2 = m_2 * x + b_2$ durch die *RMS* Fehler werden mit folgender Formel die *RMS*-Werte näherungsweise in die Streuungen umgerechnet, um einen quantitativen Vergleich zu ermöglichen:

$$\sigma'(n) \approx (RMS(n) - b_2) \frac{m_1}{m_2} + b_1 \quad (3.4)$$

Abbildung 3.10 (b) zeigt das Ergebnis der umgerechneten *RMS*-Werte, wodurch sich eine Überlagerung der Kurven ergibt. Im Folgenden kann aus den im Fahrzeug bestimmten *RMS*-Werten näherungsweise die dazugehörige Streuung berechnet werden.

Im Fahrzeug wurden verschiedene Situationen, die während des Fahrens oder im Stillstand des Fahrzeugs auftreten können, nachgestellt. Tabelle 3.3 zeigt die untersuchten Situationen mit Markierungen für die geöffneten Bereiche. Für jede Situation wurde die Bestrahlungsstärke E am Kinect One Sensor in $\frac{\mu W}{cm^2}$ gemessen und der *RMS*-Fehler mithilfe des *ICP*-Verfahrens bestimmt. Die *RMS*-Fehler sind im Diagramm 3.10 (a) als grüne Sterne dargestellt. Nach der Umrechnung in Diagramm 3.10 (b) visualisiert, sind für jede Situation die zugehörigen Standardabweichungen σ' vorhanden. Diese Werte sind in Tabelle 3.3 angegeben.

Die im Fahrzeug gemessenen Werte für das Rauschen befinden sich in Abbildung 3.10 in der Nähe der im Labor bestimmten Rauschwerte mit derselben Fremdlichtleistung. Allerdings gibt es einige Ausreißer, zum Beispiel Situation 5. Bei diesem Szenario ist das Fenster geöffnet, ansonsten ist alles geschlossen. Hier wird eine hohe Bestrahlungsstärke am Sensor gemessen, das Rauschen im Bild ist jedoch gering. Das bedeutet, dass der Sensor mit Sonnenlicht bestrahlt wird, während das Gesicht nicht angestrahlt wird. Das Rauschen der Daten für 2σ beträgt im Fahrzeug bis zu $\pm 7,8$ mm bei starker Sonnenstrahlung.

Die Verschiebung des Mittelwertes, abhängig von der Bestrahlungsstärke, ist in Abbildung 3.11 dargestellt. Mit der roten Kurve sind die durchschnittlichen systematischen Fehler (siehe Formel 3.3) dargestellt und die Balken geben die Spannweite

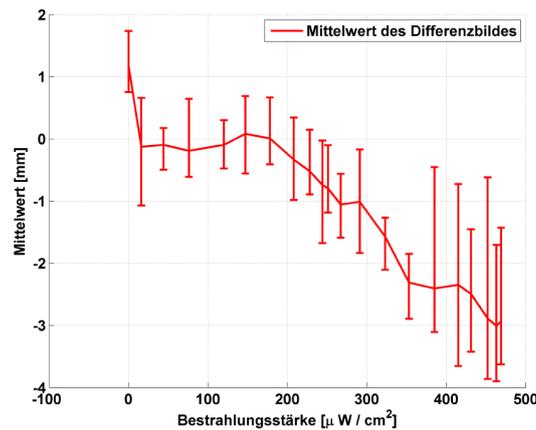


Abbildung 3.11.: Verschiebung des Mittelwertes in Abhängigkeit von der Bestrahlungsstärke (Wolman, 2014).

der systematischen Fehler an. In der Grafik zeigen die großen Spannweiten starke Abweichungen innerhalb einer Bestrahlungsstärke an. Insgesamt ist ein leichter Anstieg bei Erhöhung der Bestrahlungsstärke des Mittelwertes in die negative Richtung sichtbar. Im Fall der stärksten Bestrahlung im Gesicht in der Mittagssonne mit $E = 240 \frac{\mu\text{W}}{\text{cm}^2}$ entspricht dies einer Verschiebung der Tiefenwerte im Mittel um 2 mm.

3.3.4. Ergebnisse

Die in diesem Kapitel diskutierten Experimente zeigen, dass der Kinect One Sensor am besten für die in Abschnitt 3.3.1 aufgeführten Anforderungen geeignet ist. Zusätzlich zu den Anforderungen wurden die Sensoren noch qualitativ verglichen. Die Resultate sind im folgenden Abschnitt zusammengefasst.

- *Qualitativer Vergleich:* Beim qualitativen Vergleich wurden die Intensitäts- und Tiefenbilder der Sensoren bei der Aufnahme einer glatten Wand verglichen. Zusätzlich wurde das Tiefenbild eines Gesichtsprofils analysiert. Die Kinect One zeigt eine gleichmäßigere Ausleuchtung der glatten Flächen im Vergleich zu dem pmd nano- und DS325-Sensor. Aus dem Tiefenbild des seitlichen Gesichts ist deutlich ein Gesichtsprofil zu erkennen, während bei dem pmd nano nur grob ein Gesicht erkennbar ist und bei dem DS325 Sensor das Gesichtsprofil verwaschen ist.
- *Streuung:* Die Streuung des Kinect One-Sensors mit $\pm 2,03$ mm genügt den Anforderungen, während der DS325-Sensor eine Streuung von $\pm 17,27$ mm und der pmd nano-Sensor eine Streuung von $\pm 10,34$ mm hat, die damit die Anforderungen nicht erfüllen.
- *Laterale Auflösung:* Die laterale Auflösung des Kinect One Sensors übertrifft mit 3 mm die Anforderungen.

- *Reichweite*: Die technischen Eigenschaften des Kinect One-Sensors liefern eine Reichweite von 500 bis 1000 Millimetern und genügen damit für den Anwendungsfall im Fahrzeuginnenraum.
- *Robustheit gegenüber Umgebungslicht*: Die Streuung der Kinect One hat sich unter Einfluss eines Halogenstrahlers auf $\pm 3,04$ mm vergrößert. Im Vergleich zum DS325- und PMD nano-Sensor hat ein Halogenstrahler auf die Streuung des Tiefenbildes der Kinect One-Sensor nur einen geringen Einfluss. Bei den Experimenten mit Sonneneinstrahlung zeigt sich ein deutlich stärkerer Einfluss bei der Kinect One. Allerdings zeigen die Experimente mit starker Sonneneinstrahlung, dass ein Rauschen und ein systematischer Fehler bei den Tiefenwerten deutlich erkennbar ist. Während die Streuung der Kinect One bei starker Sonneneinstrahlung bis zu $\pm 7,8$ mm beträgt, ist der systematische Fehler mit bis zu 2 mm gering. Daraus folgt, dass die einzelnen Tiefenwerte einem Rauschen unterliegen, es allerdings nur zu einem geringen mittleren Tiefenfehler der Daten kommt.

Insgesamt haben dieser Vergleich und die durchgeführten Experimente mit dem Kinect One-Sensor gezeigt, dass dieser Sensor für die Aufnahme im Fahrzeuginnenraum geeignet ist. Das Gesichtsprofil ist deutlich erkennbar, die Streuung ausreichend gering und mit der lateralen Auflösung sind Gesichtsdetails erkennbar. Außerdem beinhaltet die Reichweite den Bereich der Kopfposition des Fahrers und Umgebungslicht beeinflusst die Streuung und den systematischen Fehler nur geringfügig. Deshalb wird dieser Sensor zur Aufnahme von Tiefen- und Infrarotbildern für die Aufnahme der Stichprobe in Kapitel 3.6 verwendet.

3.4. Methodik der Datennotation

Im folgenden Kapitel werden die Notationen der einzelnen Bilder des Datensatzes beschrieben. Zur Bestimmung der Genauigkeit der Kopfpose sind akkurate Referenzwerte der Orientierung und Position des Kopfes essentiell. Kapitel 3.4.1 definiert das Kopfkordinatensystem, welches fest im Kopf verankert ist und abhängig von diesem die Kopfpose für jedes Bild bestimmt. Anschließend beschreibt Kapitel 3.4.2 die Herleitung von der Referenzmessung der Kopfpose für jedes Bild des Datensatzes. Damit wird jedes Paar bestehend aus Infrarot- und Tiefenbild mit einer zugehörige Referenzmessung der Orientierung und Position des Kopfes ergänzt. Zur Analyse der Algorithmen auf Robustheit gegenüber Brillen, Sonnenbrillen und Verdeckungen im Gesichtsbereich sind weitere Annotationen notwendig. Diese manuellen Annotationen werden in Kapitel 3.4.3 erläutert.

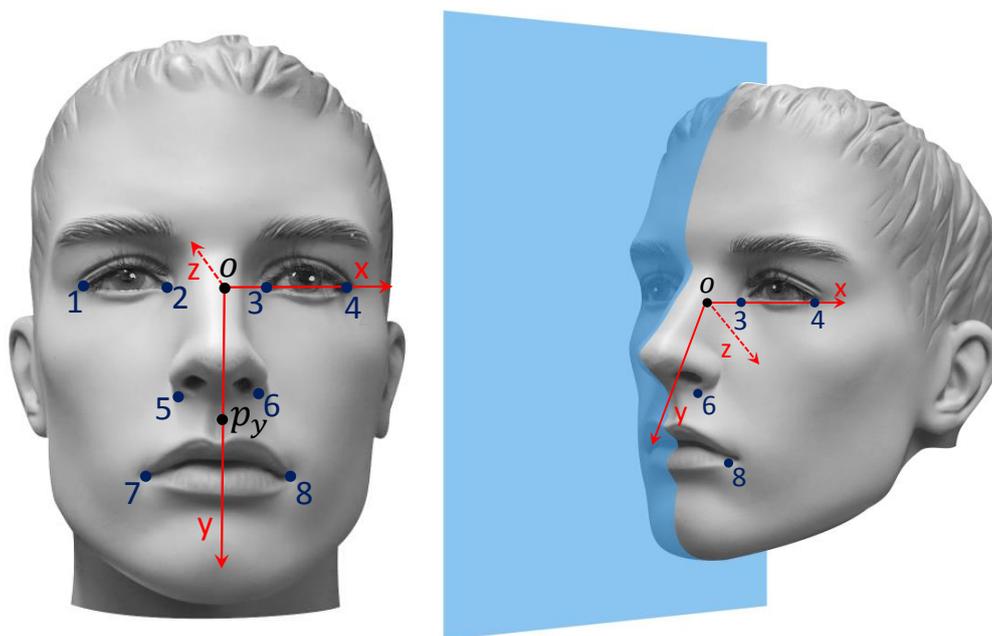


Abbildung 3.12.: Beschreibung des neu definierten Kopfkoordinatensystems basierend auf den Positionen der gemessenen 3D-Gesichtslandmarken. Aus Schwarz et al. (2017) © 2017 IEEE.

3.4.1. Definition des Kopfkoordinatensystems

Eine eindeutige Definition des Kopfkoordinatensystems ist essentiell. Hiermit wird festgelegt, um welche Achsen die Drehwinkel gemessen werden. Das im Rahmen dieser Arbeit vorgestellte Kopfkoordinatensystem basiert auf den 3D-Positionen der Gesichtslandmarken. Bei vorhandenen Arbeiten basiert das Kopfkoordinatensystem meist entweder auf der 3D-Form des kompletten Kopfes mit dem Ursprung im Kopfmittelpunkt (Fanelli et al., 2013) oder es findet ein Punkt-zu-Punkt-Matching zu 3D-Landmarken eines Durchschnittskopfes statt (Baltrušaitis et al., 2016). Im Gegensatz zu einer Definition die auf dem kompletten 3D-Kopf basiert, benötigt die hier vorgestellte Festlegung lediglich die Gesichtslandmarken, die bei einer frontalen Ausrichtung des Kopfes sichtbar sind. Gegenüber des Punkt-zu-Punkt-Matchings wird das Koordinatensystem direkt aus den individuellen Gesichtslandmarken des Probanden bestimmt. Abbildung 3.12 visualisiert das Kartesische Koordinatensystem mit den Achsen x , y und z sowie die 3D-Positionen $\mathbf{l}_{i \in \{1, \dots, 8\}} \in \mathbb{R}^3$ der Gesichtsmerkmale.

Der *Koordinatenursprung* ist definiert als der Mittelpunkt zwischen den Augen und berechnet sich damit aus dem Durchschnitt der vier Augen-Eckpunkte mit:

$$\mathbf{o} := \frac{1}{4}(\mathbf{l}_1 + \mathbf{l}_2 + \mathbf{l}_3 + \mathbf{l}_4).$$

Hiermit ist der Ursprung der einzelnen Koordinatenachsen definiert. Im Folgenden werden die Richtungen der Achsen so festgelegt, dass sie ein rechtshändiges kartesisches Koordinatensystem aufspannen mit orthogonalen Achsen.

Die Landmarken der Augen definieren die Richtung der x -Achse \mathbf{x} , die sich wie folgt definiert:

$$\mathbf{x} := \frac{\boldsymbol{\theta}_3 + \boldsymbol{\theta}_4 - (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)}{\|\boldsymbol{\theta}_3 + \boldsymbol{\theta}_4 - (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)\|}, \quad (3.5)$$

mit

$$\boldsymbol{\theta}_i := \frac{\mathbf{l}_i - \mathbf{o}}{\|\mathbf{l}_i - \mathbf{o}\|}.$$

Die Normalisierung der Vektoren $\boldsymbol{\theta}_i$ ist erforderlich, um im Fall von fehlerbehafteten Messungen der Landmarkpositionen robust zu bleiben.

Die Richtung der y -Achse definiert sich aus dem Vektor zwischen der Position des Ursprungs \mathbf{o} und einem Punkt \mathbf{p}_y , mit:

$$\mathbf{y} := \frac{\mathbf{p}_y - \mathbf{o}}{\|\mathbf{p}_y - \mathbf{o}\|} \quad (3.6)$$

Bei der Wahl von \mathbf{p}_y muss sichergestellt werden, dass die x - und y -Achsen senkrecht zueinander stehen, da es sich um ein Kartesisches Koordinatensystem handelt. Dies ist der Fall sobald die y -Achse in der Ebene liegt, die mit der x -Achse \mathbf{x} als Normalen aufgespannt wird, siehe blaue Ebene in Abbildung 3.12. Hierfür muss folgende Gleichung erfüllt sein, siehe Bronstein et al. (2012):

$$(\mathbf{o} - \mathbf{p}_y)^T \mathbf{x} = 0 \quad (3.7)$$

Als Näherungspunkt von \mathbf{p}_y wird \mathbf{p}'_y als der Mittelpunkt der Nasenflügel und Mundwinkel definiert, mit:

$$\mathbf{p}'_y := \frac{1}{4}(\mathbf{l}_5 + \mathbf{l}_6 + \mathbf{l}_7 + \mathbf{l}_8)$$

\mathbf{p}_y wird als der Punkt mit dem kleinsten Abstand zu \mathbf{p}'_y gewählt, für den Gleichung 3.7 gilt. Dies ist der Schnittpunkt zwischen der Ebene und der Geraden ausgehend von \mathbf{p}'_y mit Richtung \mathbf{x} . Für \mathbf{p}_y auf dieser Geraden gilt, siehe Bronstein et al. (2012):

$$\mathbf{p}_y = \mathbf{p}'_y + d \cdot \mathbf{x} \quad (3.8)$$

mit

$$d = \frac{(\mathbf{o} - \mathbf{p}'_y)^T \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \quad (3.9)$$

Der Abstand d leitet sich aus Einsetzen von Gleichung 3.8 in Gleichung 3.7 her.

Zum Schluss wird die z -Achse als der Vektor senkrecht zu den beiden anderen Achsen mit Richtung auf das Kopfinnere, wie folgt festgelegt:

$$\mathbf{z} := \mathbf{x} \times \mathbf{y} \quad (3.10)$$

Damit ist das Kopfkoordinatensystem h aus den Gesichtslandmarken eindeutig definiert.

3.4.2. Referenzmessung der Kopfpose

Das Ziel ist es, für jedes Bild i die zugehörige Referenz der Position und Orientierung des Kopfes zu liefern. Die Orientierung und Position des Kopfes für Bild i ist die Transformation $\mathbf{T}_i^{k \rightarrow h}$ vom Kamerakoordinatensystem k zum Kopfkoordinatensystem h .

Die Referenzmessung der Kopforientierung und -position wird durch die Kombination der Transformationen wie folgt bestimmt:

$$\mathbf{T}_i^{k \rightarrow h} := \mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t} \cdot \mathbf{T}_n^{t \rightarrow h} \quad (3.11)$$

Dabei handelt es sich um die Transformation zwischen den einzelnen Koordinatensystemen, siehe Abschnitt 3.2.2. Von links nach rechts werden hier die Transformationen vom Kamerakoordinatensystem k über das Weltkoordinatensystem w und das Targetkoordinatensystem t zum Kopfkoordinatensystem h kombiniert. Jede der Transformationen $\mathbf{T} \in \mathbf{R}^{4 \times 4}$ setzt sich zusammen aus einer Rotationsmatrix $\mathbf{R} \in \mathbf{R}^{3 \times 3}$ und einer Translation $\mathbf{t} \in \mathbf{R}^{3 \times 1}$ durch:

$$\mathbf{T} := \left(\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline 0 & 1 \end{array} \right) \quad (3.12)$$

Die einzelnen Transformationen werden wie folgt definiert.

Zunächst ist die Kalibrierung des Motion Capture-Systems und des Kamerakoordinatensystems k zu einem gemeinsamen Weltkoordinatensystem w notwendig. Diese Kalibrierung wird mithilfe eines speziell angefertigten Multi-Schachbretts (siehe Abbildung 3.4 (a)) durchgeführt. Dieses Schachbrett beinhaltet sowohl genau positionierte Marker, die das Motion Capture-System benötigt, als auch ein Schachbrettmuster, das zur Kalibrierung der Kamera notwendig ist. Mit diesem Multi-Schachbrett werden beide Systeme auf ein gemeinsames Weltkoordinatensystem kalibriert. Das Motion Capture-System definiert das Weltkoordinatensystem anhand der Marker-Position. Zur Kalibrierung des Kamerasystems wird ein Bild des Schachbretts aufgenommen und das Verfahren von Zhang (2000) angewendet. Anschließend ist die Transformation dieses Weltkoordinatensystems w zum Kamerakoordinatensystems k bekannt, bezeichnet als $\mathbf{T}^{k \rightarrow w}$.

Zweitens wird die Transformation des Targetkoordinatensystems t in das Weltkoordinatensystem w benötigt, die von dem Motion Capture-System bestimmt wird. Das Target besteht aus mehreren 3D-Kugeln, welche durch das Motion Capture-System verfolgt werden. Das System liefert für jedes Bild i die Transformation des Weltkoordinatensystems w zum Targetkoordinatensystem t , definiert als $\mathbf{T}_i^{w \rightarrow t}$.

Drittens wird die Transformation vom Targetkoordinatensystem t zum Kopfkoordinatensystem h benötigt. Diese Transformation berücksichtigt den Umstand, dass jede Person das Kopftarget an einer individuellen Position und mit einer individuellen Ausrichtung, entsprechend der persönlichen Kopfform, trägt. Diese Transformation wird für jeden Probanden n individuell bestimmt, bezeichnet als $\mathbf{T}_n^{t \rightarrow h}$.

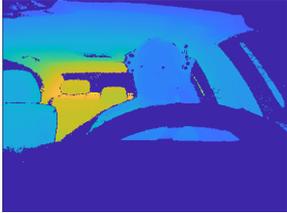
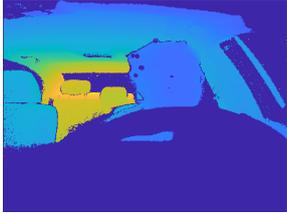
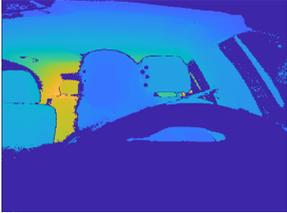
Infrarotbild			
Tiefenbild			
Verdeckung	x	x	–
Brille	x	–	–
Sonnenbrille	–	x	–
Gier-Rotation	3°	-23°	14°
Nick-Rotation	-15°	-1°	-29°
Roll-Rotation	0°	-11°	-8°
x-Translation	35 mm	106 mm	-130 mm
y-Translation	-156 mm	-97 mm	-112 mm
z-Translation	680 mm	573 mm	623 mm

Tabelle 3.4.: Infrarot- und Tiefenbilder des *DriveAhead*-Datensatzes mit den Annotationen. Für jedes Bild sind manuelle Annotationen für Brillen, Sonnenbrillen und weitere Verdeckungen vorhanden sowie Referenzmessungen für die Rotation und Translation des Kopfes.

Um $\mathbf{T}_n^{t \rightarrow h}$ zu berechnen werden mit einem speziellen 3D-Zeiger (siehe 3.4 (b)) die Gesichtslanmarken berührt und so die 3D-Positionen der Landmarken im Targetkoordinatensystem bestimmt. Wie in Abschnitt 3.4.1 beschrieben, wird daraus die Transformation des Targetkoordinatensystems t zum Kopfkoordinatensystem k berechnet.

Aus den beschriebenen Schritten werden die einzelnen Transformationen $\mathbf{T}^{k \rightarrow w}$, $\mathbf{T}_i^{w \rightarrow t}$ und $\mathbf{T}_n^{t \rightarrow h}$ für jedes Bild i bestimmt. Anschließend werden sie in Gleichung 3.11 zu der finalen Referenztransformation $\mathbf{T}_i^{k \rightarrow h}$ vom Kamerakoordinatensystem zum Kopfkoordinatensystem zusammengesetzt.

3.4.3. Manuelle Annotation

Zusätzlich zu den kontinuierlich ermittelten Werten der Orientierung und Position des Kopfes wird jedes Bild i manuell mit Eigenschaften annotiert. Speziell angeleitete Personen markierten jedes Bild mit Parametern, die angeben, ob der Fahrer eine

Sonnenbrille oder Brille trägt. Des Weiteren gibt es für jedes Bild einen binären Wert, der beschreibt, ob Verdeckungen innerhalb des Gesichts vorhanden sind. Ein Gesicht ist als verdeckt markiert, sobald mindestens eine der 68 Gesichtslanmarken, wie in Sagonas et al. (2013) definiert, durch externe Objekte verdeckt ist. Wenn das Gesicht durch starke Rotationen verdeckt ist, zählt dies nicht als Verdeckung. Analog zählen Sonnenbrillen und Brillen nicht als Verdeckungen, da sie in separaten Parametern annotiert sind.

Tabelle 3.4 zeigt drei Bildpaare bestehend aus Infrarot- und Tiefenbild mit den zugehörigen Annotationen. Für jedes Bild ist die Referenzorientierung und -position des Kopfes angegeben sowie die manuellen Parameter für Verdeckungen, Brillen und Sonnenbrillen. Die Translation ist in x-, y- und z-Richtung angegeben. Bei der Rotation werden für den aufgenommenen Datensatz Quaternionen gewählt, wie in Kapitel 2.5 beschriebenen. In Tabelle 3.4 sind die Quaternionen zur besseren Anschaulichkeit in die Winkel-Richtungen Gier, Nick und Roll umgerechnet.

3.5. Charakterisierung des Referenzsystems

Bei der Referenzmessung kann es zu einem systematischen und einem zufälligen Fehler bei jeder Messung kommen. Während der systematische Fehler zu einer konstanten Verschiebung der Referenzwerte führt, kommt es bei dem zufälligen Fehler zu einem unvorhersehbaren Rauschen der Daten. Der systematische Fehler des Referenzsystems wird durch die Wahl eines hoch-genauen Motion Capture-Systems und die Verwendung eines präzisen Multi-Funktions-Kalibrierungsbrettes als gering eingeschätzt, kann allerdings nicht exakt für dieses Referenzsystem bestimmt werden. Für die Bestimmung des *systematischen Fehlers* ist ein System notwendig, welches genauer als das Referenzsystem ist. Ein solches System stand im Rahmen dieser Arbeit nicht zur Verfügung. Neben dem *systematischen Fehler* ist es wichtig, den *zufälligen Fehler* zu kennen, um die Streuung der einzelnen Messungen abzuschätzen. Der *zufällige Fehler* wird durch die Wiederholung von Messungen bei gleichbleibenden Bedingungen bestimmt. Hierdurch erhält man ein Maß für die Streuung der gemessenen Referenzmessungen während der Datenaufnahmen. Dieser Wert definiert die Präzision der Messung.

Im Rahmen dieser Arbeit werden die *zufälligen Fehler* der folgenden Transformationen bestimmt:

- der Transformation $\mathbf{T}_n^{t \rightarrow h}$ des Targetkoordinatensystems t zum Kopfkoordinatensystem h für unterschiedliche Probanden n .
- der Transformation $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$ des Kamerakoordinatensystems k über das Weltkoordinatensystem w zum Targetkoordinatensystem t für das Bild i .
- der gesamten Transformation des Referenzsystems vom Kamerakoordinatensystem k zum Kopfkoordinatensystem h $\mathbf{T}_i^{k \rightarrow h} = \mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t} \cdot \mathbf{T}_n^{t \rightarrow h}$.

Die im folgenden Abschnitt durchgeführte Analyse untersucht experimentell die Präzision der Koordinatentransformationen $\mathbf{T}_n^{t \rightarrow h}$ und $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$. Hierfür werden unter gleichbleibenden Bedingungen diese Transformationen mehrmals gemessen. Anschließend wird hieraus die Präzision der Transformationen bestimmt. Durch die Kombination aller Messungen der beiden experimentell erzeugten Koordinatentransformationen werden Werte der Gesamtreferenz $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t} \cdot \mathbf{T}_n^{t \rightarrow h}$ erzeugt. Mithilfe der kombinierten Werte wird die Streuung der Gesamtreferenzmessung abgeschätzt.

Im folgenden Abschnitt 3.5.1 wird zuerst der experimentelle Aufbau zur Messung der einzelnen Transformationen beschrieben. Anschließend wird die statistische Analyse der Messungen in Abschnitt 3.5.2 durchgeführt.

3.5.1. Experimenteller Aufbau

Um die einzelnen Transformationen $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$ und $\mathbf{T}_n^{t \rightarrow h}$ statistisch zu untersuchen, wurden diese Transformationen unter gleichbleibenden Bedingungen mehrmals gemessen. Im Folgenden wird der Aufbau der Experimente für die Messung der jeweiligen Transformationen genauer erläutert und die Verteilung der gemessenen Translationen und Rotationen visualisiert. Die Rotationen werden bei allen Rechenschritten in Quaternionen ausgedrückt. Zur besseren Anschaulichkeit werden sie in den Abbildungen mit Eulerwinkel in Gier-, Nick- und Roll-Richtungen umgerechnet. Hierbei wird folgende Reihenfolge und Bezeichnung verwendet: Drehung um die Y-Achse als Gier-Richtung, Drehung um die X-Achse als Nick-Richtung und Drehung um die Z-Achse als Roll-Richtung. Diese Bezeichnung entspricht den Drehrichtungen des Kopfes in dem definierten Koordinatensystem aus Kapitel 3.4.1.

Experimentelle Untersuchung der Koordinatentransformation $\mathbf{T}_n^{t \rightarrow h}$

Die Transformation $\mathbf{T}_n^{t \rightarrow h}$ beinhaltet die Rotation und Translation des Targetkoordinatensystems t in das Koordinatensystem des Kopfes h abhängig vom Probanden n . Diese Koordinatentransformation wird anhand von Gesichtslanmarken definiert (siehe Abbildung 3.12). Um die Streuung dieser Transformation zu bestimmen, haben vier Probanden mehrmals die Positionen der Gesichtslanmarken mit dem im vorherigen Kapitel beschriebenen Messzeiger berührt und damit vermessen. Aus den vermessenen Gesichtslanmarken wird für jeden Durchlauf $\mathbf{T}_n^{t \rightarrow h}$ bestimmt (siehe Kapitel 3.4.1). Abbildung 3.13 zeigt die Verteilung der gemessenen Transformationen normalisiert mit der mittleren Transformation jedes Probanden n . Da jeder Proband das Target unterschiedlich auf dem Kopf trägt, ist die Transformation $\mathbf{T}_n^{t \rightarrow h}$ personenabhängig.

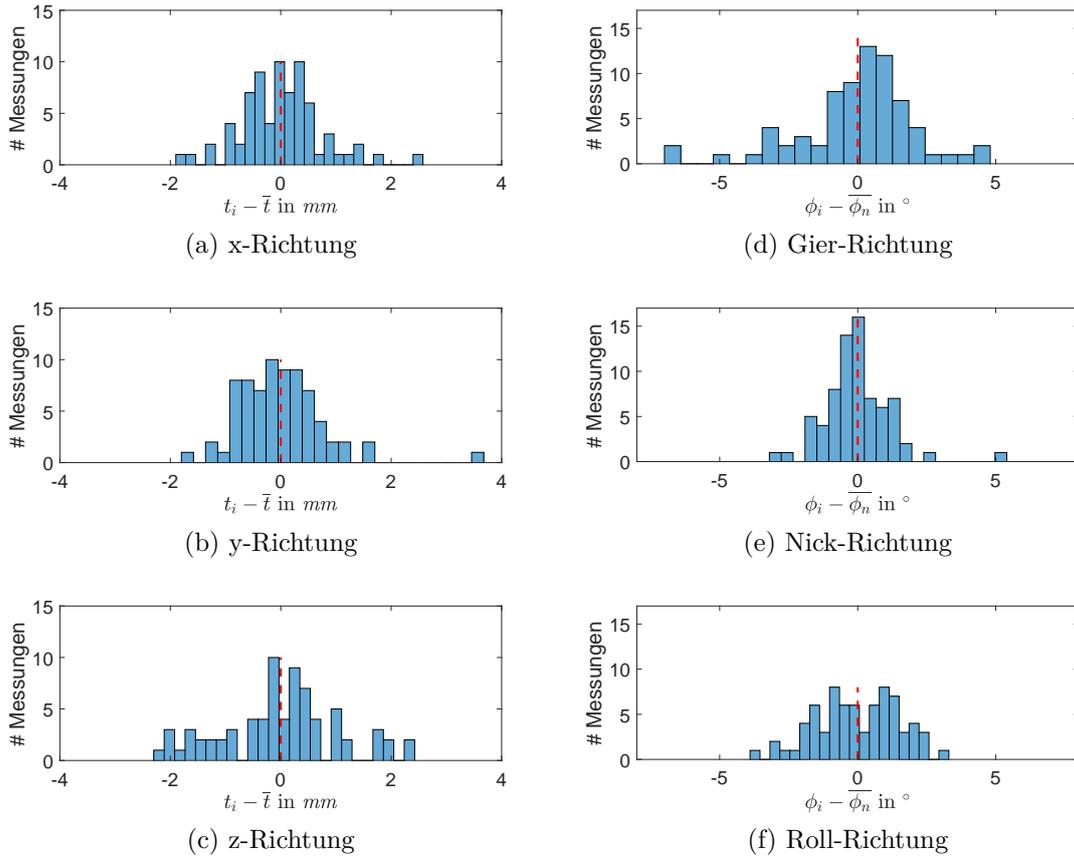


Abbildung 3.13.: Histogramme der gemessenen Transformationen $T_n^{t \rightarrow h}$ normiert auf die mittlere Transformation. Die Komponenten der Transformation sind aufgeteilt in die Translationen x-, y- und z-Richtung (linke Spalte) sowie in die Winkel Gier-, Nick- und Roll-Richtung (rechte Spalte).

Experimentelle Untersuchung der Koordinatentransformation $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$

Die Koordinatentransformation $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$ vom Kamerakoordinatensystem k in das Koordinatensystem des Targets t wird experimentell untersucht, indem unter gleichbleibenden Bedingungen diese Transformation mehrmals vermessen wird. Abbildung 3.14 zeigt den Aufbau des Experiments. Während des Experiments sind das Schachbrett zur Kalibrierung des Motion Capture-Systems und das Kopf-Target, dessen Position und Orientierung durch das Motion Capture-System bestimmt werden, fixiert. Die Kalibrierung des Kamerasystems zum Weltkoordinatensystem bestimmt die Transformation $T^{k \rightarrow w}$. Diese ist während der Aufnahme konstant, da das Schachbrett fixiert ist und damit das Bild des Schachbretts unverändert bleibt. Die Kalibrierung des Motion Capture-Systems wird 38-mal wiederholt und im Anschluss werden jeweils die Orientierung und Position des Targets bestimmt. Diese einzelnen Messungen M_1, \dots, M_{38} beinhalten den zufälligen Fehler der Transformation vom Weltkoordinatensystem w zum Targetkoordinatensystem t $T_i^{w \rightarrow t}$. In Abbildung 3.15 sind die gemessenen Transformationen normalisiert auf die mittlere Transformation in Histogrammen dargestellt.



Abbildung 3.14.: Versuchsaufbau zur Bestimmung der Streuung der Transformation des Kamerakoordinatensystems k in das Koordinatensystem des Motion-Capture-Targets t über das Weltkoordinatensystem w , $\mathbf{T}^{k \rightarrow t} = \mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$. Das Schachbrett und das Motion Capture-System sind während der Messwiederholungen fixiert. Bei jeder Messung wird die Kalibrierung neu durchgeführt und anschließend Orientierung und Position des Targets gemessen.

3.5.2. Statistische Analyse der Streuung des Referenzsystems

Das Ziel der statistischen Analyse ist die Streuung der Referenzmessung, um so den zufälligen Fehler während der Datenaufnahme herauszufinden. Wie in [Fahrmeir et al. \(2016\)](#) erläutert, kann man bei etwa normalverteilten Daten annehmen, dass im Intervall $\bar{x} \pm 2\sigma$ ca. 95% der Daten liegen. Deshalb werden im Folgenden die Werte von 2σ der einzelnen Transformationen sowie der Gesamttransformation bestimmt (siehe Tabelle 3.5).

Die Abbildungen 3.13 und 3.15 zeigen die Messungen der Translationen in x-, y- und z-Richtung und die gemessenen Rotationen in Gier-, Nick- und Roll-Richtung. Da es sich bei den Verteilungen in den Histogrammen näherungsweise um Normalverteilungen handelt (siehe Abbildung 3.13 und 3.15) wird im Folgenden eine statistische Analyse basierend auf [Fahrmeir et al. \(2016\)](#) für normalverteilte Messungen durchgeführt.

Der Mittelwert \bar{v} von K Messungen berechnet sich aus:

$$\bar{v} = \frac{1}{K} \sum_{j=1}^K v_j \quad (3.13)$$

Für die Rotationen wird der Mittelwert aus Quaternionen mit den Rechenregeln aus [Markley et al. \(2007\)](#) bestimmt, dieser Mittelwert ist eindeutig festgelegt.

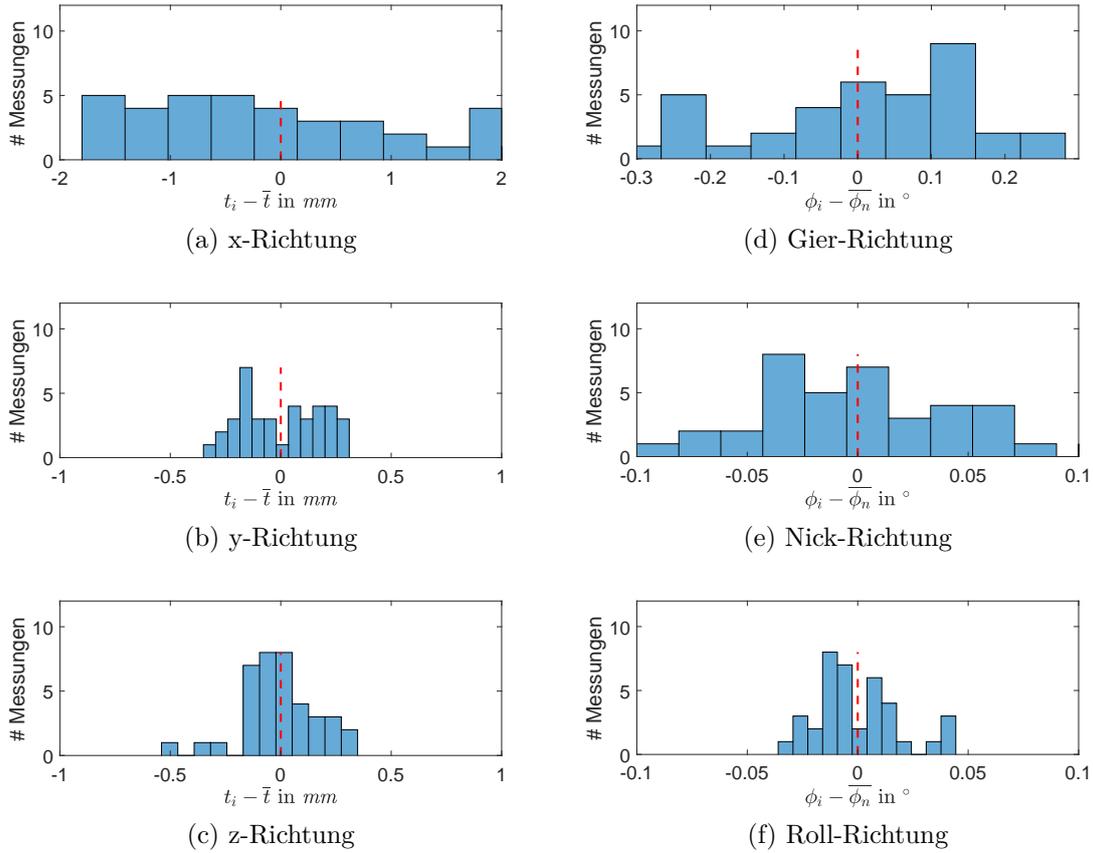


Abbildung 3.15.: Histogramme der gemessenen Transformation des Kamerakoordinatensystems in das Targetkoordinatensystem $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$. Verteilung der gemessenen Translation in x-, y- und z- Richtung (linke Spalte), Verteilung der gemessenen Rotation in Gier-, Nick-, und Roll-Richtung (rechte Spalte).

Für die Streuung der normalverteilten Messungen gilt im allgemeinen folgende Formel:

$$\sigma_x = \sqrt{\frac{1}{(K-1)} \sum_{j=1}^K (v_j - \bar{v})^2} \quad (3.14)$$

Mit dieser Formel wird die Streuung der Translation und Rotation bestimmt. Die Streuung der Translation σ_t berechnet sich aus dem Mittelwert der euklidischen Distanz von jeder Messung t_j zum Mittelwert der Translation \bar{t} :

$$\sigma_t = \sqrt{\frac{1}{(K-1)} \sum_{j=1}^K (t_j - \bar{t})^2} \quad (3.15)$$

Für die Rotation wird der Abstand der gemessenen Rotation q_i in Quaternionen von der mittleren Rotation \bar{q} in Quaternionen mit der Formel $2 \cdot \arccos(\langle \mathbf{q}_j, \bar{\mathbf{q}} \rangle)$

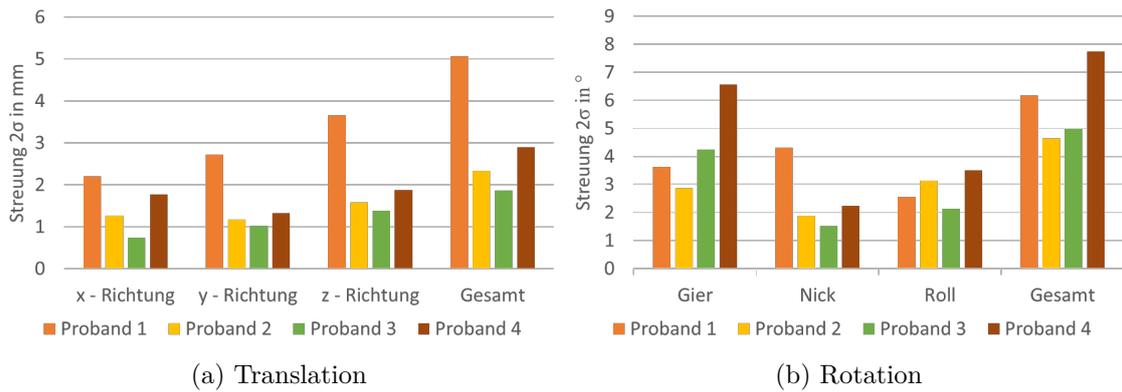


Abbildung 3.16.: Personenabhängige Streuung $\mathbf{T}_n^{t \rightarrow h}$. In (a) ist für jeden Probanden einzeln die Streuung $2\sigma_t$ der Translation für die x-, y- und z-Richtung sowie die Gesamtstreuung dargestellt. Abbildung (b) visualisiert die Streuung $2\sigma_\phi$ der Eulerwinkel für die einzelnen Probanden.

bestimmt (Collet et al., 2011). Die einfache Streuung σ_ϕ der Rotation berechnet sich aus dem gemittelten Rotationsfehler:

$$\sigma_\phi = \sqrt{\frac{1}{(K-1)} \sum_{j=1}^K (2 \cdot \arccos(\langle \mathbf{q}_j, \bar{\mathbf{q}} \rangle))^2} \quad (3.16)$$

Streuung der Koordinatentransformation $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$

Die Streuung der Transformation $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$ berechnet sich aus den Gleichungen (3.13-3.16). Die erste Zeile von Tabelle 3.5 zeigt die Werte der Streuung 2σ für die Translation und Rotation. Die Streuung beträgt 2,57 mm und 0,34 °.

Streuung der Koordinatentransformation $\mathbf{T}_n^{t \rightarrow h}$

Zur Analyse der Streuung der Transformation vom Koordinatensystems des Targets t in das des Kopfes h wird sowohl die Streuung der einzelnen Probanden n als auch die mittlere Streuung von allen Probanden bestimmt.

Streuung der Transformation der einzelnen Probanden. Für die einzelnen Probanden wird die Streuung der Translation und Rotation aus Gleichung (3.13) - (3.16) berechnet. Abbildung 3.16 zeigt die unterschiedlichen Streuungen der bestimmten Transformationen abhängig von den Probanden. Es ist eine geringe Abweichung der Werte sichtbar. Daraus folgt, dass es wichtiger ist viele Wiederholungen mit demselben Probanden durchzuführen als die Anzahl von Probanden zu erhöhen, um eine Abschätzung für den Mittelwert der Transformation zu erhalten.

Streuung der Transformation gemittelt über alle Probanden. Um eine Streuung gemittelt über alle Probanden zu erhalten wird im Folgenden die Streuung aus allen Messungen bestimmt. Dabei ist es notwendig, die Transformationen der einzelnen Probanden mit der mittleren Transformation des jeweiligen Probanden $\overline{\mathbf{T}}_n^{t \rightarrow h}$ zu normalisieren. Durch die Normalisierung werden die Gleichungen 3.15 und 3.16 wie folgt umformuliert:

$$\sigma_t = \sqrt{\frac{1}{(\sum K_n - 1)} \sum_{n=1}^{\#subj} \sum_{j=1}^N (\mathbf{t}_j - \bar{\mathbf{t}}_n)^2} \quad (3.17)$$

$$\sigma_\phi = \sqrt{\frac{1}{(\sum K_n - 1)} \sum_{n=1}^{\#subj} \sum_{j=1}^{K_n} (2 \cdot \arccos(\langle \mathbf{q}_j, \bar{\mathbf{q}}_n \rangle))^2} \quad (3.18)$$

Hierbei wird für jeden Probanden n der individuelle Mittelwert der Rotation $\bar{\mathbf{q}}_n$ beziehungsweise der Translation $\bar{\mathbf{t}}_n$ betrachtet. Es ist wichtig zu beachten, dass die mittlere Transformation auch bei exakter Messung von der Person abhängig ist, da jeder Proband das Target an einer anderen Stelle trägt. Die Ergebnisse der Streuung 2σ sind in der zweiten Zeile von Tabelle 3.5 dargestellt. Es ist evident, dass die Streuung dieser Transformation mit 5.77 mm und 3.04° die Gesamtstreuung dominiert.

Gesamtstreuung der Referenzmessung

Die Gesamtstreuung der Referenzmessung setzt sich aus den in den vorherigen Abschnitten diskutierten Streuungen zusammen.

Mit $\mathbf{T}_1 = \mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t}$ und $\mathbf{T}_2 = \mathbf{T}_n^{t \rightarrow h}$ ergibt sich:

$$\mathbf{T}_{ges} = \mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t} \cdot \mathbf{T}_n^{t \rightarrow h} = \mathbf{T}_1 \cdot \mathbf{T}_2 \quad (3.19)$$

Die experimentell erzeugten Messungen der Transformation \mathbf{T}_1 und \mathbf{T}_2 werden zur Bestimmung der Gesamtstreuung kombiniert. Die Messung der Transformation \mathbf{T}_1 wurde experimentell $K_1 = 38$ mal wiederholt, dadurch ergibt sich hierfür die Transformation \mathbf{T}_i mit $i \in \{1, \dots, K_1\}$. Die Transformation \mathbf{T}_2 wurde für $n = 4$ Probanden experimentell 18 bis 21 Mal erzeugt. Es ergeben sich für diese Transformation insgesamt $K_2 = 73$ Messungen, die im Folgenden als \mathbf{T}_j mit $j \in \{1, \dots, K_2\}$ bezeichnet werden. Für \mathbf{T}_{ges} ergeben sich die Messungen aus allen Kombinationen der gemessenen Transformationen \mathbf{T}_i und \mathbf{T}_j , insgesamt entstehen damit $K_1 * K_2 = 2774$ Kombinationen für die Analyse der Referenzmessung. Die zusammengesetzten Transformationen \mathbf{T}_l mit $l \in \{1, \dots, K_1 * K_2\}$ ergeben sich aus:

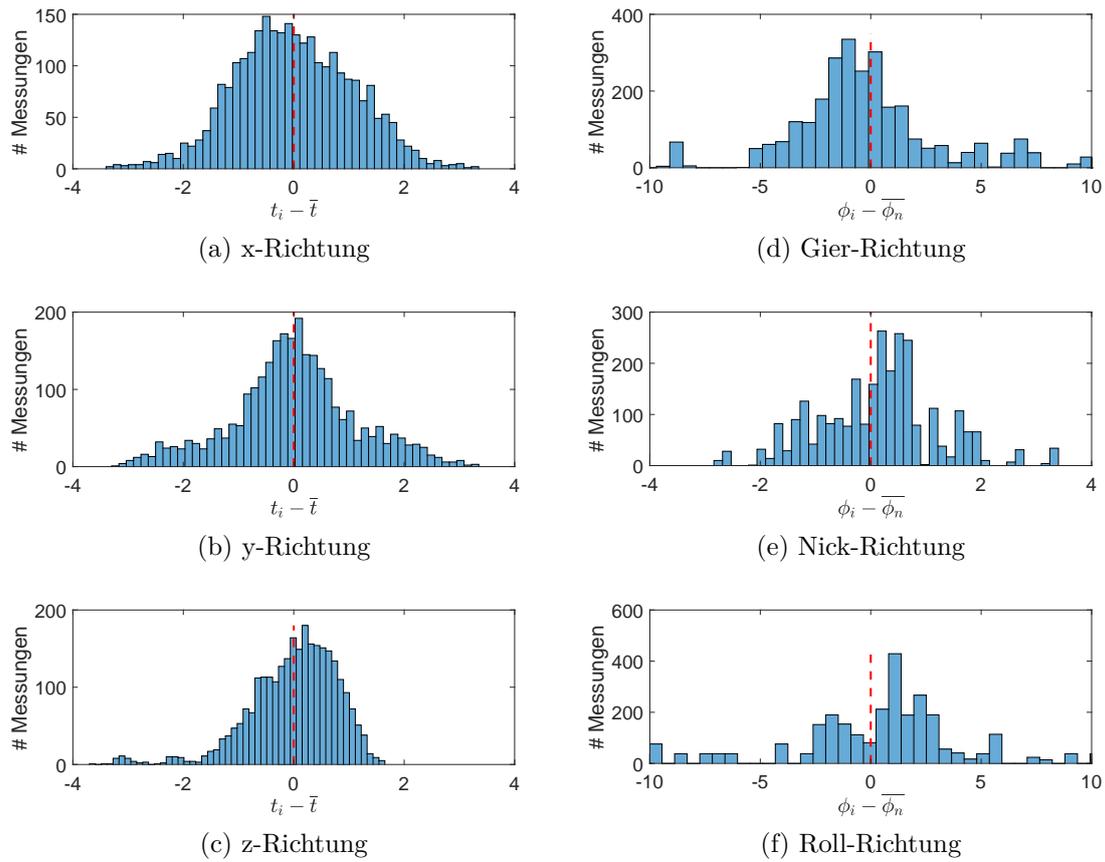


Abbildung 3.17.: Histogramme der Referenzmessung $\mathbf{T}_i^{k \rightarrow h}$ durch die Kombination der Transformation $\mathbf{T}^{k \rightarrow w} \cdot \mathbf{T}_i^{w \rightarrow t} \cdot \mathbf{T}_n^{t \rightarrow h}$. Verteilung der Translation in x-, y- und z- Richtung (linke Spalte), Verteilung der gemessenen Rotation in Gier-, Nick-, und Roll-Richtung (rechte Spalte).

$$\mathbf{T}_l = \mathbf{T}_i \cdot \mathbf{T}_j = \begin{pmatrix} \mathbf{R}_i & | & \mathbf{t}_i \\ 0 & 0 & 0 & | & 1 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{R}_j & | & \mathbf{t}_j \\ 0 & 0 & 0 & | & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_i \cdot \mathbf{R}_j & | & \mathbf{R}_i \cdot \mathbf{t}_j + \mathbf{t}_i \\ 0 & 0 & 0 & | & 1 \end{pmatrix} \quad (3.20)$$

In Abbildung 3.17 ist die Verteilung der zusammengesetzten Transformationen in Histogrammen für die x-, y- und z-Richtung sowie der Gier-, Nick- und Roll-Richtung visuell dargestellt.

Für die Streuung der Rotation ergibt sich:

$$\sigma_\phi = \sqrt{\frac{1}{(K_1 - 1) \cdot (K_2 - 1)} \sum_{j=1}^{K_2} \sum_{i=1}^{K_1} (2 \cdot \arccos(\langle \mathbf{q}_i \cdot \mathbf{q}_j, \overline{\mathbf{q}_i \cdot \mathbf{q}_j} \rangle))} \quad (3.21)$$

Hierbei entspricht \cdot der Multiplikation von Quaternionen. Dies ist die Hintereinanderausführung von zwei Rotationen.

	Streuung $2\sigma_\phi$ in $^\circ$	Streuung $2\sigma_t$ in mm
$T^{k \rightarrow w} \cdot T_i^{w \rightarrow t}$	0,34	2,57
$T_n^{t \rightarrow h}$	5,77	3,04
$T^{k \rightarrow w} \cdot T_i^{w \rightarrow t} \cdot T_n^{t \rightarrow h}$	5,74	3,49

Tabelle 3.5.: Streuung der Translationen und Rotationen. $T^{k \rightarrow w} \cdot T_i^{w \rightarrow t}$ und $T_n^{t \rightarrow h}$ sowie die Gesamttransformation $T^{k \rightarrow w} \cdot T_i^{w \rightarrow t} \cdot T_n^{t \rightarrow h}$. Die Messunsicherheit wurde mithilfe einer statistischen Analyse experimentell bestimmt, indem jede Messung unter gleichbleibenden Bedingungen wiederholt wurde.

Für die Streuung der Translation folgt:

$$\sigma_t = \sqrt{\frac{1}{(K_1 - 1) \cdot (K_2 - 1)} \sum_{j=1}^{K_2} \sum_{i=1}^{K_1} (t_l - \bar{t}_l)^2} \quad (3.22)$$

Zur Vereinfachung der Darstellung wurden die unterschiedlichen Personen n in Gleichung 3.21 und 3.22 nicht dargestellt. Bei der Berechnung wird für \bar{T}_j der Mittelwert der entsprechenden Person n für alle $j \in \{1, \dots, K_2\}$ verwendet. Tabelle 3.5 zeigt die Streuung 2σ der gesamten Referenzmessung in der letzten Zeile. Für die Translation und Rotation ergeben sich *statistische Fehler* von $\pm 3.49mm$ und $\pm 5.74^\circ$.

3.5.3. Fazit

In diesem Kapitel wurde die Referenzmessung charakterisiert, indem der *zufällige Fehler* der Transformationen bestimmt wurde, welcher die Streuung der Referenzmessung angibt. Die Transformation der Referenzmessung setzt sich aus drei einzelnen Transformationen zusammen. Bei der experimentellen Durchführung wurden die einzelnen Transformationen unter gleichbleibenden Bedingungen mehrmals bestimmt. Anschließend wurden diese Messungen zur statischen Analyse der Gesamtstreuung kombiniert. Das Kapitel beinhaltet sowohl den experimentellen Aufbau zur Bestimmung der einzelnen Messung als auch die statistische Analyse des *zufälligen Fehlers* der Referenzmessung.

Tabelle 3.5 zeigt den experimentell bestimmten Wert der einzelnen Transformationen $T_n^{t \rightarrow h}$ und $T^{k \rightarrow w} \cdot T_i^{w \rightarrow t}$. Insgesamt ergibt sich durch das Kombinieren aller Messungen der einzelnen Transformationen ein *zufälliger Fehler* von $\pm 5,74^\circ$ für die Rotation und $\pm 3,49 mm$ für die Translation der Referenzmessung $T^{k \rightarrow w} \cdot T_i^{w \rightarrow t} \cdot T_n^{t \rightarrow h}$. Es ist deutlich sichtbar, dass die Kalibrierung des Systems und die Lagebestimmung des Targets vom Motion Capture-System, welche durch die Transformationen $T^{k \rightarrow w} \cdot T_i^{w \rightarrow t}$ beschrieben werden mit $\pm 0,34^\circ$ und $\pm 2,57 mm$, nur zu einer geringen Streuung führen. Dagegen ist die Transformation $T_n^{t \rightarrow h}$ vom Koordinatensystem des Targets t zu dem des Kopfes k mit einer Streuung von $\pm 5,77^\circ$ und $\pm 3,04 mm$ für den *zufälligen Fehler* ausschlaggebend. Einen großen Einfluss auf diese Streuung

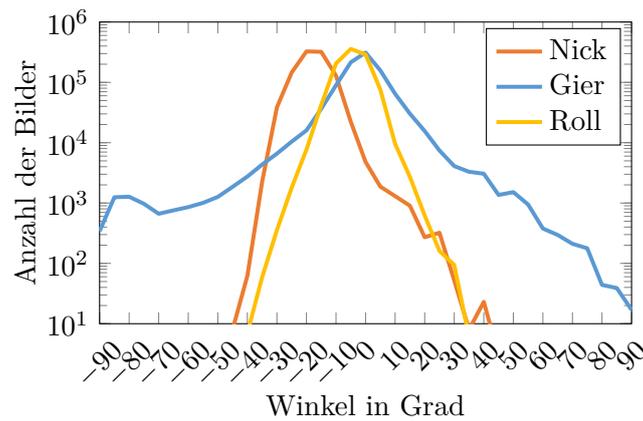


Abbildung 3.18.: Histogramm der Rotationswinkel in Grad um die x-Achse (Nickbewegung), die y-Achse (Gierbewegung) und die z-Achse (Rollbewegung). Aus Schwarz et al. (2017) © 2017 IEEE.

hat die Bestimmung der Gesichtslanmarken. Bei der Markierung der Gesichtslanmarken kommt es zu Ungenauigkeiten. Die Positionen der Lanmarken beeinflussen direkt die Berechnung der Transformation des Target-Koordinatensystems zum Kopf-Koordinatensystem, dadurch kommt es zu einer Streuung dieser Transformation.

3.6. Beschreibung der Stichprobe

Im folgenden Abschnitt werden die Eigenschaften der Probanden beschrieben, die an der Datenaufnahme teilgenommen haben. Zusätzlich wird die Verteilung der aufgenommenen Daten analysiert, hierbei wird auf die Anzahl der Bilder abhängig von den Rotationswinkeln und von den unterschiedlichen Verdeckungen eingegangen.

weiblich / männlich	Brillen	Sonnenbrillen	Alter	Körpergröße	ethnische Erscheinungsbilder
4/16	6/14 ^a	16/4 ^a	21-52 Jahre	1,6-1,95 m	Lateinamerika/Asien/Europa/Afrika

^a vorhanden/nicht vorhanden

Tabelle 3.6.: Eigenschaften der Probanden, die an der Datenaufnahme teilgenommen haben.

Tabelle 3.6 zeigt die Eigenschaften der Probanden, die an der Studie teilgenommen haben. Bei der Studie wurden insgesamt Fahrten von 20 Probanden aufgenommen, davon 16 männliche Fahrer und 4 weibliche Fahrerinnen. Während der Fahrten trugen 6 der Probanden mindestens teilweise eine Brille, während 16 der Probanden einen Teil der Fahrt eine Sonnenbrille aufhatten. Bei den Brillen sowie Sonnenbrillen handelt es sich um jeweils unterschiedliche Modelle. Das Alter der Teilnehmer liegt zwischen 21 und 52 Jahren und die Größe zwischen 1,6 Meter und 1,95 Meter.

■ Brillen ■ Sonnenbrillen ■ Ohne ■ Mit Verdeckung ■ Ohne Verdeckung

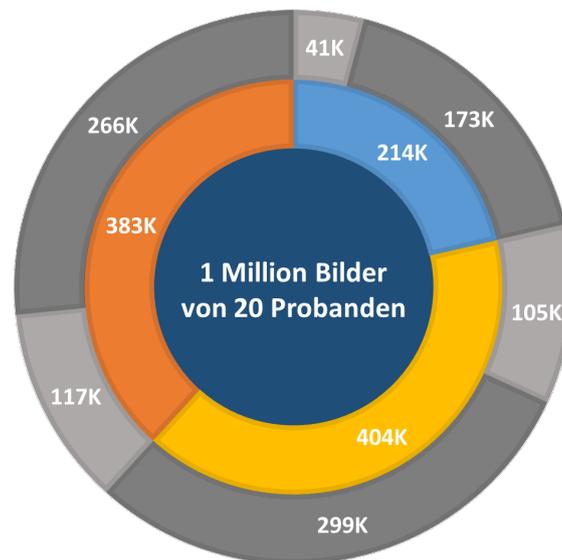


Abbildung 3.19.: Die Zahlen des inneren Ringes zeigen die Anzahl der Bilder des Datensatzes mit Personen, welche Brillen, Sonnenbrillen oder keines von beiden ('Ohne') tragen. Der äußere Ring beschreibt die Verdeckungen zusätzlich zu den im inneren Ring beschriebenen Brillen. Aus Schwarz et al. (2017) © 2017 IEEE.

Die Wahl der Probanden wurde so getroffen, dass möglichst viele unterschiedliche ethnische Erscheinungsbilder vorhanden sind.

In Abbildung 3.18 sind die Histogramme der einzelnen Rotationswinkel über alle Fahrten aufgezeichnet. Durch die Datenaufnahme während realer Autofahrten bei denen der Fahrer den Kopf die meiste Zeit frontal ausrichtet, ist ein Häufungspunkt bei allen Rotationen zu erkennen. Während die Gier- und Rollbewegung ihr Maximum bei 0 Grad haben, liegt es bei der Nickbewegung bei -20 Grad. Die Ursache dieser Verschiebung um 20 Grad ist die Ausrichtung der Kamera. Dadurch ist die Nickbewegung aus Sicht der Kamera bei frontaler Kopforientierung -20 Grad. Während die Roll- und Nickbewegung zwischen -45° und $+40^\circ$ ist, bewegen die Probanden ihren Kopf um die y-Achse nach links und rechts mit mehr als -90° und über $+90^\circ$ Grad. Diese starke Gierbewegung findet hauptsächlich bei Schulterblicken und Spurwechseln statt.

Die Tortengrafik in Abbildung 3.19 beschreibt die Aufteilung der Bilder des Datensatzes entsprechend der unterschiedlichen Verdeckungen durch Brillen, Sonnenbrillen sowie zusätzliche Verdeckungen. Insgesamt beinhaltet der Datensatz mehr als eine Million Infrarot- und Tiefenbilder von 20 unterschiedlichen Probanden. Mehr als ein Drittel der Daten zeigen Fahrer mit Sonnenbrillen, während bei etwa 20 Prozent der Gesichter Brillen vorhanden sind. Zusätzlich werden bei etwa 30 Prozent der Daten weitere Verdeckungen, beispielsweise durch das Lenkrad oder die Hand, notiert.

3.7. DriveAHead-Datensatz

Der in diesem Kapitel vorgestellte *DriveAHead*-Datensatz ist als Datensatz frei verfügbar unter <https://cvhci.anthropomatik.kit.edu/data/DriveAHead/>. Die Personen 6 bis 20 dienen als Trainingsdatensatz auf denen neue Algorithmen trainiert werden können, wobei Person 6 und 7 als Validierungsdaten dienen. Der Testdatensatz besteht aus Person 1 bis 5. Seit der Veröffentlichung im Juli 2017 haben sich bereits 5 unterschiedliche Gruppen aus weltweit verteilten Universitäten zur Evaluation auf dem Datensatz registriert.

Kapitel 4

Regressions-basierte Bestimmung der Orientierung und Position des Kopfes

Dieses Kapitel präsentiert den im Rahmen dieser Arbeit entstandenen regressionsbasierten Algorithmus zur Bestimmung der Kopfpose aus Tiefendaten. Das Kapitel basiert auf der Veröffentlichung von [Schwarz et al. \(2016\)](#), in der das Verfahren und die Ergebnisse bereits vorgestellt wurden.

Im Fahrzeugkontext sind die Rechenkapazitäten limitiert. Dadurch spielt die Effizienz des Algorithmus eine ausschlaggebende Rolle. Kapitel [2.4.2](#) gibt einen Überblick von bereits vorhandenen Algorithmen, die aus Tiefendaten die Kopfpose bestimmen. Insbesondere Verfahren mit Entscheidungsbäumen, wie zum Beispiel der Ansatz von [Fanelli et al. \(2013\)](#), zeigten vielversprechende Ergebnisse und eine schnelle Laufzeit. Im Vergleich zu diesem Verfahren verwendet die im Folgenden vorgestellte Methode eine intelligente Kombination aus Entscheidungsbäumen und einer linearen globalen Regression. Dadurch erreicht sie zum Einen für die Translation des Kopfes genauere Ergebnisse und zum Anderen wird der Rechenaufwand minimiert, indem die Entscheidungsbäume an weniger Bildpunkten ausgewertet werden.

Im folgenden Kapitel wird ein neuartiger effizienter Algorithmus zur Bestimmung der Position und Orientierung des Kopfes aus Tiefendaten vorgestellt. Der Algorithmus zeigt bei weniger Rechenaufwand Ergebnisse, die vergleichbar zum Stand der Forschung sind. Das hier beschriebene Verfahren ist inspiriert von der Veröffentlichung von [Ren et al. \(2014\)](#), siehe [Abbildung 2.3](#). Durch die Kombination von Entscheidungsbäumen und einer linearen Regressionsmatrix ist der Algorithmus sehr effizient, deshalb wird die Methode *Highly efficient Head Orientation and Position*-Verfahren (*HeHOP*) genannt.

Im ersten Abschnitt wird der Aufbau des Verfahrens genauer beschrieben, während der zweite Abschnitt die Ergebnisse des Algorithmus auf dem öffentlich zur Verfügung stehenden Kopfposendatensatz BIWI ([Fanelli et al., 2013](#)) aufzeigt. Das

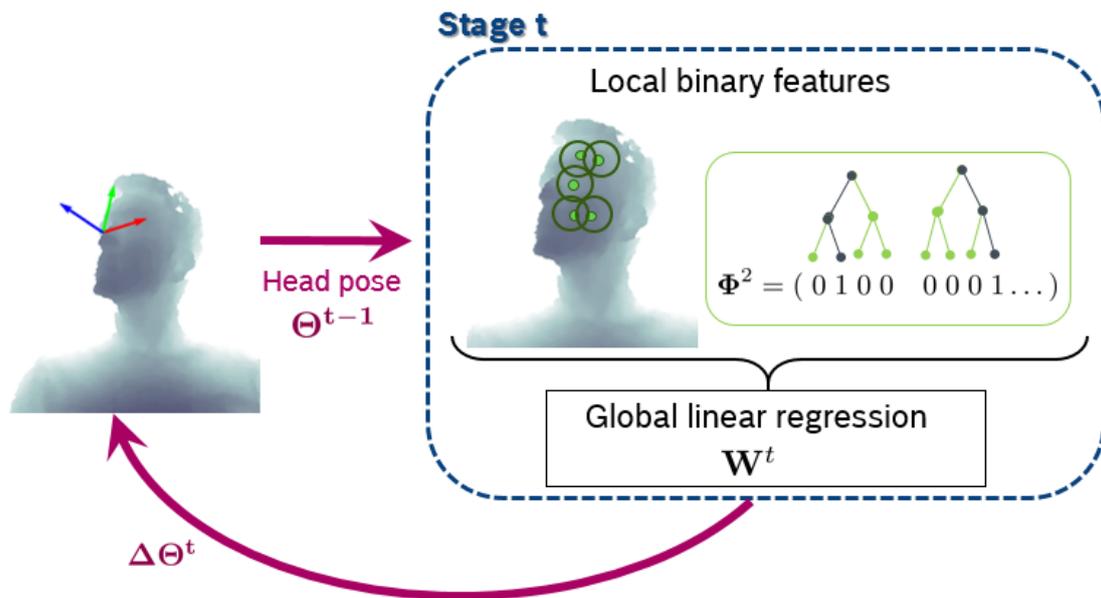


Abbildung 4.1.: Die Abbildung zeigt eine Iteration des Algorithmus zur Bestimmung der Kopfposition und -orientierung. In jedem Schritt wird die vorherige Kopfpose Θ^{t-1} aktualisiert. Zuerst werden lokale Regionen mit dieser Rotation und Translation rückprojiziert und mithilfe von Entscheidungsbäumen an diesen Positionen binäre Features gewonnen. Mit einer linearen Regression berechnet sich anschließend die Aktualisierung $\Delta\Theta^t$ der Position und Orientierung des Kopfes.

Verfahren wird in Kapitel 6 auf dem im Rahmen dieser Arbeit entwickelten Datensatz zur Bestimmung der Kopfpose im Fahrzeuginnenraum analysiert. Hierbei wird die Genauigkeit der Methode unter den während realer Autofahrten auftretenden Einflüssen untersucht.

4.1. Ziel

Das Ziel des Verfahrens ist die Bestimmung der 3D-Position des Kopfes und der Rotation aus einem einzelnen Tiefenbild. Die Bestimmung erfolgt stufenweise, wobei in jeder Stufe die vorherige Position und Orientierung des Kopfes in Richtung der finalen Kopfpose aktualisiert wird.

In jeder Stufe t werden zuerst lokale binäre Features ϕ^t berechnet und anschließend eine globale lineare Regression \mathbf{W}^t darauf angewendet, um die Position und Orientierung des Kopfes zu aktualisieren. Sowohl die Entscheidungsbäume, welche die lokalen binären Features berechnen, als auch die globale Regression benötigen eine Lernphase, in der die Parameter bestimmt werden. In Abschnitt 4.2 wird das Training beschrieben, während in 4.3 die gelernten Entscheidungsbäume sowie die globale Regression auf die unbekanntenen Testdaten angewendet werden.

4.2. Trainingsphase

Die Lernphase ist überwacht, das bedeutet, für die Trainingsdaten gibt es Referenzdaten, die zum Training der Parameter verwendet werden. Jedes Trainingselement besteht aus einem Tiefenbild D_i und einem zugehörigen Referenzpaar $\hat{\Theta}_i = \{\hat{l}_i, \hat{\theta}_i\}$, bestehend aus der Position des Kopfes $\hat{l}_i = \{\hat{x}, \hat{y}, \hat{z}\}$ und der Orientierung des Kopfes in Quaternionen $\hat{\theta}_i = \{\hat{q}_1, \hat{q}_2, \hat{q}_3, \hat{q}_4\}$.

Bevor die Kopfpose aus den Tiefendaten berechnet wird, bestimmt ein Gesichtserkennungsalgorithmus in der Vorverarbeitung die Region des Kopfes. Anschließend bestimmt das hier vorgestellte Verfahren aus der Gesichtsregion die Position und Orientierung des Kopfes. Um diese zu berechnen, werden in der Lernphase für jede Stufe t mehrere Regressionsbäume aufgebaut, zur Bestimmung des lokalen binären Eigenschafts-Vektors Φ^t . Die Parameter der globalen Regressionsmatrix W^t werden für jede Stufe t trainiert.

4.2.1. Erzeugung der lokalen binären Eigenschaftsvektoren

Die lokalen binären Eigenschaftsvektoren Φ^t sind die Ausgabe von mehreren Entscheidungsbäumen, genannt Entscheidungswald. Für jede Stufe t der Kopfposenbestimmung werden in der Trainingsphase diese Entscheidungswälder aufgestellt, die anschließend in der Testphase auf lokale Regionen innerhalb des Gesichts angewendet werden, um die binären Eigenschafts-Vektoren zu erhalten. Dieser Wald wird nach dem in Ren et al. (2014) vorgestellten Prinzip erzeugt, wobei hier Tiefendaten D_i anstatt Grauwerte verwendet werden. Die Tiefenwerte D_i durchlaufen die Bäume bestehend aus internen Knoten und Blattknoten. An den internen Knoten spaltet eine Funktion die Eingangsdaten basierend auf ihren Nachbarwerten auf. Anhand der Blattknoten wird ein binärer Vektor generiert, indem die Stelle des erreichten Blattes auf 1 gesetzt wird und die restlichen auf 0.

Das Ziel der Bäume ist es, die lokalen Positionen anhand ihrer Tiefenwerte in der Nachbarschaft zu klassifizieren. Um dies zu erreichen, werden während des Trainings für jeden Baum optimale Spaltungsfunktionen, bestehend aus einem Grenzwert und der zu untersuchenden Nachbarregionen, definiert. Die Spaltungsfunktionen werden während des Trainings so gewählt, dass die Positionen abhängig von den Abständen zu den Referenzpositionen aufgeteilt werden. Damit die Spaltungsfunktionen die Daten entsprechend der Abstände zu den Referenzpositionen teilen, sind für das Training für jedes Bild i die 2D-Referenzpositionen $\hat{S}_i \in \mathbb{R}^{2 \times k}$ der k lokalen Regionen erforderlich. Die 2D-Referenzpositionen $\hat{S}_i \in \mathbb{R}^{2 \times k}$ sind Positionen innerhalb der Gesichtsregionen, die aus 3D-Gesichtslandmarken eines starren Modellgesichts mithilfe der bekannten Referenz der Kopforientierung und -position zurückprojiziert werden. Die Entscheidungsbäume werden an den 2D-Positionen $S_i^{t-1} \in \mathbb{R}^{2 \times k}$ aus der vorherigen Stufe $t - 1$ angewendet. Zur Bestimmung dieser lokalen Regionen

$\mathbf{S}_i^{t-1} \in \mathbb{R}^{k \times 2}$ der vorherigen Stufe werden dieselben Regionen mit der vom Algorithmus geschätzten Translation und Orientierung des Kopfes aus der vorherigen Stufe $t - 1$ zurückprojiziert. Mit den bekannten Referenzpositionen der Regionen werden während des Trainings Spaltungsfunktionen gewählt, die eine Klassifikation der Daten anhand der Tiefendaten ermöglichen.

Die aufgebauten Entscheidungswälder, angewendet auf die Positionen der vorherigen Stufe \mathbf{S}_i^{t-1} , ergeben den binären Featurevektor Φ^t . Da für jede lokale Region k die einzelnen Regressionswälder keine Information von den übrigen Regionen verwenden, sind die Regressionsbäume voneinander unabhängig. Die Werte des binären Vektors werden an den Stellen auf 1 gesetzt, an denen ein Blatt erreicht wird, die anderen Werte sind 0. Dadurch handelt sich um einen dünnbesetzten Vektor, bei dem pro Baum ein Wert auf 1 ist.

4.2.2. Bestimmung der linearen Regressionsmatrix

Die globale lineare Regressionmatrix \mathbf{W}^t bestimmt den Vektor $\Delta\Theta_i^t$, der den Abstand in Richtung der finalen Position und Orientierung des Kopfes angibt. In der Trainingsphase werden die Parameter der Matrix \mathbf{W}^t optimiert, indem der binäre Vektor aus der vorherigen Stufe und das Residuum $\Delta\hat{\Theta}_{i,s}^t$ in Richtung der Referenzposition und -orientierung des Kopfes verwendet wird. Hierbei wird eine lineare Regression verwendet. Mithilfe der folgenden Formel wird die optimale Regressionsmatrix \mathbf{W}^t bestimmt:

$$\min_{\mathbf{W}^t} \sum_{i=1}^N \|\Delta\hat{\Theta}_i^t - \mathbf{W}^t \Phi_i^t\|_2^2 + \lambda \|\mathbf{W}^t\|_2^2 \quad (4.1)$$

Wobei der erste Teil sicherstellt, dass das Ergebnis nach der linearen Regression möglichst nahe am Referenzwert ist, während der zweite Teil mit einer L2-Regularisierung die Überanpassung einzelner Matrixelemente verhindert, indem der Betrag der Werte mit einem Gewichtungsfaktor λ berücksichtigt wird. Die für Gleichung 4.1 erforderlichen Residuen zu den Referenzwerten $\Delta\hat{\Theta}_i^t$ berechnen sich wie folgt:

$$\Delta\hat{\Theta}_i^t = \begin{pmatrix} \Delta\hat{\mathbf{l}}_i^t \\ \Delta\hat{\boldsymbol{\theta}}_i^t \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{l}}_i - \mathbf{l}_i^{t-1} \\ Q(\hat{\mathbf{R}}_i(\mathbf{R}_i^{t-1})^{-1}) \end{pmatrix} \quad (4.2)$$

$\hat{\mathbf{R}}_i$ ist die Rotationsmatrix, welche aus dem Quaternionenvektor $\hat{\boldsymbol{\theta}}_i$ bestimmt ist und \mathbf{R}_i^{t-1} die Rotationsmatrix des Quaternionenvektors $\boldsymbol{\theta}_i^{t-1}$ zum vorherigen Zeitpunkt $t - 1$. Die Funktion $Q()$ definiert die Funktion, welche eine Rotationsmatrix in den eindeutigen Quaternionenvektor umrechnet.

Der Vorteil der Verwendung von Quaternionen ist die kompakte Darstellung der Orientierung durch einen vierdimensionalen Vektor. Außerdem kann eine Rotation durch mehrere Kombinationen von Euler Winkeln beschrieben werden, dadurch

kann es zu einem Gimbal Lock kommen. Im Gegensatz dazu ist die Repräsentation einer Rotation durch Quaternionen eindeutig bestimmt (Dam et al., 1998). Diese Eigenschaft der Quaternionen-Representation ist für die Gleichung 4.1 essentiell.

4.2.3. Stufenweise Aktualisierung der Kopfpose

Nach jeder Stufe t findet für jedes Bild i eine Aktualisierung der Kopfposition und -orientierung statt. Die Kopfpose Θ_i^t ergibt sich aus:

$$\Theta_i^t = f(\Theta_i^{t-1}, \Delta\Theta_i^t) = \begin{pmatrix} \mathbf{l}_i^{t-1} + \Delta\mathbf{l}_i \\ Q(\Delta\mathbf{R}_i^t \mathbf{R}_i^{t-1}) \end{pmatrix} \quad (4.3)$$

mit

$$\Delta\Theta_i^t = \begin{pmatrix} \Delta\mathbf{l}_i^t \\ \Delta\boldsymbol{\theta}_i^t \end{pmatrix} = \mathbf{W}^t \Phi_i^t \quad (4.4)$$

Die Rotationsmatrix $\Delta\mathbf{R}_i^t$ berechnet sich aus dem Quaternionenvektor $\Delta\boldsymbol{\theta}_i^t$ zum Zeitpunkt t und \mathbf{R}_i^{t-1} aus dem Quaternionenvektor $\boldsymbol{\theta}_i^{t-1}$ der vorherigen Stufe zum Zeitpunkt $t - 1$.

4.3. Testphase: Bestimmung der Kopfpose

Das Ziel des Verfahrens ist die Bestimmung der Position und Orientierung des Kopfes $\Theta_i = \{\mathbf{l}_i, \boldsymbol{\theta}_i\}$ aus einem unbekanntem Tiefenbild D_i . Der Algorithmus berechnet stufenweise eine Aktualisierung $\Delta\Theta_i^t$ in Richtung der finalen Orientierung und Position. Beginnend mit der frontalen Rotation, welche der Einheitsmatrix entspricht, und der Kopfposition an dem Mittelpunkt der Kopfreion, besteht jede Stufe aus einem lokalen und einem globalen Schritt. Mit der Rotation und Translation der vorherigen Stufe werden die Gesichtslanmarken des im Training verwendeten 3D-Modellgesichts rück-projiziert. An diesen Positionen bestimmen im lokalen Schritt Entscheidungsbäume binäre Eigenschafts-Vektoren Φ_i^t aus den lokalen Nachbarschaftsregionen. Anschließend ermittelt die globale lineare Regressionsmatrix \mathbf{W}^t die Aktualisierung der Orientierung und Position des Kopfes $\Delta\Theta_i^t = \{\Delta\mathbf{l}_i^t, \Delta\boldsymbol{\theta}_i^t\}$ mithilfe von Gleichung 4.4. Nach jeder Stufe aktualisiert sich die Kopfpose mit $\Theta_i^t = f(\Theta_i^{t-1}, \Delta\Theta_i^t)$ aus Gleichung 4.3.

4.4. Evaluation auf Labordaten

Dieser Abschnitt beschreibt die Evaluation der Laufzeit und Genauigkeit des vorgestellten Verfahrens mit unterschiedlichen Parameterkonfigurationen. Es findet ein Vergleich zu aktuellen Ansätzen, die den Stand der Technik darstellen, statt. Der öffentlich zur Verfügung stehende BIWI-Kopfposen-Datensatz (Fanelli et al., 2013)

ermöglicht diese Validierung. Hierbei sind zum Einen der mittlere Fehler und zum Anderen die Varianz der Position des Kopfes, repräsentiert durch die Nasenposition, und der Orientierung des Kopfes von Interesse. Die Metrik zur Bestimmung des Fehlers der Kopfposition ist die euklidische Distanz zwischen der geschätzten Position und der Referenzposition. Der Fehler der Orientierung des Kopfes bestimmt sich, wie in [Papazov et al. \(2015\)](#) beschrieben, aus dem Fehler zwischen den Orientierungsvektoren. Zur Validierung wurde ein unoptimiertes Programm basierend auf C++ auf einem CPU-Kern mit 2,80 GHz ausgeführt.

4.4.1. BIWI Kinect-Datensatz

Der BIWI-Datensatz ([Fanelli et al., 2013](#)) zur Validierung der Kopfpose beinhaltet mehr als 15 Tausend Farb- und Tiefenbilder von 20 unterschiedlichen Personen. Davon sind sechs der Probanden weiblich und 14 Probanden männlich. Sowohl die Farbbilder als auch die Tiefendaten wurden mit dem Kinect-Sensor xbox 360 von Microsoft aufgenommen. Die Auflösung der Tiefenbilder beträgt 640×480 Pixel, wobei die Kopfregion im Durchschnitt circa 90×110 Pixel umfasst. Die Kopforientierung beträgt $\pm 75^\circ$ bei der Gierbewegung und $\pm 60^\circ$ bei der Nickbewegung. Für jedes Bild sind die Referenzwerte der Nasenposition und der Kopforientierung vorhanden.

4.4.2. Referenzpositionen innerhalb des Gesichtes

Das Training der Parameter der Entscheidungsbäume basiert auf k Referenzpositionen im Gesicht. Diese Referenzpositionen erschließen sich aus der Rückprojektion von $k = 68$ Gesichtslanmarken eines 3D-Durchschnittskopfes mit der Referenzorientierung und -position des BIWI-Datensatzes, wodurch sich für jedes Bild i 68 Referenzpositionen ergeben.

Sowohl für die Trainings- als auch für die Testphase ist eine Initialposition und -rotation des Kopfes notwendig. Hierfür bestimmt der Gesichtserkennungs-Algorithmus von [Viola und Jones \(2004\)](#) die Region des Gesichts auf dem Farbbild. Eine Projektion der Farbwerte auf die Tiefendaten ist erforderlich, da die Pixel der beiden Daten nicht exakt übereinstimmen. Die Initialposition des Kopfes ist definiert als der Mittelwert der 3D-Voxels innerhalb der Gesichtsregion, während die Initialrotation die frontale Orientierung mit allen Winkeln auf null ist, welche der Einheitsmatrix entspricht. Für den Fall, dass der Gesichtserkennungsalgorithmus fehlschlägt, wird die Gesichtsregion des vorherigen Bildes verwendet.

4.4.3. Optimierung der Parameter

In der Trainingsphase der Entscheidungsbäume gibt es Grundparameter, die im Vorhinein festgelegt werden müssen. Hierzu gehören die Anzahl der Entscheidungsbäu-

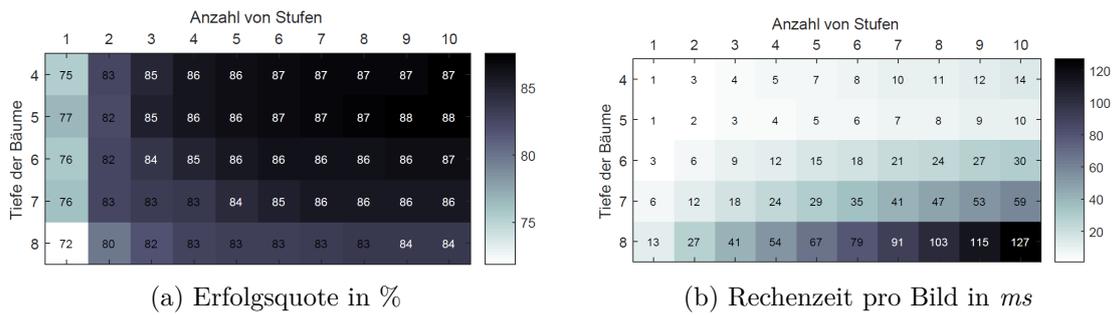


Abbildung 4.2.: Ergebnisse des Algorithmus bei Erhöhung der Anzahl der Stufen und abhängig von der Tiefe der Bäume. (a) zeigt die Erfolgsquote. (b) die Rechenzeit pro Bild in ms. Die Anzahl der Bäume wurde konstant auf 11 gesetzt.

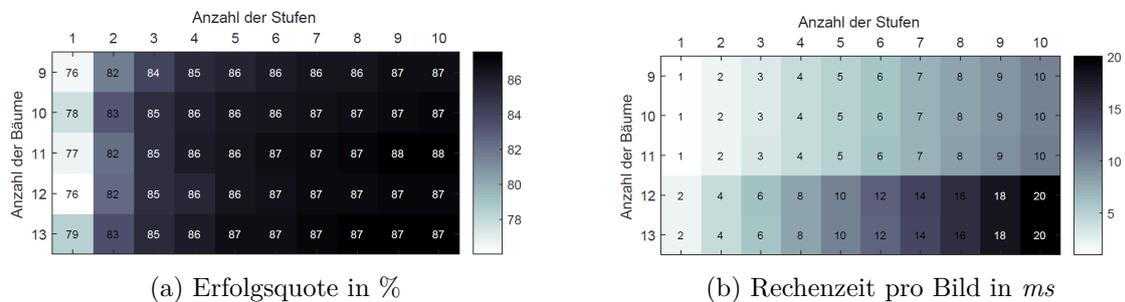


Abbildung 4.3.: Ergebnisse des Algorithmus bei Erhöhung der Anzahl der Stufen und abhängig von der Anzahl der der Bäume. (a) zeigt die Erfolgsquote. (b) die Rechenzeit pro Bild in ms. Die Tiefe der Bäume wurde konstant auf 5 gesetzt. Aus Schwarz et al. (2016) © 2016 IEEE.

me, die Tiefe der Entscheidungsbäume und die Anzahl der Stufen des Verfahrens. Um die optimalen Werte dieser Parameter zu finden, wird ein Validierungsexperiment wie in Fanelli et al. (2011a) durchgeführt, bei dem auf Grundlage der Daten von Person 1 und 12 validiert wird während der Algorithmus auf Grundlage der Daten der restlichen Personen trainiert wird. Bei jedem Bild wird die Erkennung der Orientierung und Position des Kopfes als erfolgreich klassifiziert entsprechend der Definition von Fanelli et al. (2013), wenn der Rotationsfehler kleiner als 15 Grad ist und der Translationsfehler unter 20 Millimetern liegt. Von allen Testdaten spiegelt die Erfolgsquote den prozentualen Wert der erfolgreichen Detektionen wider. Im folgenden Abschnitt wird das Validierungsexperiment zur Bestimmung der optimalen Anzahl der Entscheidungsbäume, der Tiefe der Entscheidungsbäume und der Anzahl der Stufen beschrieben.

Zuerst wird abhängig von der Tiefe der Bäume und der Anzahl der Bäume mit variierender Stufenanzahl des Verfahrens die Erfolgsquote und die Rechenzeit des Algorithmus ausgewertet. Die Abbildungen in 4.2 zeigen die Erfolgsquote und die Geschwindigkeit des Verfahrens mit variierender Tiefe der Bäume und Anzahl der Stufen bei gleichbleibender Anzahl von Bäumen. Die Abbildung zeigt, dass der Algo-

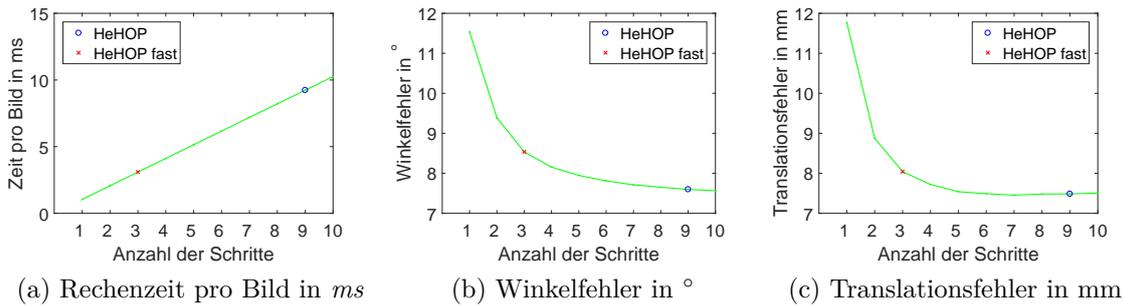


Abbildung 4.4.: Abhängig von der Anzahl der Stufen zeigt (a) die Rechenzeit pro Bild, (b) den mittleren Winkelfehler in $^{\circ}$ und (c) den mittleren Fehler der Translation in Millimeter. Der *blaue* Kreis und das *rote* Kreuz zeigen die Ergebnisse des Validierungs-Experimentes für die Algorithmen *HeHOP* und *HeHOP fast*. Aus Schwarz et al. (2016) © 2016 IEEE.

rithmus bei zu tiefen Bäumen die Parameter überanpasst, wodurch die Erfolgsquote sich verringert, während mehr Stufen die Erfolgsquote des Verfahrens kontinuierlich verbessern. Im Gegensatz dazu erhöht sich die Rechenzeit durch mehr Stufen linear, während tiefere Bäume die Rechenzeit quadratisch steigern. In Grafik 4.3 ist die Performance des Algorithmus abhängig von der Anzahl der Bäume sowie der Anzahl der Stufen dargestellt. Eine höhere Anzahl an Entscheidungsbäume verbessert das Ergebnis minimal, wobei es durch zu viele Bäume zu einer Überanpassung kommen kann, hierbei erhöht sich die Rechenzeit linear. Während die Erhöhung der Tiefe der einzelnen Entscheidungsbäume schnell zu einer Überanpassung der Parameter führt, verbessert die Anzahl der Bäume die Erfolgsquote bis es zu einer Überanpassung der Parameter kommt, gleichzeitig führt die Erhöhung beider Werte zu einem Anstieg der Rechenzeit.

Zur genaueren Analyse der Stufenanzahl wird abhängig von der Anzahl der Stufen die Rechenzeit pro Bild, der Winkelfehler und der Translationsfehler quantitativ untersucht. Die Ergebnisse sind grafisch in Abbildung 4.4 visualisiert, wobei (a) die Rechenzeit pro Bild, (b) die Winkelfehler und (c) die Translationsfehler abhängig von der Anzahl der Stufen zeigt. Während die Rechenzeit linear mit der Erhöhung der Stufenanzahl ansteigt, verringert sich der Winkel- und Translationsfehler mindestens quadratisch. Daraus folgt: Bei der Durchführung von mehr Stufen wird die Rechenzeit linear erhöht, allerdings sinkt der Winkel- und Translationsfehler der Kopfpose quadratisch.

Aus den durchgeführten Validierungsexperimenten werden Werte für die Anzahl der Stufen, die Tiefe der Bäume und die Anzahl der Bäume festgelegt. Insgesamt folgt, aus den Grafiken 4.2, 4.3 und 4.4, dass mit mehr Stufen die Erfolgsquote erhöht und damit der Winkel- und Rotationsfehler minimiert wird, allerdings die Rechenzeit sich ebenfalls erhöht. Deshalb werden im Folgenden zwei Versionen des Verfahrens betrachtet. Ein schnelles Verfahren, genannt *HeHOP fast*, bei dem es nur $T = 3$ Stufen gibt, und ein langsames *HeHOP*, dafür jedoch genaueres mit $T = 9$ Stufen. Bei beiden Versionen besteht der Entscheidungswald aus 11 Bäumen für jede lokale

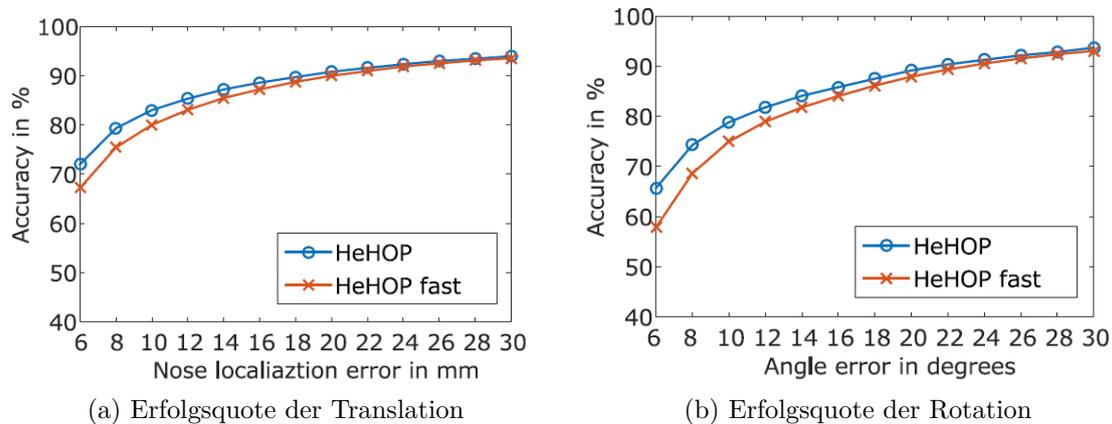


Abbildung 4.5.: Erfolgsquote in %. (a) der Rotation von *HeHOP* und *HeHOP fast* (b) der Translation von *HeHOP* und *HeHOP fast*. Aus Schwarz et al. (2016) © 2016 IEEE.

Region k und die Baumhöhe beträgt 5. In Grafik 4.4 sind die Resultate dieser Verfahren ausgewertet auf den Validierungsdaten für die Rechenzeit, Winkelfehler und Translationsfehler mit einem roten Kreuz für *HeHOP fast* und einem blauen Kreis für *HeHOP* gekennzeichnet. Im nächsten Abschnitt werden die beiden Varianten des Algorithmus detaillierter analysiert.

4.4.4. Ergebnisse der 4-fachen Kreuzvalidierung

Zur Evaluation des Verfahrens wird eine 4-fache Kreuzvalidierung durchgeführt. Für diese Methode wird der komplette Datensatz in vier Teile geteilt, das bedeutet in jedem Teil befinden sich die Daten von 5 Personen. Anschließend werden vier Evaluationen durchgeführt, wobei immer auf einem Teil die Ergebnisse berechnet werden während das Verfahren auf den restlichen 3 Teilen trainiert wird. Am Ende berechnet sich das Ergebnis aus dem Mittelwert der vier Validierungen.

Für die beiden Varianten *HeHOP fast* und *HeHOP* des in diesem Kapitel vorgestellten regressions-basierten Verfahrens wird die Erfolgsquote bei Veränderung des Grenzwertes für den Translations- und Winkelfehler untersucht. Abbildung 4.5 visualisiert die Erfolgsquote in Prozent der beiden Versionen des Algorithmus. In Abbildung 4.5 (a) ist der Prozentsatz der Daten dargestellt deren Translationsfehler kleiner ist als der auf der x -Achse angegebene Grenzwert. Abbildung 4.5 (b) zeigt den Prozentsatz der Daten, die einen bestimmten Rotationsfehler nicht überschreiten. Sowohl für die Translation als auch für die Rotation ist deutlich zu erkennen, dass *HeHOP* eine höhere Erfolgsquote erreicht als *HeHOP fast*.

Die Ergebnisse von zahlreichen Algorithmen auf dem BIWI-Datensatz (Fanelli et al., 2013) wurden veröffentlicht. Im Folgenden handelt es sich bei allen Verfahren um *frame-by-frame*-Algorithmen, welche die Kopfpose für jedes Bild unabhängig bestimmen, im Gegensatz zu *tracking*-Verfahren, die basierend auf dem vorherigen

	Position in mm	Gier in °	Nick in °	Roll in °	Verfügbarkeit in %	Orientierung in °	Zeit in ms
Papazov et al. (2015) ^a $\Delta = 200$	8,4 ± 22,2	3,0 ± 9,6	2,5 ± 7,4	3,8 ± 16,0	100	3,9 ± 10,3	75,9
Papazov et al. (2015) ^a $\Delta = 100$	10,6 ± 28,1	4,3 ± 14,3	3,2 ± 8,8	5,4 ± 20,8	100	5,4 ± 14,3	37,9
Riegler et al. (2014) ^b	8,1 ± 5,3	3,8 ± 3,7	6,7 ± 6,6	4,3 ± 4,9	99	9,8 ± 8,0	50
Fanelli et al. (2013) , Schrittweite 5	12,2 ± 22,8	3,8 ± 6,5	3,5 ± 5,8	5,4 ± 6,0	93,4	5,9 ± 8,1	44,7
Fanelli et al. (2013) , Schrittweite 15	13,4 ± 26,9	4,2 ± 7,8	3,8 ± 6,4	5,5 ± 6,2	93,5	6,4 ± 9,4	10,7
HeHOP	9,6 ± 27,5	5,1 ± 9,5	3,9 ± 6,1	4,2 ± 7,3	100	8,4 ± 11,8	9
HeHOP fast	10,0 ± 27,2	5,9 ± 9,6	4,0 ± 6,1	4,3 ± 7,4	100	9,3 ± 11,9	3,1

^a Das Training findet auf synthetische Daten statt.

^b Die Rechenzeit ist auf einer GPU gemessen.

Tabelle 4.1.: Vergleich der Ergebnisse von Algorithmen zur Bestimmung der Position und Orientierung des Kopfes, evaluiert auf dem BIWI-Datensatz (Fanelli et al., 2013). Für die Bestimmung der Nase, die einzelnen Orientierungswinkel und die Richtung der Kopforientierung ist der mittlere Fehler und die Varianz angegeben. Zusätzlich sind die Bilder bei denen keine Schätzung für die Kopfpose vorhanden ist in Prozent angegeben und die Rechenzeit in Millisekunden.

Bild die Kopfpose bestimmen. Allerdings ist die Verwendung der Trainingsdaten bei den einzelnen Veröffentlichungen unterschiedlich, was einen objektiven Vergleich erschwert. Während das hier vorgestellte Verfahren sowie die Methoden von Fanelli et al. (2013) und Riegler et al. (2014) eine 4-fache Kreuzvalidierung durchführen, trainiert der Algorithmus von Papazov et al. (2015) auf synthetischen Daten und die Ergebnisse sind der Mittelwert des kompletten BIWI-Datensatzes. In Tabelle 4.1 werden zusätzlich zu den Verfahren, die eine 4-fache Kreuzvalidierung durchgeführt haben, die Resultate von Papazov et al. (2015) auf den kompletten Daten angegeben.

Zum Kompromiss zwischen Rechenzeit und Genauigkeit werden neben dem hier vorgestellten Verfahren auch für andere Methoden verschiedene Varianten vorgestellt. Während Fanelli et al. (2011a) unterschiedliche Rechenzeiten erreicht, indem die Abstände der ausgewerteten Pixel variiert werden, verwendet Papazov et al. (2015) unterschiedliche Anzahlen von Dreiecken. Bei den Verfahren von Fanelli et al. (2013) und Papazov et al. (2015) gibt es mehrere Versionen, wobei Tabelle 4.1 jeweils die Ergebnisse der Version mit der schnellsten Rechenzeit sowie diejenige mit der höchsten Genauigkeit zeigt.

Anhand der Ergebnisse der 4-fachen Kreuzvalidierung für den mittleren Winkel- und Translationsfehler werden die beiden Varianten des vorgestellten Verfahrens zu Algorithmen mit dem Stand der Forschung verglichen. In Tabelle 4.1 sind die Ergebnisse von HeHOP und HeHOP fast im Vergleich zu aktuellen Algorithmen dargestellt. Hier ist es wichtig hervorzuheben, dass die Verfügbarkeit für alle Daten gegeben ist, während es bei den anderen Verfahren für bis zu 6 Prozent der Daten keine Resultate gibt. Diese nicht vorhandenen Resultate haben dadurch auch keinen Einfluss auf das Ergebnis des mittleren Fehlers der Winkel und der Nasenposition. Ein weiteres Vorteil des HeHOP-Verfahrens ist die Rechenzeit, die mit über 300 FPS mehr als 3

Mal schneller ist als alternative Algorithmen. Insgesamt folgt aus dem Vergleich zu vorhandenen Algorithmen für das in diesem Kapitel vorgestellte Verfahren eine mit dem Stand der Forschung vergleichbare Genauigkeit der Translation und Rotation des Kopfes bei einer geringeren Rechenzeit.

4.5. Fazit

Dieses Kapitel stellt das im Rahmen der vorliegenden Arbeit entwickelte *Highly efficient Head Orientation and Position (HeHOP)*-Verfahren zur Bestimmung der Kopfpose aus Tiefendaten vor. Die Orientierung und Position des Kopfes wird für jedes Bild unabhängig berechnet. Um wenig Rechenkapazitäten zu verwenden, wird eine regressions-basierte Methode, die stufenweise die Kopfpose bestimmt, vorgestellt. Zum Vergleich des Verfahrens gegenüber aktuellen Methoden nach dem Stand der Forschung wird eine 4-fache Kreuzvalidierung auf dem öffentlich vorhandenen BIWI-Kopfposendatensatz (Fanelli et al., 2013) durchgeführt. Die Evaluation zeigt, dass die vorgestellte Methode vergleichbare Ergebnisse bei der Genauigkeit erreicht mit einer höheren Verfügbarkeit und bei geringerer Rechenzeit pro Bild. In Kapitel 6 wird der hier vorgestellte Algorithmus auf dem im Rahmen dieser Arbeit entwickelten Kopfposen-Datensatz im Fahrzeuginnenraum unter verschiedenen Einflüssen während realer Autofahrten evaluiert.

Kapitel 5

Tiefes neuronales Netz zur Bestimmung der Orientierung des Kopfes

Dieses Kapitel präsentiert die im Rahmen dieser Arbeit entstandene Deep Learning Architektur zur Bestimmung der Kopforientierung und die Fusionsverfahren zur Kombination von Infrarot- und Tiefendaten. Die Architektur und die Fusionsverfahren wurden in der Veröffentlichung von [Schwarz et al. \(2017\)](#) bereits vorgestellt und in Zusammenarbeit mit Monica Haurilet entwickelt.

In der Bildverarbeitung zeigen tiefe neuronale Netze im Vergleich zu anderen Verfahren weitaus bessere Ergebnisse als konventionelle Methoden und sind damit vielversprechend für zahlreiche Anwendungen. Beispielsweise bei der *ImageNet Visual Recognition Challenge* (ILSVRC) ([Russakovsky et al., 2015](#)) erreichen tiefe Neuronale Netze enorme Verbesserungen bei der Klassifizierung von Bildern in Objektkategorien. AlexNet ([Krizhevsky et al., 2012](#)) ist das erste Netzwerk, das signifikant bessere Ergebnisse erzielt als andere Verfahren auf dem Datensatz. Danach übertrafen die tiefen neuronalen Netzarchitekturen VGG ([Simonyan und Zisserman, 2015](#)), GoogLeNet ([Szegedy et al., 2015](#)) und ResNet ([He et al., 2016](#)) diese Resultate. Um die Performance von neuronalen Netzen auf die Bestimmung der Orientierung des Kopfes im Fahrzeuginnenraum zu untersuchen, wird im Rahmen dieser Arbeit eine neue neuronale Netzarchitektur entwickelt, die auf der VGG-Architektur basiert.

Zusätzlich entwickeln [Misra et al. \(2016\)](#) eine neue *stitch*-Einheit, welche den Austausch zwischen neuronalen Netzen ermöglicht. Diese *stitch*-Einheit tauscht Informationen zwischen zwei Modellen aus und verbessert damit die Ergebnisse der Modelle zur semantischen Segmentierung und Berechnung von Oberflächennormalen. Hierbei werden dieselben Eingangsdaten verwendet, um zwei unterschiedliche Anwendungen zu lösen. Inspiriert von dieser Methode wird ein *stitch-basiertes* Fusionsverfahren von Grau- und Tiefenwerten vorgestellt. Im Gegensatz zu dem Algorithmus von [Misra et al. \(2016\)](#) handelt es sich hierbei um zwei unterschiedliche Eingangsdaten für

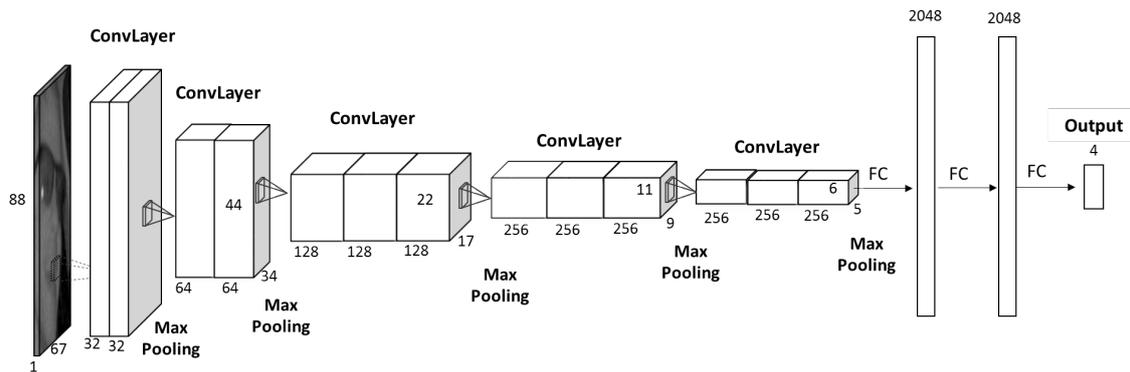


Abbildung 5.1.: Architektur des tiefen neuronalen Netzes zur Bestimmung der Orientierung des Kopfes aus einer einzelnen Modalität, entweder Infrarot- oder Tiefendaten.

dieselbe Anwendung der Bestimmung der Kopforientierung. Neben diesem Fusionsverfahren wird die Architektur zur *frühen* und *späten* Fusion angepasst.

Der erste Teil des Kapitels (siehe Kapitel 5.2) gibt einen Überblick über die generelle Architektur des *Head Pose Networks (HPN)*. Dieses Modell kann entweder aus 2D- oder Tiefendaten die Orientierung des Kopfes bestimmen. Um beide Modalitäten zu fusionieren, werden anschließend unterschiedliche Verfahren vorgestellt (siehe Kapitel 5.3). Bei den Fusionsverfahren wird eine Architektur zur *frühen* Fusion in Abschnitt 5.3.1, eine zur *späten* Fusion in Abschnitt 5.3.2 und eine *stitch-basierte* Fusion in Abschnitt 5.3.3 vorgestellt.

5.1. Ziel

Das Ziel des neuronalen Netzes ist die Bestimmung der Orientierung des Kopfes aus einem einzelnen Tiefen- oder Infrarotbild sowie aus der Kombination beider Modalitäten. Der Input ist die Gesichtsregion des Tiefenbildes D_i , des Infrarotbildes I_i oder der Kombination $\{D_i, I_i\}$. Mit dem tiefen neuronalen Netz wird die Orientierung des Kopfes in Quaternionen \mathbf{q}_i ausgegeben. Die Architektur des neuronalen Netzes wird einheitlich gewählt und mit Fusionsverfahren erweitert. Dadurch ist eine direkte Vergleichbarkeit des Einflusses der Eingangsdaten auf die Ergebnisse möglich.

5.2. Architektur des Head Pose Networks (HPN)

In Abbildung 5.1 ist die Architektur des tiefen neuronalen Netzes zur Bestimmung der Kopforientierung grafisch dargestellt. Ähnlich zu Simonyan und Zisserman (2015) besteht das Netzwerk aus insgesamt 16 *Convolutional* und *Fully Connected Layers* mit Max Pooling Layern zwischen den *Convolutional Layers*. Die *Convolutional Layer* bestehen aus trainierten Filtern die auf lokalen Regionen des Inputbildes

agieren. Im Anschluss an zwei bis drei *Convolutional Layers* verkleinern die *Max Pooling Layer* mit einer Größe von 2×2 den Input, indem mit einer Schrittweite von zwei das Maximum dieser Region gewählt wird. Zum Schluss folgen drei Fully Connected Layers, um einen globalen Ausgabewert zu erhalten. Als Aktivierungsfunktion wird in jedem Layer Rectified Linear Units (ReLU) (Nair und Hinton, 2010) mit $f(\mathbf{x}) = \max(0, \mathbf{x})$ angewendet. Im Vergleich zu Simonyan und Zisserman (2015) ist die Anzahl der Neuronen halbiert, dadurch entsteht folgende Struktur:

$2 \times \text{ConvLayer}(32)$, *MaxPool*, $2 \times \text{ConvLayer}(64)$, *MaxPool*, $3 \times \text{ConvLayer}(128)$, *MaxPool*, $3 \times \text{ConvLayer}(256)$, *MaxPool*, $3 \times \text{ConvLayer}(256)$, *MaxPool*, *FC(2048)*, *FC(2048)*, *FC(4)*.

Das Netzwerk verwendet während des Trainings als Eingangsgrößen zufällige Regionen der Größe 88×67 innerhalb der Gesichtsregion, die von dem Gesichtsdetektor King (2009) detektiert wird. Dabei sind die Gesichtsregionen auf eine einheitliche Größe von 91×70 Pixels skaliert. Als Ausgangsgröße entstehen Quaternionen $\mathbf{q}_{est} \in \mathbb{R}^4$, welche die Rotation des Kopfes beschreiben, wie in Kapitel 2.5 beschrieben.

Ein wichtiger Aspekt der Trainingsphase ist die Wahl der Kostenfunktion, die minimiert wird, um die Parameter des Netzwerkes zu finden. Da es sich hier um ein Regressionsproblem und kein Klassifizierungsproblem handelt, kann die häufig verwendete *cross-entropy*-Funktion nicht angewendet werden. Ahn et al. (2014) verwenden zur Optimierung des neuronalen Netzes für die Schätzung der Kopforientierung eine Kostenfunktion basierend auf den Eulerwinkeln. Allerdings kommt es bei der Verwendung von Eulerwinkel zur Beschreibung einer Rotation zu dem Problem des Gimbal Locks. Deshalb beinhaltet dieses Netzwerk als Ausgangsgrößen Quaternionen. Basierend auf Quaternionen wird eine Loss-Funktion für das Regressionsproblem definiert, welche ähnlich zu der Kostenfunktion von Kendall et al. (2015) zum Training eines tiefen neuronalen Netzes zur Bestimmung von Objektposen ist. Im Gegensatz zur Loss-Funktion von Kendall et al. (2015) beinhaltet sie einen zusätzlichen Regularisierungsterm. Während des Trainings wird folgende Kostenfunktion minimiert, um die Parameter des tiefen neuronalen Netzes zu optimieren:

$$\ell_{\gamma}(\mathbf{q}_{gt}, \mathbf{q}_{est}) := \left\| \mathbf{q}_{gt} - \frac{\mathbf{q}_{est}}{\|\mathbf{q}_{est}\|} \right\| + \gamma \cdot \alpha(\mathbf{q}_{est}) \quad (5.1)$$

Hierbei bezeichnet $\mathbf{q}_{est} \in \mathbb{R}^4$ die geschätzten Quaternionen und $\mathbf{q}_{gt} \in \mathbb{R}^4$ die Referenzwerte der Quaternionen. Im Vergleich zu Kendall et al. (2015) wird zusätzlich der Normalisierungsterm $\alpha(\mathbf{q}) = \|\mathbf{q}\|$ verwendet, um sicherzustellen, dass die einzelnen Werte innerhalb des Quaternionenvektors \mathbf{q}_{est} klein bleiben. Das geschätzte Quaternion \mathbf{q}_{est} wird mit $\|\mathbf{q}_{est}\|$ normiert, da zur Beschreibung von Rotationen Quaternionen mit der Länge 1 verwendet werden.

Zur Initialisierung der Gewichte des Modells werden zufällige normalverteilte Werte gesetzt mit einem Bias von null. Während des Trainings wurden sowohl Modelle mit Regularisierungsterm, d.h. $\gamma > 0$, als auch ohne Regularisierung, d.h. $\gamma = 0$ trainiert. Bei den Modellen ohne Regularisierungsterm kam es zu sehr hohen Werten

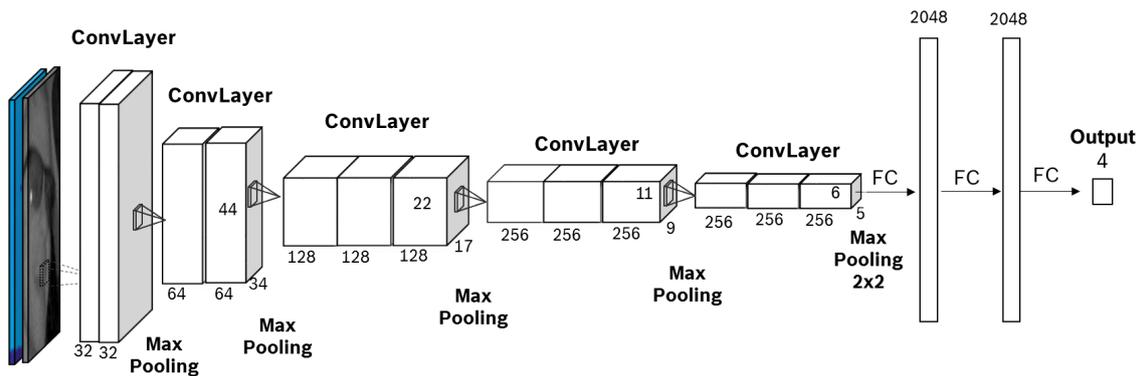


Abbildung 5.2.: *Frühe Fusion*. Architektur des tiefen neuronalen Netzes zur frühen Kombination von 2D- und Tiefendaten. Bei dieser Fusion werden die Eingangsbilder fusioniert.

innerhalb der Quaternionen. Deshalb werden im Folgenden nur Ergebnisse präsentiert, bei denen das Modell mit einem Regularisierungsterm trainiert wurde. Für die Parameteroptimierung wird Adam (Kingma und Ba, 2014a) verwendet, mit einer Anfangslernrate von 0.001. Das Modell wird mit 600k Iterationen trainiert mit Mini-Stapeln der Größe 32.

In diesem Abschnitt ist die allgemeine Struktur des *Head Pose Networks (HPN)* dargestellt. Mit dieser Struktur kann entweder aus 2D- oder Tiefendaten die Kopforientierung bestimmt werden. Hierdurch können die beiden Modalitäten als Eingangswerte einzeln verglichen werden.

5.3. Fusionsverfahren

Neben dem Vergleich von 2D- und Tiefenwerten als Eingangswerten ist es von großem Interesse, ob die Fusion der beiden Modalitäten das Ergebnis verbessern kann. Hierzu ist es notwendig die Modalitäten auf intelligente Weise miteinander zu fusionieren. Außerdem muss es möglich sein, die neuen Fusionsverfahren mit den Verfahren, welche nur eine Modalität verwenden, zu vergleichen. Im folgenden Abschnitt wird eine Architektur zur frühen Fusion (siehe Abschnitt 5.3.1), eine Fusion zur späten Fusion (siehe Abschnitt 5.3.2) und eine stitch-basierte Fusion (siehe Abschnitt 5.3.3) vorgestellt. Diese Fusionen verwenden als Basis die im vorherigen Abschnitts 5.2 vorgestellte tiefe neuronale Netzarchitektur des *HPN*. Dadurch ist ein direkter Vergleich abhängig von den verwendeten Modalitäten im Kapitel 6 möglich.

5.3.1. Frühe Fusion

Abbildung 5.2 zeigt die Architektur des frühen Fusionsmodells. Bei der frühen Fusion werden bereits die Eingangswerte fusioniert. Im Vergleich zu der Basis-Architektur (siehe Abbildung 5.1) des *HPN* beinhaltet die Eingangsmatrix zwei Kanäle mit den

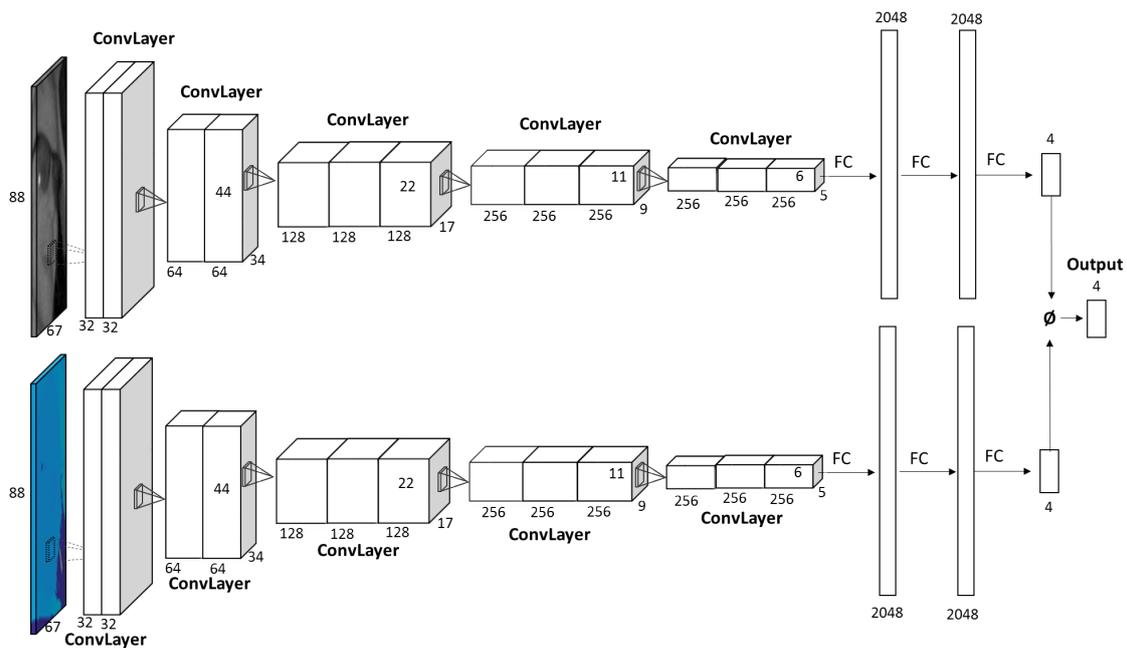


Abbildung 5.3.: *Späte Fusion*. Architektur des tiefen neuronalen Netzes zur späten Kombination von 2D- und Tiefendaten. Bei dieser Fusion wird der Mittelwert der Ausgaben bestimmt.

beiden Modalitäten. Dadurch wird an jedem Pixel ein Grauwert und ein Tiefenwert zusammengefügt. Anschließend ist die Architektur des Netzwerkes dieselbe wie die Basis-Architektur des *HPN* bestehend aus *Convolutional Layern*, *Max Pooling Layern* und *Fully Connected Layern*. Während der Trainingsphase werden die Parameter des Modells optimiert. Die Ausgabe des Netzes ist die Rotation des Kopfes in Quaternionen.

Mit diesem Netzwerk wird aus den zusammengeführten Pixeln beider Modalitäten die Rotation des Kopfes in Quaternionen bestimmt. Die Besonderheit dieser Fusion ist, dass durch das Zusammenfügen der Pixel entweder beide Werte Einfluss auf das Ergebnis haben oder keiner der beiden.

5.3.2. Späte Fusion

Die Architektur der späten Fusion ist in Abbildung 5.3 visualisiert. Als Eingabe werden die beiden Modalitäten als einzelne Matrizen betrachtet und für jede Modalität folgt ein tiefes neuronales Netz, welches die Architektur des *HPN* hat. Während der Trainingsphase werden die Parameter der beiden neuronalen Netze getrennt voneinander optimiert. Die Ausgaben der Netze sind jeweils Quaternionen für die Rotation des Kopfes. Anschließend fusioniert das Netzwerk die Ausgaben der beiden Modelle, indem der Durchschnitt gebildet wird.

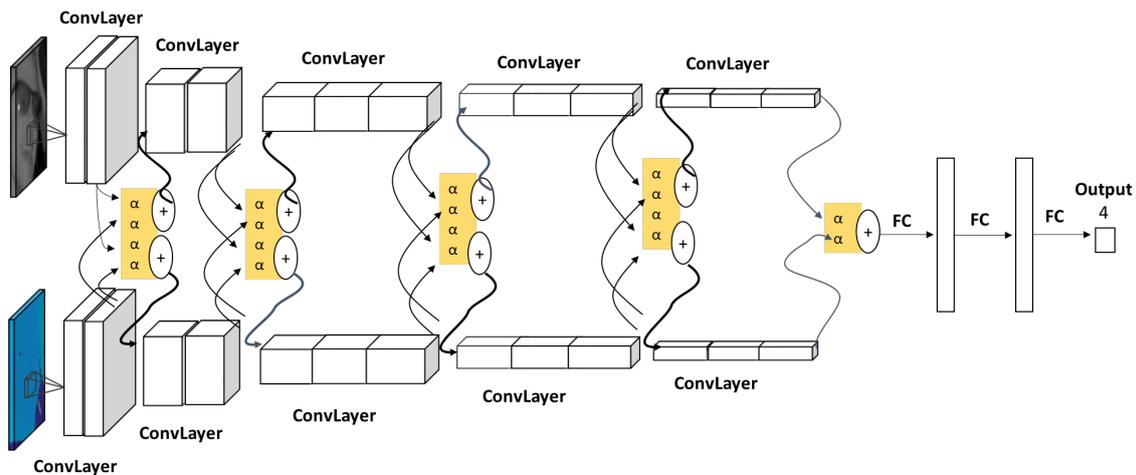


Abbildung 5.4.: *Stitch-basierte Fusion*. Architektur des tiefen neuronalen Netzes zur Kombination von 2D- und Tiefendaten mit stitch-basierten Gewichten.

Bei diesem Netzwerk werden für die beiden Modalitäten jeweils getrennt die Parameter der Modelle trainiert. Durch das Trennen der Modalitäten während der Trainingsphase, werden die Netze nicht voneinander beeinflusst. Dadurch ist der Einfluss einer lokalen Region einer Modalität auf das Ergebnis unabhängig von dem Wert der lokalen Region in der Matrix der anderen Modalität.

5.3.3. Stitch-basierte Fusion

Abbildung 5.4 stellt die Architektur der stitch-basierten Fusion grafisch dar. Die Fusion dieser Methode ist inspiriert von dem Algorithmus von Misra et al. (2016), bei dem ein Austausch zwischen zwei neuronalen Netzen stattfindet. Bei dem Verfahren von Misra et al. (2016) stammen die Eingangswerte der beiden Netze von demselben Bild, die einzelnen Netze verfolgen allerdings zwei unterschiedliche Zielaufgaben. Das eine neuronale Netz verfolgt einen Klassifizierungsansatz während das andere eine Segmentierung löst. Ähnlich zu diesem Ansatz findet in der hier präsentierten Fusionsmethode ein Austausch zwischen zwei neuronalen Netzen mithilfe von *stitch* statt. Für die Architektur der beiden Netzwerke wurde die Struktur des HPN-Netzwerkes gewählt. Im Gegensatz zu dem Verfahren von Misra et al. (2016) handelt es sich bei den Daten der Fusionsmethode um unterschiedliche Eingangswerte, jedes Netzwerk hat die Bildmatrix einer Modalität als Eingangswerte. Des Weiteren sind die beiden Teilnetzwerke im Gegensatz zu Misra et al. (2016) auf dieselbe Aufgabe spezialisiert, nämlich die Orientierung des Kopfes herauszufinden.

Der Vorteil dieses Verfahrens ist, dass während des Trainings automatisch die Gewichte optimiert werden, welche die beiden Modalitäten fusionieren. Diese Fusion findet nach jedem *Max Pooling Layer* statt. Durch diese automatisierte Fusion wird während des Trainings die optimale Fusion der beiden Modalitäten gefunden.

5.4. Evaluation

Bei der Evaluation ist der Vergleich der einzelnen Eingangsdaten, Grauwert- oder Tiefendaten, gegenüber der Kombination der Eingangsdaten mit den vorgestellten Fusionsmethoden von Interesse. Um die Fusionsverfahren anzuwenden sind pixelweise übereinstimmende Grauwert- und Tiefendaten notwendig. Diese pixelweise Übereinstimmung ist bei dem in Kapitel 3 vorgestellten Kopfposendatensatz im Fahrzeuginnenraum vorhanden. Die ausführliche Evaluation des *HPN*-Netzwerkes auf Infrarot- oder Tiefenbilder sowie die Anwendung der Fusionsverfahren wird in Kapitel 6 präsentiert. Zusätzlich wird die Genauigkeit der Verfahren beim Einfluss von Brillen, Sonnenbrillen und weiteren Verdeckungen im Gesichtsbereich analysiert.

5.4.1. Heatmaps zur Visualisierung des Einflusses von Pixel-Regionen.

Um den Einfluss der einzelnen Pixel-Regionen auf die Genauigkeit der Kopforientierung zu bestimmen, werden in Abbildung 5.5 *Heatmaps*, basierend auf der Methode von Zeiler und Fergus (2014), visualisiert. Im folgenden Abschnitt wird die Berechnung der *Heatmaps* beschrieben und es erfolgt die Analyse der visualisierten *Heatmaps*.

Der Einfluss von $d \times d$ großen Pixel-Regionen auf die Schätzung der Kopfpose soll im Folgenden ermittelt werden. Hierzu werden die Pixel-Werte an den Regionen, zentriert auf ein gegebenes Pixel i , ersetzt und anschließend bestimmt wie sich der Fehler des tiefen neuronalen Netzes verändert. Die Pixel-Regionen werden mit Werten ersetzt, die vom tiefen neuronalen Netz möglichst unbeachtet bleiben. Hierzu wird in dieser Arbeit das mittlere Gesicht aus Grauwert- und Tiefenbild der gesamten Trainingsdaten erzeugt und die abgedeckten Pixel-Regionen mit diesen Werten ersetzt. Diese Manipulation der Eingangsdaten verursacht einen Fehler in der Schätzung der Kopfpose. Die Amplitude des Fehlers wird für jedes Pixel i farblich kodiert in einer Heatmap dargestellt. Hierbei repräsentieren blaue Pixel einen geringen und rote Pixel einen großen Fehler. In Abbildung 5.5 werden gemittelte *Heatmaps* von Szenen des *DriveAhead*-Datensatzes dargestellt. Für jede Szene wurden 20 Infrarot- und Tiefenbilder gewählt mit ähnlichen Kopforientierungen. Die erste Spalte beinhaltet Personen mit Brille, die zweite Spalte Personen ohne Brille, die dritte Spalte Personen mit nach links ausgerichteter Kopforientierung und die vierte Spalte Personen mit nach rechts ausgerichteter Kopforientierung und sichtbaren Markern.

Insgesamt verdeutlichen die *Heatmaps*, dass bei denselben Eingangsbildern, die verschiedenen *HPN*-Verfahren unterschiedliche Gesichtsbereiche als Eigenschaften für die Schätzung der Kopforientierung verwenden. Während die *Heatmaps* der zweiten Spalte für die verschiedenen *HPN*-Verfahren unterschiedlich sind, ist in der ersten Spalte der Brillenbügel für alle Verfahren eine wichtige Eigenschaft. In der dritten

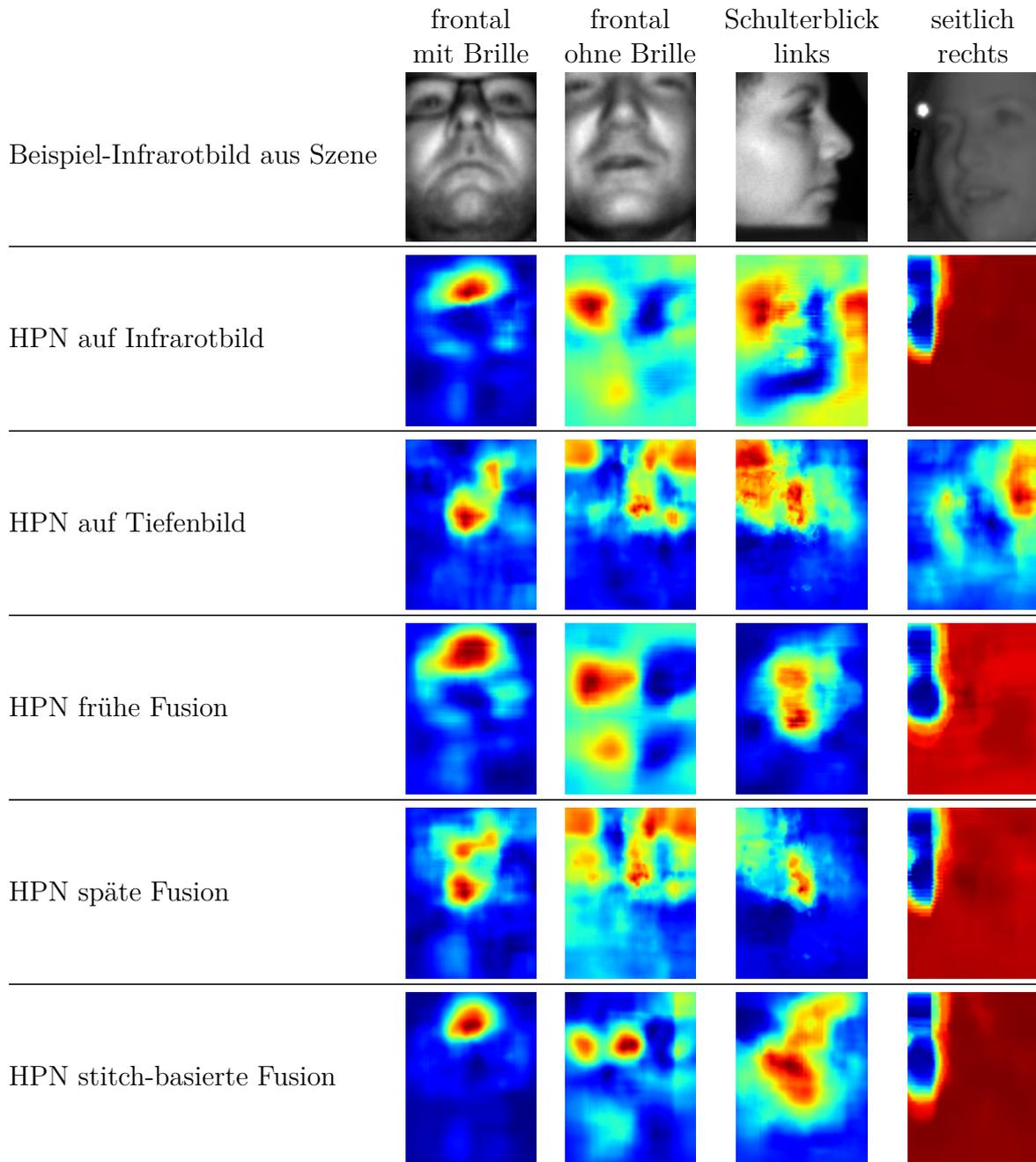


Abbildung 5.5.: Die Spalten zeigen unterschiedliche Szenen des *DriveAhead*-Datensatzes: Frontal mit Brille, frontal ohne Brille, Schulterblick links, leichte Drehung rechts mit sichtbaren Markern. In den Zeilen sind die gemittelten *Heatmaps* aus den Szenen visualisiert bei Anwendung von *HPN* auf Infrarot- oder Tiefenbildern sowie Anwendung der frühen, späten und stitch-basierten Fusion. Die Heatmaps präsentieren farblich kodiert die Fehler der Kopforientierung bei Verdeckung des Eingangsbildes mit Regionen der Größe $d \times d$. Das Farbspektrum geht von *blau* für geringe Fehlerwerte bis *rot* für große Fehlerwerte.

Spalte, welche die *Heatmaps* einer starken Kopfdrotation darstellt, liegen für alle Verfahren bis auf das rein infrarot-basierte Verfahren die wichtigen Merkmale innerhalb des Gesichts. Aus den ersten drei Spalten wird deutlich, dass durch die Verwendung von Tiefenwerten die Formunterschiede des Gesichts aussagekräftige Merkmale sind, wie zum Beispiel die Nase.

Die in der letzten Spalte dargestellte Szene beinhaltet seitliche Kopfdrehungen mit sichtbaren Markern. Bei sichtbaren Markern kommt es bei Anwendung der Infrarot-basierten *HPN*-Verfahren häufig zu einer Fehlbestimmung der Kopforientierung. Um diesen Effekt zu untersuchen wurden in der letzten Spalte Zeitpunkte gewählt bei denen die Infrarot-basierte Bestimmung fehlerbehaftet ist. Anhand der *Heatmaps* wird deutlich, dass bei Abdeckung der sichtbaren Marker der Fehler kleiner wird. Daraus folgt, dass die tiefen neuronalen Netze die Kopforientierung nicht basierend auf den Markerpositionen bestimmen. Dies wurde sichergestellt, indem die sichtbaren Marker in den Trainingsdaten mithilfe einer Vorverarbeitung entfernt wurden, siehe Kapitel 6.2. Allerdings wird durch die *Heatmaps* deutlich, dass die sichtbaren Marker, welche ungelernen Störungen in der Nähe des Gesichts entsprechen, die Infrarot-basierten tiefen neuronalen Netze beeinflussen und das Ergebnis verschlechtern. Im Gegensatz dazu ist die Abdeckung der Marker-Region bei der *Heatmap* des tiefen-basierten *HPN*-Verfahren nicht sichtbar. Daraus folgt, dass das rein auf Tiefendaten gelernte *HPN*-Verfahren robuster gegenüber ungelernen Störungen in der Nähe des Gesichtsbereiches ist.

5.5. Fazit

Dieses Kapitel präsentiert eine neue tiefe neuronale Netzwerk-Architektur, das *Head Pose Network (HPN)*, zur Bestimmung der Kopforientierung. Die Architektur basiert auf der *VGG-Architektur Simonyan und Zisserman (2015)*, welche den Stand der Forschung auf zahlreichen Anwendungen repräsentiert. Es erfolgt im Rahmen der vorliegenden Arbeit die Weiterentwicklung dieser Architektur, indem die Anzahl der Neuronen und die Kostenfunktion zur Optimierung der Parameter des Netzwerks auf die Kopforientierung angepasst wird. Zusätzlich werden Fusionsverfahren zur Kombination von Grau- und Tiefenwerten diskutiert, die das *HPN* als Basis-Netzwerk verwenden. Als Fusionsverfahren werden die *frühe*, *späte* und *stitch-basierte* Fusion vorgestellt. Durch die gleichbleibende Basis-Architektur der Netzwerke ist eine direkte Vergleichbarkeit der Verfahren möglich. Mit der *HPN*-Architektur kann ermittelt werden, welche Art von Eingangsdaten, Grau- oder Tiefenwerte, bessere Ergebnisse liefern. Zusätzlich kann mit den Fusionsverfahren analysiert werden, ob eine Kombination der Modalitäten die Genauigkeit der Bestimmung der Kopforientierung weiter verbessern kann.

Kapitel 6

Evaluation auf Kopfposendatensatz im Fahrzeuginnenraum

Die in diesem Kapitel präsentierten Ergebnisse ausgewertet auf dem *DriveAHead*-Datensatz wurden bereits teilweise im Rahmen der Publikation [Schwarz et al. \(2017\)](#) veröffentlicht. In der Veröffentlichung wurde die Vorverarbeitung genannt, die Metriken beschrieben und die Ergebnisse der Analyse von Orientierung und Translation auf allen Daten abhängig von den Methoden und Modalitäten sowie auf Daten mit Verdeckungen im Gesicht erläutert. Zusätzlich wurden die Resultate der Analyse der Kopforientierung mit Sonnenbrillen und Brillen diskutiert.

Für die Anwendung von Verfahren zur Bestimmung der Kopfpose im Fahrzeuginnenraum ist die Analyse auf realen Daten essentiell. Die Entwicklung des *DriveAHead*-Datensatzes aus Kapitel 3 ermöglicht eine Evaluation auf der Basis von Tiefen- und Infrarotdaten, die während realer Autofahrten im Straßenverkehr aufgenommen wurden. Das folgende Kapitel liefert eine ausführliche Evaluation der Bestimmung der Kopfpose im Fahrzeuginnenraum auf der Basis des *DriveAHead*-Datensatzes.

Beginnend mit Kapitel 6.1, werden die Metriken für die Bewertung der Orientierung und Position des Kopfes genannt. Für die Orientierung des Kopfes wird eine speziell für die Fahrerbeobachtung entwickelte Metrik, der Balanced Mean Angular Error (BMAE), vorgestellt. Dieses Fehlermaß berücksichtigt, dass die Kopfdrehung bei Autofahrten häufig frontal ist und stellt sicher, dass die Verfahren auch für große Kopfdrehungen generalisieren können. Für die Evaluation ist eine Vorverarbeitung der *DriveAHead*-Daten notwendig, die in Abschnitt 6.2 präsentiert wird. Kapitel 6.3 nennt die Methoden, welche mit dieser Metrik auf den Bildern des *DriveAHead*-Datensatz ausgewertet werden. Als Methoden werden aus der Literatur Verfahren nach dem Stand der Technik ([Ahn et al., 2014](#); [Baltrušaitis et al., 2016](#)) gewählt und den in dieser Arbeit (siehe Kapitel 4 und Kapitel 5) entwickelten Verfahren gegenüber gestellt. Diese Verfahren werden ausführlich zur Bestimmung der Orientierung

des Kopfes in Abschnitt 6.4 und Translation des Kopfes in Abschnitt 6.5 untersucht. Bei der Untersuchung werden unterschiedliche Einflüsse, wie die Wahl der Methoden, Verdeckungen in der Gesichtsregion, Brillen und Sonnenbrillen, analysiert.

6.1. Evaluationsmetrik

Dieser Abschnitt stellt die Metriken vor, die bei der Auswertung angewendet werden. Abschnitt 6.1.1 präsentiert die im Rahmen dieser Arbeit entstandene Metrik *Balanced Mean Angular Error* zur Evaluation der Orientierung des Kopfes. Die Translation wird über die euklidische Distanz mit der in Abschnitt 6.1.2 genannten Formel ausgewertet.

6.1.1. Balanced Mean Angular Error (BMAE)

Ein häufig verwendetes Fehlermaß der Rotation ist der mittlere Winkelfehler der Eulerwinkel (z.B. Fanelli et al. (2013)). Da dieser Datensatz im Labor aufgenommen wurde, sind dort alle Rotationen mit derselben Häufigkeit abgedeckt. Im Gegensatz zu diesem Datensatz wurde der hier vorgestellte *DriveAhead*-Datensatz während realer Autofahrten aufgenommen, wodurch der Kopf meistens frontal ausgerichtet ist.

Um eine Robustheit der Verfahren über einen großen Winkelbereich sicherzustellen, wird im Folgenden eine neue Metrik, der *Balanced Mean Angular Error (BMAE)*, zur Bestimmung des Rotationsfehlers vorgestellt. Diese Metrik schafft einen Ausgleich zwischen häufiger und seltener vorkommenden Winkeln, indem jeder Winkelbereich denselben Einfluss auf die Metrik hat. Hierfür wird der gesamte Winkelbereich in gleich große Segmente geteilt und für jedes Segment der mittlere Winkelfehler bestimmt. Anschließend bildet der Mittelwert über die einzelnen Winkelfehler den *BMAE*. Damit gewichtet der *BMAE* die Winkelsegmente und berücksichtigt die einzelnen Winkelbereiche in gleichem Maß.

Abbildung 6.1 zeigt grafisch die Berechnung des *Balanced Mean Angular Error (BMAE)*. Der *BMAE* berechnet sich nach folgender Formel:

$$\text{BMAE} := \frac{d}{k} \sum_m \Delta\Phi_{m,m+d}, m \in d\mathbb{N} \cap [0, k], \quad (6.1)$$

wobei $\Delta\Phi_{m,m+d}$ der Winkelfehler im Intervall $[m, m+d]$, d die Segmentgröße und k die maximal berücksichtigte Rotation sind. Für den mittleren Winkelfehler $\Delta\Phi_{m,m+d}$ im Intervall $[m, m+d]$ ergibt sich:

$$\Delta\Phi_{m,m+d} := \frac{1}{N} \sum_{\mathbf{q}_{gt} \in [m, m+d]} \Delta\phi(\mathbf{q}_{est}, \mathbf{q}_{gt}), \quad (6.2)$$

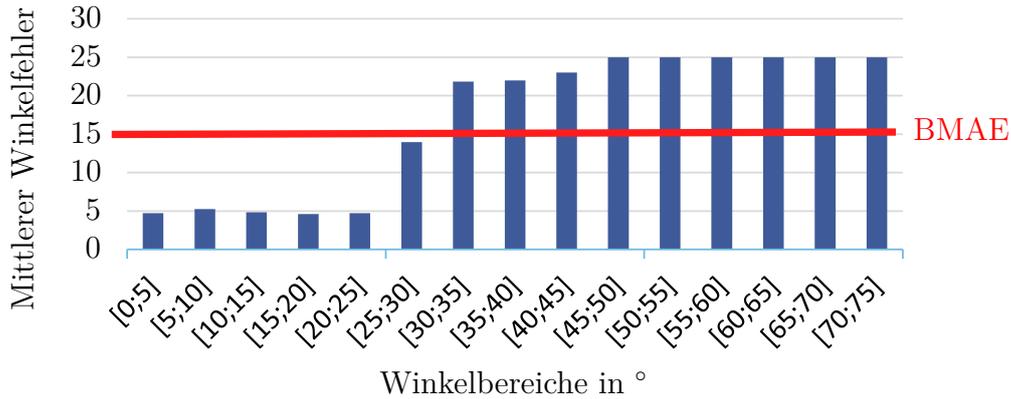


Abbildung 6.1.: Exemplarische Darstellung der Berechnung des *Balanced Mean Angular Error (BMAE)*, der BMAE ist in *rot* visualisiert. Dieser berechnet sich aus den mittleren Winkelfehlern $\Phi_{m,m+d}$ der einzelnen Segmente $[m, m+d]$, wobei die mittleren Winkelfehler in *blauen* Balken dargestellt sind.

wobei N der Anzahl von $\mathbf{q}_{gt} \in [m, m+d]$ entspricht.

Der Winkelfehler $\Delta\phi$ der geschätzten Rotation in Quaternionen \mathbf{q}_{est} zur Referenzmessung \mathbf{q}_{gt} berechnet sich aus :

$$\Delta\phi(\mathbf{q}_{est}, \mathbf{q}_{gt}) := 2 \cdot \arccos(\langle \mathbf{q}_{est}, \mathbf{q}_{gt} \rangle)$$

Der Vorteil von Quaternionen gegenüber Eulerwinkeln ist, dass sie für eine Rotation eine eindeutige Darstellung bereithalten. Deshalb wird diese Darstellung in der Praxis häufig verwendet, wie zum Beispiel für Rotationen von Objekten in Collet et al. (2011).

In der folgenden Analyse werden Rotationen bis zu einem Winkel von $k = 75^\circ$ untersucht und die Segmentgröße d auf 5° gesetzt.

6.1.2. Translation

Der Fehler der Translation Δd bestimmt sich aus der euklidischen Distanz der Referenztranslation \mathbf{t}_{gt} und der geschätzten Translation \mathbf{t}_{est} , mit:

$$\Delta d(\mathbf{t}_{gt}, \mathbf{t}_{est}) := \|\mathbf{t}_{gt} - \mathbf{t}_{est}\|_2 \quad (6.3)$$

6.2. Vorverarbeitung

Die Referenzmessung des *DriveAhead*-Datensatzes verwendet ein Motion-Capture-System, wie in Kapitel 3 detailliert dargestellt, welches die Orientierung und Position

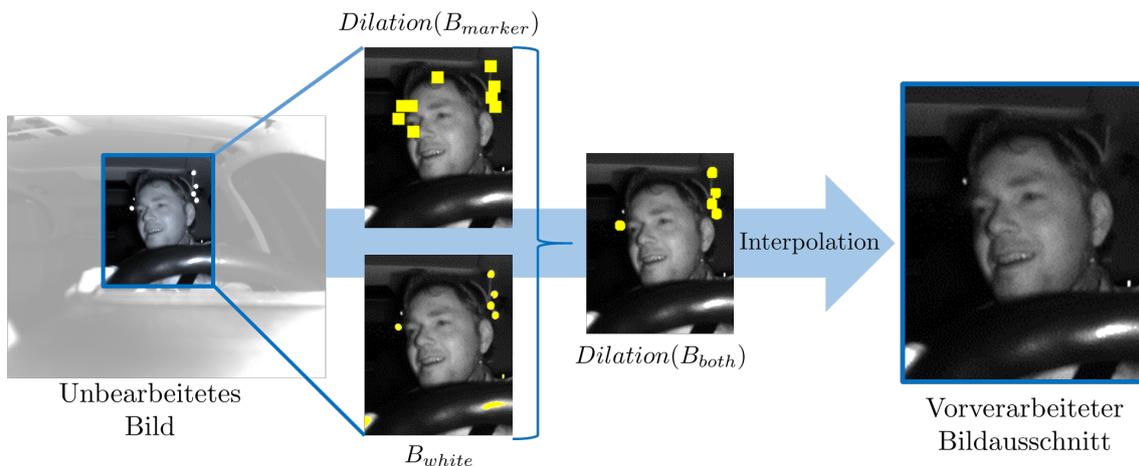


Abbildung 6.2.: Ablauf der Vorverarbeitung zur Entfernung der sichtbaren Marker in den Infrarot- und Tiefenbildern.

des Targets auf dem Kopf des Fahrers bestimmt. Auf diesem Target sind sichtbare Marker befestigt, die vom Motion-Capture-System verfolgt werden. Ein Nachteil dieses Verfahrens ist, dass die Marker in den von der Kamera aufgenommenen Daten ebenfalls sichtbar sind, siehe 6.3 Spalte 1 und 3. Um sicherzustellen, dass die Bestimmung der Kopfpose von den Algorithmen nicht auf den Markern basiert, ist eine Vorverarbeitung der Bilder notwendig.

6.2.1. Entfernung der sichtbaren Marker

Die Vorverarbeitung reduziert die Sichtbarkeit der Marker in den Bildern des Datensatzes. Das Bild wird an den Positionen der Marker mit interpolierten Werten aus der Umgebung korrigiert. Abbildung 6.2 gibt einen Überblick über den Ablauf der Vorverarbeitung. Folgende Schritte werden dabei durchgeführt:

- Bestimmung einer binären Matrix \mathbf{B}_{marker} , welche die Rückprojektion der 3D-Positionen der Kugeln in das Bild beinhaltet.
- Bestimmung einer binären Matrix \mathbf{B}_{white} , welche die weißen Pixel des Bildes enthält.
- Zusammenführung der binären Matrizen \mathbf{B}_{marker} und \mathbf{B}_{white} zu einer binären Matrix \mathbf{B}_{both} , bei der beide Bedingungen erfüllt sind.
- Ersetzen der Pixel aus \mathbf{B}_{both} mit interpolierten Pixeln.

Die im Folgenden vorgestellten binären Matrizen entsprechen alle der Größe des Bildes \mathbf{I} , d.h. es gilt $\forall \mathbf{B}, size(\mathbf{B}) = size(\mathbf{I})$.

Binäre Bildmaske \mathbf{B}_{marker} der Marker Positionen im Bild. Der erste Schritt bestimmt eine binäre Matrix \mathbf{B}_{marker} , welche die Projektion der Marker des Motion-Capture-Systems in das Bild beinhaltet. Dieser Schritt ist notwendig, um herauszufinden an welchen Pixelpositionen sich weiße Stellen aufgrund der reflektierenden Marker befinden können. Die 3D-Positionen der Marker $\mathbf{k}_l \in \mathbb{R}^3$ auf dem Target sind zueinander konstant und bilden dadurch ein bekanntes starres Modell $\mathbf{M} = (\mathbf{k}_1, \dots, \mathbf{k}_9) \in \mathbb{R}^{3 \times 9}$. Das Motion Capture-System verfolgt während der gesamten Aufnahme das Modell \mathbf{M} . Es misst für jedes Bild i die Transformation $\mathbf{T}_i^{w \rightarrow t}$ des Weltkoordinatensystems w in das Targetkoordinatensystem t . Zusätzlich liefert die Kalibrierung die intrinsische Kameramatrix \mathbf{K} und die Transformation $\mathbf{T}^{k \rightarrow w}$ des Kamerakoordinatensystems k in das Weltkoordinatensystem w , welche nach Gleichung 3.12 die Rotationsmatrix $\mathbf{R}_i^{k \rightarrow w}$ und die Translation $\mathbf{t}_i^{k \rightarrow w}$ beinhaltet. Dadurch projizieren sich die 3D-Position des Targets mit folgender Formel in das Bild:

$$\begin{pmatrix} \mathbf{m}_{im} \\ 1 \end{pmatrix} \equiv \mathbf{K} \cdot \left(\mathbf{R}^{k \rightarrow w} \mid \mathbf{t}^{k \rightarrow w} \right) \cdot \mathbf{T}_i^{w \rightarrow t} \cdot \begin{pmatrix} \mathbf{M} \\ \mathbb{1}_{1 \times 9} \end{pmatrix}$$

Daraus berechnen sich die Pixel $\mathbf{m}_{im} \in \mathbb{R}^{2 \times 9}$ im Bild, an denen sich die Mittelpunkte der Marker befinden, falls sie nicht verdeckt sind.

Die binäre Matrix \mathbf{B}_{marker} klassifiziert mögliche Markerpositionen, indem die Pixel der Marker-Mittelpunkte den Wert 1 erhalten während die restlichen 0 sind.

$$\mathbf{B}_{marker}(x, y) := \begin{cases} 1, & \text{falls } (x, y) \in \mathbf{m}_{im} \\ 0, & \text{sonst} \end{cases}$$

Die sichtbaren Kugeln \mathbf{k}_l haben einen Durchmesser von 10 Millimetern, deshalb wird zusätzlich eine Dilatation (Soille, 2013) mit einem Durchmesser d von 13 Pixeln angewendet. Damit enthält die Bildmaske $Dilatation(\mathbf{B}_{marker})$ die möglichen Markerpositionen und deren Umgebungspixel.

Binäre Bildmaske \mathbf{B}_{white} der weißen Pixel des Bildes. Eine zusätzliche binäre Bildmaske \mathbf{B}_{white} filtert die weißen Stellen aus dem Grauwertbild des Infrarotbildes. An diesen weißen Stellen können Markerpositionen sein. In dieser Matrix werden alle Pixel des Bildes \mathbf{I} auf 1 gesetzt, deren Wert größer als ein Grenzwert γ ist, mit:

$$\mathbf{B}_{white}(c, y) := \begin{cases} 1, & \text{falls } \mathbf{I}(x, y) > \gamma \\ 0, & \text{sonst} \end{cases}$$

Binäre Bildmaske \mathbf{B}_{both} der kombinierten Bildmasken. Die binäre Bildmaske \mathbf{B}_{both} kombiniert die beiden vorherigen Matrizen \mathbf{B}_{marker} und \mathbf{B}_{white} , mit:

$$\mathbf{B}_{both}(x, y) := \begin{cases} 1, & \text{falls } Dilatation(\mathbf{B}_{marker}(x, y)) = 1 \text{ und } \mathbf{B}_{white}(x, y) = 1 \\ 0, & \text{sonst} \end{cases}$$

Daraus entsteht eine Matrix mit 1 an den zu interpolierenden Stellen. Anschließend folgt eine Dilatation mit einem Durchmesser von 5 Pixeln, um den Rand der Kugeln

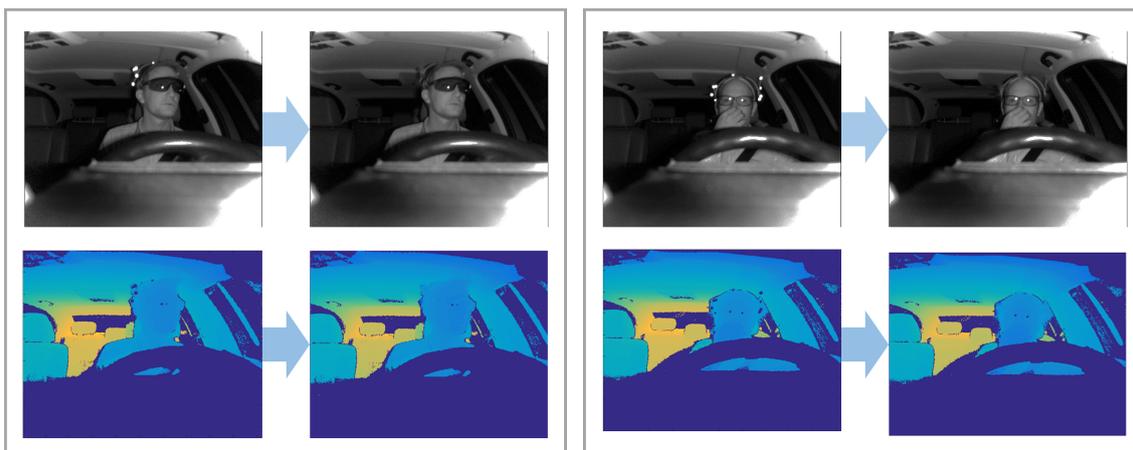


Abbildung 6.3.: Beispielbilder des *DriveAhead*-Datensatzes vor (erste und dritte Spalte) und nach (zweite und vierte Spalte) der Vorverarbeitung. Durch die Vorverarbeitung sind die im Bild als weiße Kugeln sichtbaren Marker mithilfe einer Interpolation der Umgebung nicht mehr sichtbar. Aus Schwarz et al. (2017) © 2017 IEEE.

mit einzuschließen, welcher häufig nicht reflektiert und dadurch in der Matrix \mathbf{B}_{white} nicht enthalten ist. Es entsteht die finale Maske $Dilation(\mathbf{B}_{both})$, mit deren Hilfe die Interpolation der Grau- und Tiefenwerte durchgeführt wird.

Interpolation der Pixel. Die Pixel, an denen die binäre Bildmaske $Dilation(\mathbf{B}_{both})$ die Werte 1 hat, werden mit interpolierten Grauwerten der Umgebungspixel ersetzt. Hierfür wird die in der *Image Processing Toolbox MATLAB (R2016a)* vorhandene Implementierung *regionfill* angewendet, welche eine Interpolation von außen nach innen durchführt. Um eine möglichst glatte Interpolation zu erhalten, werden diskrete Laplace Funktionen auf den Regionen bestimmt und ein Dirichlet-Randproblem (Gilbarg und Trudinger, 2015) gelöst. Abbildung 6.3 zeigt Beispielbilder des *DriveAhead*-Datensatzes vor (in der ersten und dritten Spalte) und nach der Vorverarbeitung (in der zweiten und vierten Spalte).

Das in diesem Abschnitt vorgestellte Verfahren zur Vorverarbeitung der Daten wird auf die gesamten Trainings- und Validierungsdaten des *DriveAhead*-Datensatzes angewendet. Damit wird ausgeschlossen, dass die Verfahren auf den sichtbaren Markern als Merkmale trainieren. Die Testdaten werden als Rohdaten verwendet, das bedeutet, die Marker sind hier sichtbar. Dieser Vorgang stellt sicher, dass die bei der Vorverarbeitung interpolierten Regionen nicht als Merkmale in den Testdaten sichtbar sind. Diese Unterscheidung ist wichtig, damit sowohl die interpolierten Regionen als auch die reflektierenden Marker als Merkmale für die Algorithmen ausgeschlossen sind.

6.2.2. Detektion der Gesichtsregion

Während der Vorverarbeitung wird die Gesichtsregion aus jedem Bild ausgeschnitten. Die ausgeschnittenen Gesichtsregionen dienen als Eingangsbild für die unter-

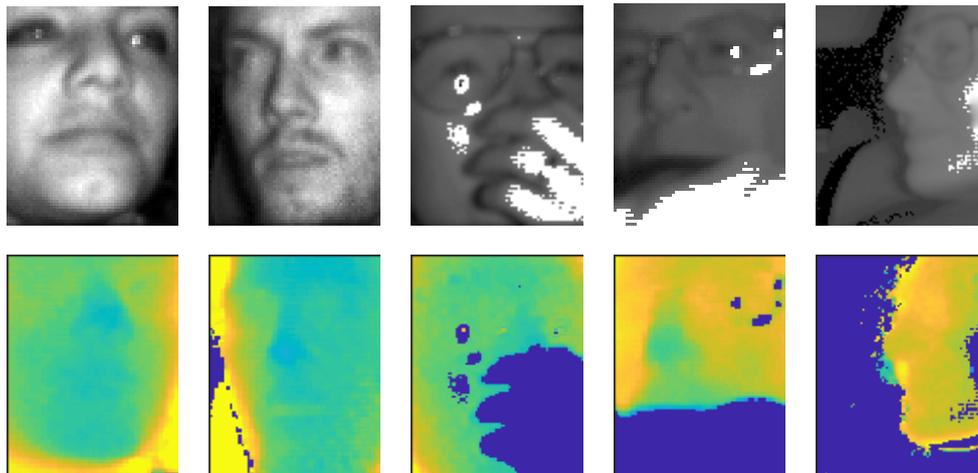


Abbildung 6.4.: Beispielbilder der detektierten Gesichtsregionen. Die erste Zeile zeigt die Infrarotbilder und die zweite Zeile die Tiefenbilder.

schiedlichen Verfahren, siehe Kapitel 6.3. Zur Detektion der Gesichtsregion wird das Verfahren von King (2009) angewendet. Für den Fall, dass kein Gesicht erkannt wird, wird die vorherige Gesichtsregion verwendet. Es wird dadurch angenommen, dass sich der Kopf von einem Bild zum nächsten nicht stark bewegt. Durch den Anwendungsbereich im Fahrzeuginnenraum kann dies angenommen werden. Abbildung B.1 visualisiert Beispielbilder der ausgeschnittenen Gesichtsregionen.

6.3. Angewendete Methoden

In diesem Abschnitt werden die Methoden zur Bestimmung der Kopfpose dargestellt, die anschließend auf dem *DriveAhead*-Datensatz evaluiert werden. Hierbei wird zwischen den aus der Literatur bekannten Methoden nach dem Stand der Forschung und den in dieser Arbeit vorgestellten Verfahren unterschieden.

Für die neuronalen Netze werden einheitlich skalierte Eingangsdaten verwendet. Die im vorherigen Kapitel detektierten Gesichtsregionen werden auf eine einheitliche Größe skaliert, damit die tiefen neuronalen Netze direkt angewendet werden können. Nach der Skalierung beträgt die einheitliche Größe der Gesichtsregionen 91×70 Pixel. Von diesen Gesichtsregionen werden zufällig Regionen mit einer Größe von 88×67 ausgeschnitten. Während diese skalierten Bilder als Eingangsdaten für die tiefen neuronalen Netze dienen, werden die anderen Verfahren direkt auf die im vorherigen Kapitel beschriebenen Gesichtsregionen angewendet.

Für alle Methoden werden nachfolgend die Details der Implementierung zusammengefasst.

6.3.1. Prior

Bei diesem Verfahren handelt es sich um eine einfache Methode ohne Training. Der *Prior* bestimmt, unabhängig von den Eingangsbildern, immer dieselbe Rotation und Translation. Es wird jeweils der Mittelwert der Rotation und Translation der Trainingsdaten ausgegeben. Für die Trainingsdaten des *DriveAhead*-Datensatzes ergeben sich für die Rotation folgende Winkel: *Gier* = $0,9^\circ$, *Nick* = $-15,3^\circ$, *Roll* = $-1,8^\circ$ und für die Translation: $t_x = 50,7$ mm, $t_y = -143,2$ mm, $t_z = 679,4$ mm. Es dient als *Baseline* und stellt die Schwierigkeit des Datensatzes dar.

6.3.2. Verfahren mit dem Stand der Forschung

Basierend auf dem Stand der Forschung aus Kapitel 2 wurden zwei Verfahren unter Berücksichtigung der Möglichkeit einer Nachimplementierung oder der Verfügbarkeit des Quellcodes ausgewählt. Während die eine Methode anhand der Positionen von Gesichtslanmarken aus Grauwertbildern die Kopfpose bestimmt, verwendet die andere ein tiefes neuronales Netz. Nachfolgend werden die zwei Verfahren genannt und Details zur Implementierung gegeben.

Openface (Baltrušaitis et al., 2016). Dieses Verfahren stellt den Stand der Forschung zur Bestimmung der Kopfpose aus 2D-Grauwerten dar. Es bestimmt die Kopfpose basierend auf den Positionen der Gesichtslanmarken. Zur Detektion der Gesichtslanmarken werden *Conditional Local Neural Field (CLNF)* (Baltrušaitis et al., 2012) angewendet. Anschließend wird die Kopfpose basierend auf einer Punkt-zu-Punkt-Übereinstimmung (Dementhon und Davis, 1995) berechnet. Für die Evaluation wird die öffentlich verfügbare Implementierung *Openface wild* verwendet, die auf zahlreichen RGB-Datensätzen, Multi-PIE (Gross et al., 2010), LFPW (Belhumeur et al., 2013) und Helen (Le et al., 2012), trainiert wurde. Um einen fairen Vergleich zu erhalten wurde eine Transformation bestimmt die das 3D Modell der Gesichtslanmarken in das in Kapitel 3.4.2 definierte Kopfkoordinatensystem transformiert.

N2 (Ahn et al., 2014). Das aus der Literatur bekannte Verfahren von Ahn et al. (2014) ist ein tiefes neuronales Netz zur Bestimmung der Kopfpose aus RGB-Bildern und erzielte Ergebnisse mit dem Stand der Forschung auf den Farbbildern des BIWI-Datensatzes (Fanelli et al., 2013). Im Vergleich zu dem im Rahmen dieser Arbeit entstandenen Netzwerk (siehe Kapitel 5) ist die Anzahl der *Convolutional Layer* gering, wie in Abbildung 2.2 schematisch anhand der Architektur des Netzwerkes zu sehen ist. Zur Evaluation dieses Verfahrens auf dem *DriveAhead*-Datensatz wurde es reimplementiert und auf Quaternionen angepasst. Als Eingangsdaten werden sowohl Infrarot- als auch Tiefendaten verwendet. Zusätzlich werden die Fusionsverfahren (siehe Kapitel 5) mit diesem Netzwerk als Basis implementiert.

6.3.3. Verfahren aus Kapitel 4 und 5

Die folgenden Methoden wurden im Rahmen dieser Arbeit entwickelt und in Kapitel 4 und Kapitel 5 vorgestellt.

HeHOP, Kapitel 4. Das in Kapitel 4 vorgestellte Verfahren zur Bestimmung der Kopfpose aus Tiefendaten wird in diesem Kapitel im Fahrzeuginnenraum evaluiert. Dieses effiziente Verfahren bestimmt die Kopfpose iterativ, beginnend mit einer frontalen Orientierung. Für jeden Schritt werden Zufallswälder auf lokalen Bereichen trainiert, um binäre Merkmale zu erhalten. Anschließend bestimmt eine globale lineare Transformation die Orientierung und Position des Kopfes. Die Parameter werden, wie in Kapitel 4 dargelegt, gewählt. Die Stufenanzahl beträgt $T = 9$, der Entscheidungswald besteht aus 11 Bäumen und die Baumtiefe beträgt 5.

HPN, Kapitel 5. Das in Kapitel 5 vorgestellte tiefe neuronale Netz wird ebenfalls auf dem *DriveAhead* Datensatz evaluiert. Ähnlich zu VGG (Simonyan und Zisserman, 2015) werden *Convolutions-Layers* mit Filtern der Größe 3×3 verwendet. Allerdings besteht das Netzwerk aus einer geringeren Anzahl von Parametern, da die Anzahl der neuronalen Filter halbiert wurde. Das Netzwerk besteht aus insgesamt 16 Convolutional und Fully Connected Layern mit Max Pooling Layern zwischen den Convolutional Layern. Es werden sowohl die Ergebnisse auf einer Modalität, Infrarot- oder Tiefendaten, präsentiert, als auch die Fusionsmöglichkeiten (siehe Kapitel 5). Bei den Fusionsverfahren handelt es sich um eine *frühe*, eine *späte* und eine *stitch-basierte* Fusion.

6.4. Orientierung des Kopfes

In diesem Kapitel werden die Ergebnisse der einzelnen Methoden für die Orientierung des Kopfes evaluiert. Die ersten beiden Abschnitte 6.4.1 und 6.4.2 diskutieren die Ergebnisse der Methoden auf dem gesamten Testdatensatz. Zum einen werden die unterschiedlichen Verfahren analysiert (siehe Abschnitt 6.4.1) und zum anderen die verwendeten Modalitäten (siehe Abschnitt 6.4.2). Bei den Modalitäten handelt es sich um Tiefen-, Nahinfrarotdaten oder die Fusion beider Modalitäten. Anschließend wird in Abschnitt 6.4.3 der Einfluss von Verdeckungen analysiert und in Abschnitt 6.4.4 der Einfluss von Brillen und Sonnenbrillen.

6.4.1. Evaluation der Methoden

Die erste Zeile von Tabelle 6.1 zeigt den *Balanced Mean Angular Error (BMAE)* des Priors. Als Prior ist die durchschnittliche Rotation der Trainingsdaten definiert, welche unabhängig vom Bild ausgegeben wird. Das in Kapitel 4 vorgestellte effiziente Verfahren, welches auf Tiefenbildern basiert, erreicht einen deutlich geringeren Fehlerwert des *BMAE* als der Prior. Im Vergleich dazu sind die Ergebnisse von

Methode	Modalität	BMAE in °
Prior	–	35.7
Openface* (Baltrušaitis et al., 2016)	Infrarotbild	20.6
N2 (Ahn et al., 2014)	Infrarotbild	19.2
HPN (Kapitel 5)	Infrarotbild	16.4
HeHOP (Kapitel 4)	Tiefenwerte	26.3
N2 (Ahn et al., 2014)	Tiefenwerte	15.9
HPN (Kapitel 5)	Tiefenwerte	14.2

* vor-trainierte Version von Baltrušaitis et al. (2016).

Tabelle 6.1.: Vergleich der unterschiedlichen Methoden. Die Tabelle zeigt den Balanced Mean Angular Error (*BMAE*) in ° der verschiedene Methoden für die Orientierung des Kopfes auf den gesamten Testdaten. Als Eingangsbild wird zwischen Infrarot- und Tiefenwerten unterschieden.

Openface (Baltrušaitis et al., 2016) mit einem *BMAE* von 20.6° noch besser, wobei das Verfahren auf mehreren RGB-Datensätzen trainiert wurde. Weitaus bessere Ergebnisse liefert das in Kapitel 5 vorgestellte tiefe neuronale Netz *HPN* (*Head Pose Network*) mit einem *BMAE* von 16.4° für Infrarot- und 14.2° für Tiefendaten. In den folgenden Abschnitten werden die einzelnen Verfahren abhängig von den Eingangsmodalitäten genauer untersucht. Hierfür werden zusätzlich zu dem in Tabelle 6.1 angezeigten *BMAE* Fehlermaß die Ergebnisse der Verfahren für die einzelnen Segmente untersucht.

Infrarotdaten. Abbildung 6.5 zeigt die Fehler der Infrarot-basierten Verfahren abhängig von der Kopfdrehung. Die Referenzwinkel sind in Segmente der Größe $d = 5^\circ$ geteilt und für jeden Winkelbereich $[m, m+5]$ wird der mittlere Winkelfehler $\Phi_{m,m+5}$ visualisiert. Die blaue Fläche zeigt die Verteilung der Trainingsdaten des *DriveA-Head*-Datensatzes. Daraus folgt, dass für große Kopforientierungen prozentual wenige Trainingsdaten vorhanden sind. Obwohl *Openface* (Baltrušaitis et al., 2016), in *orange* dargestellt, nicht auf den Trainingsdaten von *DriveAHead* trainiert wurde, ist auch bei diesem Verfahren ein deutlicher Anstieg des Fehlers für große Kopfdrehungen sichtbar. Das Verfahren basiert auf einer Detektion der Gesichtslandmarken, welche für starke Kopfdrehungen schwieriger wird. Im Vergleich zu diesem Verfahren liefern die beiden tiefen neuronalen Netze, N2 (Ahn et al., 2014) in *gelb* und HPN (siehe Kapitel 5) in *grün*, bessere Ergebnisse für große Kopfdrehungen. Bei einer Kopfdrehung zwischen 30 und 35 ° und bei einer Kopfdrehung zwischen 60 und 65 ° ist ein Anstieg des mittleren Winkelfehlers zu erkennen. Der zweite Anstieg zwischen 60 und 65 ° ist bei allen Verfahren zu sehen und lässt auf geringfügig schwierigere Szenen in diesem Winkelbereich schließen. Zur Analyse des ersten Anstiegs zwischen 30 und 35 ° zeigt Abbildung 5.5 in der letzten Spalte die mittlere *Heatmap* von Bildern dieses Winkelbereichs mit einem hohen Winkelfehler. In diesem Winkelbereich ist häufig ein Marker in der Nähe des Gesichts erkennbar, der zu einem hohen Feh-

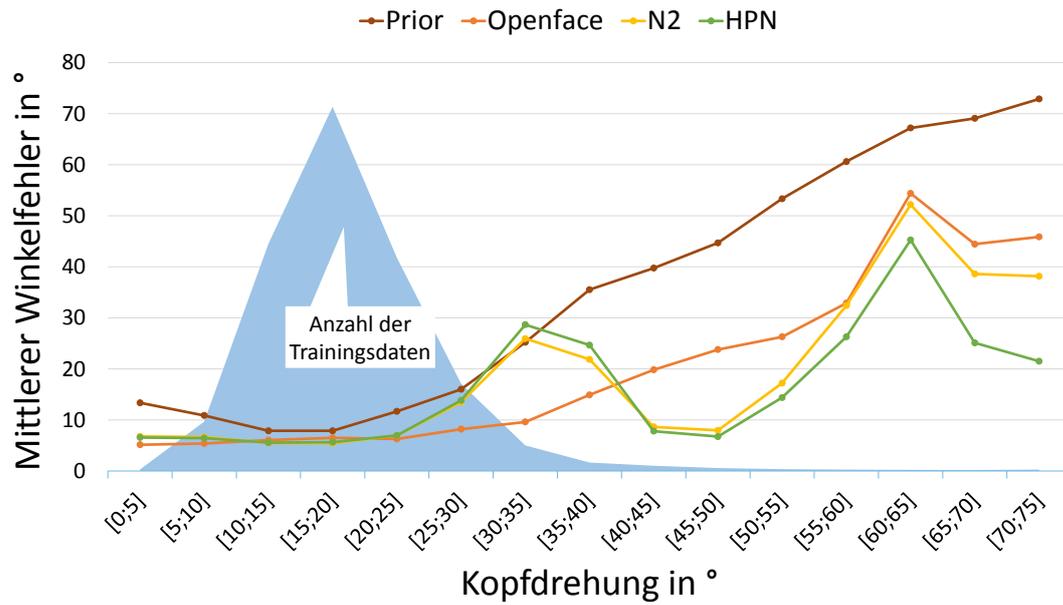


Abbildung 6.5.: Mittlerer Winkelfehler der Methoden basierend auf Infrarotbildern abhängig von der Kopfdrehung. Die Referenzwerte der Orientierung des Kopfes sind in Winkelabschnitte von 5° geteilt, für jeden Winkelabschnitt sind die mittleren Fehler der Verfahren basierend auf Infrarotdaten dargestellt. Im Hintergrund ist die Anzahl der Trainingsdaten in *hellblau* visualisiert.

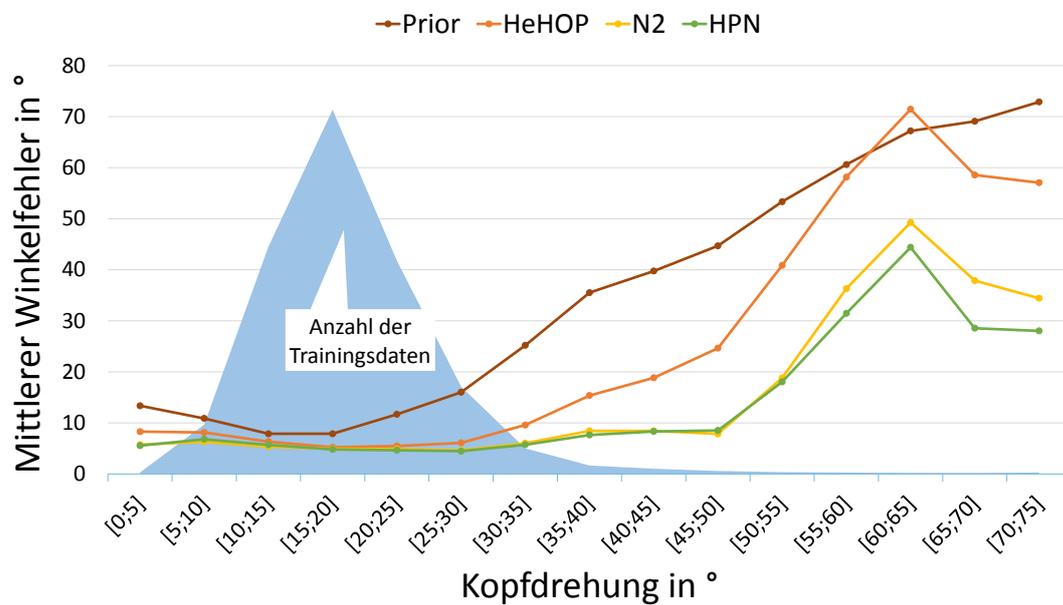


Abbildung 6.6.: Mittlerer Winkelfehler der Methoden basierend auf Tiefenbildern abhängig von der Kopfdrehung. Die Referenzwerte der Orientierung des Kopfes sind in Winkelabschnitte von 5° geteilt, für jeden Winkelabschnitt sind die mittleren Fehler der Verfahren basierend auf Tiefendaten dargestellt. Die *hellblaue* Fläche im Hintergrund zeigt die Anzahl der Trainingsdaten.

ler der Bestimmung der Kopforientierung aus Infrarotdaten führt. Aus der *Heatmap* ist erkennbar, dass bei Abdeckung des Markers der Fehler gering wird, was darauf hinweist, dass die tiefen neuronalen Netze bei Infrarotbildern sensibel gegenüber Veränderungen in der Nähe des Gesichts reagieren.

Tiefendaten. Abbildung 6.6 stellt den mittleren Winkelfehler $\Phi_{m,m+d}$ der Tiefenbasierten Verfahren abhängig von dem Rotationswinkel des Kopfes in Segmentgrößen d von 5° dar. Dabei befinden sich die Referenzwinkel für $\Phi_{m,m+5}$ im Winkelbereich $[m, m + 5]$. Die *hellblaue* Fläche zeigt die Anzahl der Trainingsdaten innerhalb der Winkelsegmente. In *rot* ist das Ergebnis des Priors dargestellt, bei dem kein Training abhängig von den Winkeln stattgefunden hat. Deshalb steigt der mittlere Winkelfehler des Priors für Winkelsegmente, die von dem mittleren Winkel der Trainingsdaten abweichen, an. Während für Segmente mit kleinen Winkeln, bei denen die Anzahl der Trainingsdaten hoch ist, das in Kapitel 4 vorgestellte effiziente Verfahren exakte Ergebnisse liefert, in *orange* dargestellt, steigt die Abweichung zu den Referenzwerten für große Kopfdrehungen. Im Vergleich zu diesem Verfahren generalisieren die beiden tiefen neuronalen Netze, N2 (Ahn et al., 2014) in *gelb* und HPN (siehe Kapitel 5) in *grün*, für große Kopfdrehungen weitaus besser. Der mittlere Winkelfehler für Verfahren angewendet auf Tiefendaten steigt im Bereich zwischen 30 und 35° nicht an, im Vergleich zu den Verfahren angewendet auf Infrarotdaten. In Abbildung 5.5 ist zu sehen, dass die neuronalen Netze angewendet auf Tiefendaten robust gegenüber Veränderungen in der Nähe des Gesichts reagieren.

6.4.2. Einfluss der Modalitäten

Für die Anwendung im Fahrzeuginnenraum ist von enormem Interesse, ob durch Verwendung oder Hinzunahme von Tiefenbildern die Ergebnisse im Vergleich zu Grauwert-basierten Verfahren verbessert werden kann. Der im Rahmen dieser Arbeit entstandene Datensatz zur Bestimmung der Kopfpose beinhaltet als erster Datensatz im Fahrzeuginnenraum Grauwertbilder und Tiefenbilder bei denen jedes Pixel übereinstimmt. Diese Eigenschaft ermöglicht eine direkte Vergleichbarkeit der einzelnen Modalitäten. Für einen fairen Vergleich wird im Folgenden dasselbe Verfahren mit denselben Freiheitsgraden sowohl auf den Infrarot-, als auch auf den Tiefendaten trainiert. Zusätzlich werden die Fusionsverfahren aus Kapitel 5 auf die gewählten Verfahren angewendet, um herauszufinden ob eine Fusion beider Modalitäten die Ergebnisse verbessert.

Eine geeignete Methode wird aus der Analyse des vorherigen Abschnitts gewählt. Sowohl auf Infrarot- als auch auf Tiefendaten erzielten tiefe neuronale Netze die besten Ergebnisse. Deshalb werden im Folgenden die beiden tiefen neuronalen Netze von Ahn et al. (2014) und das in Kapitel 5 vorgestellte Verfahren untersucht. Hierfür werden die tiefen neuronalen Netze jeweils mit derselben Anzahl von Filtern für die unterschiedlichen Eingangsbilder trainiert. Zuerst wird der Einfluss der einzelnen Modalitäten, Infrarot- und Tiefendaten, verglichen. Anschließend werden die

Methode	Modalität	$BMAE$ in $^\circ$
Prior	–	35.7
N2 (Ahn et al., 2014)	Infrarotbild	19.2
N2 (Ahn et al., 2014)	Tiefenbild	15.9
N2 (Ahn et al., 2014)	Frühe Fusion	19.0
N2 (Ahn et al., 2014)	Späte Fusion	16.7
HPN (Kapitel 5)	Infrarotbild	16.4
HPN (Kapitel 5)	Tiefenbild	14.2
HPN (Kapitel 5.3.1)	Frühe Fusion	17.4
HPN (Kapitel 5.3.2)	Späte Fusion	13.4
HPN (Kapitel 5.3.3)	Stitch Fusion	13.7

Tabelle 6.2.: *Balanced Mean Angular Error (BMAE)* in $^\circ$ auf den gesamten Testdaten bei Anwendung auf Infrarot- oder Tiefendaten, sowie der Fusion beider Modalitäten. Aus Schwarz et al. (2017) © 2017 IEEE.

in Kapitel 5 vorgestellten Fusionsmöglichkeiten angewendet, gemeint sind die *späte Fusion*, die **frühe Fusion** und die *stitch-basierte Fusion*.

Infrarot- versus Tiefenbilder. Tabelle 6.2 zeigt die Ergebnisse der tiefen neuronalen Netze allein auf Infrarot- oder Tiefenbildern ausgewertet. Für alle Modalitäten sind die Ergebnisse des in Kapitel 5 vorgestellten *Head Pose Network (HPN)* besser als die des tiefen neuronalen Netzes von Ahn et al. (2014). Bei den einzelnen Modalitäten sind die Ergebnisse auf Tiefendaten deutlich besser als auf Infrarotdaten, mit einem Unterschied von 3.3° des $BMAE$ für das N2 Netzwerk (Ahn et al., 2014) und 2.2° für das in Kapitel 5 vorgestellte HPN Netzwerk. Insgesamt ergibt sich daraus, dass aus Tiefendaten eine etwas genauere Bestimmung der Orientierung des Kopfes möglich ist, als aus Grauwerten eines Infrarotbildes.

Fusionsverfahren. Die in Kapitel 5.3 vorgestellten Fusionsmöglichkeiten ermöglichen eine weitere Verbesserung der Ergebnisse in Tabelle 6.2. Allerdings liefert die in Kapitel 5.3.1 vorgestellte *frühe Fusion* einen höheren $BMAE$. Im Gegensatz dazu verbessert die in Kapitel 5.3.2 vorgestellte *späte Fusion* die Ergebnisse. Mit der *stitch-basierten Fusion* (siehe 5.3.3) sind die Ergebnisse vergleichbar mit einem niedrigen $BMAE$ -Wert von 13.7° zu dem $BMAE$ der *späten Fusion* mit 13.4° . Deshalb werden im Folgenden diese beiden Verfahren noch genauer anhand von unterschiedlichen Einflüssen untersucht. Hierbei wird untersucht, wie gut die Performance der Methoden bei Personen mit Verdeckungen im Vergleich zu Personen ohne Verdeckungen ist. Des Weiteren werden die Ergebnisse auf Gesichtern mit Brillen, Sonnenbrillen und ohne Brillen verglichen.

Methode	Modalität	<i>BMAE</i> in °	
		Keine Verdeckungen	Mit Verdeckungen
Prior	–	34.4	37.7
HPN (Kapitel 5.3.2)	Späte Fusion	9.2	17.0
HPN (Kapitel 5.3.3)	Stitch Fusion	12.3	16.0

Tabelle 6.3.: Einfluss von Verdeckungen. *Balanced Mean Angular Error (BMAE)* in ° der späten Fusion und frühen Fusion auf Gesichtern mit und ohne Verdeckung.

6.4.3. Einfluss von Verdeckungen

Der *DriveAhead*-Datensatz beinhaltet manuelle Notationen für verdeckte Gesichter, wobei eine Verdeckung als jegliche Verdeckung durch andere Objekte innerhalb des Gesichts definiert ist. Hierbei zählen allerdings Brillen, Sonnenbrillen oder nicht sichtbare Gesichtsbereiche aufgrund von starken Kopfdrehungen nicht als verdeckt, siehe Kapitel 3.4.3. Diese Notationen ermöglichen es, den Einfluss von Verdeckungen auf die Genauigkeit der Algorithmen zu analysieren. Im vorherigen Abschnitt zeigten die beiden *Head Pose Network*-Netzwerke (*HPN*) mit *später Fusion* aus Kapitel 5.3.2 und mit *stitch-basierte Fusion* aus Kapitel 5.3.3 die besten Ergebnisse. Deshalb wird im Folgenden für das *HPN* mit *später Fusion* und das *HPN* mit *stitch-basierte Fusion* der Einfluss von Verdeckungen im Gesichtsbereich untersucht.

Zusätzlich zu den Ergebnissen der beiden Fusionsverfahren, *HPN* mit *später Fusion* und *stitch-basierte Fusion*, zeigt Tabelle 6.3 die Ergebnisse des *Priors* auf Testdaten mit und ohne Verdeckungen. Die Fehler des *Priors* spiegeln die Verteilung der Testdaten mit und ohne Verdeckungen wieder. Die Ergebnisse des *Priors* zeigen einen Anstieg von 2° des *BMAE* für Gesichter mit Verdeckungen im Vergleich zu Gesichtern ohne Verdeckungen. Aus diesem geringen Unterschied folgt, dass Gesichter mit Verdeckungen im Durchschnitt eine 2° höhere Abweichung von der mittleren Rotation der Trainingsbilder haben als Gesichter ohne Verdeckungen.

Im Vergleich zu den Ergebnissen des *Priors* beeinflussen Verdeckungen die Genauigkeiten der tiefen neuronalen Netze stark. Während es bei dem *Head Pose Network (HPN)* mit *später Fusion* zu einem 7, 8° höheren *BMAE* durch Verdeckungen im Gesicht kommt, steigt der *BMAE* für das *HPN* mit *stitch-basierte Fusion* nur um 3, 7° an. Insgesamt liefert bei Verdeckungen das *HPN* mit *stitch-basierte Fusion* einen *BMAE* von 16° im Vergleich zu dem *HPN* mit *später Fusion*, welches einen *BMAE* von 17° erreicht. Daraus folgt, dass durch das Training mit *stitch-basierten* Gewichten ein tiefes neuronales Netzwerk entsteht, welches im Fall von Verdeckungen besser generalisieren kann.

Methode	Modalität	<i>BMAE</i> in °		
		Mit Brillen	Mit Sonnenbrillen	Ohne
Prior	–	33.8	36.5	36.3
HPN (Kapitel 5.3.2)	Späte Fusion	9.7	17.8	9.4
HPN (Kapitel 5.3.3)	Stitch Fusion	10.9	18.4	11.4

Tabelle 6.4.: Einfluss von Brillen und Sonnenbrillen. *Balanced Mean Angular Error (BMAE)* in ° der *späten Fusion* und *frühen Fusion* von Bildern bei denen der Fahrer eine Brille, eine Sonnenbrille oder keines von beiden trägt.

6.4.4. Einfluss von Brillen und Sonnenbrillen

Eine Abschätzung der Blickrichtung anhand der Kopforientierung ist genau dann von Interesse, wenn die Augen oder die Pupillen der Fahrer nicht sichtbar sind. Dies kommt insbesondere vor, wenn der Fahrer eine Brille oder Sonnenbrille trägt. Auch wenn diese durchsichtig ist und damit die Grauwerte der Augen sichtbar sind, kann es zu Reflexionen kommen, die eine exakte Bestimmung der Blickrichtung nicht ermöglichen. Der *DriveAhead*-Datensatz enthält neben Notationen für Verdeckungen auch Notationen, ob der Fahrer in dem Bild eine Brille oder Sonnenbrille trägt. Damit ermöglicht dieser Datensatz eine Analyse der Verfahren auf die Robustheit gegenüber Brillen und Sonnenbrillen. Alle Fahrer tragen unterschiedliche Modelle von Brillen und Sonnenbrillen. Hierbei erscheinen einige Modelle der Sonnenbrillen in den Infrarot- und Tiefenbildern als nahezu durchsichtig, siehe Anhang B. Der folgende Abschnitt untersucht den Einfluss von Brillen und Sonnenbrillen auf die Ergebnisse der Methoden.

Der *Prior* in der ersten Zeile von Tabelle 6.4 gibt ein Maß für die Rotationen in den Testdaten bei denen der Fahrer eine Brille, eine Sonnenbrille oder keine Brille trägt. Er zeigt für Personen mit Brillen einen um 3° geringeren *BMAE* als für Personen ohne Brillen oder mit Sonnenbrillen. Daraus folgt, dass Personen, die eine Brille tragen, durchschnittlich den Kopf geringfügig weniger von der mittleren Rotation der Trainingsdaten wegbewegen als diejenigen, die eine Sonnenbrille oder keine Brille tragen.

Die tiefen neuronalen Netze, *HPN* mit *später Fusion* und *stitch-Fusion* zeigen in Tabelle 6.4 beide für Personen mit Brillen signifikant bessere Ergebnisse als für Gesichter mit Sonnenbrillen. Im Vergleich zu Gesichtern ohne Brillen, ist der Fehler auf Gesichtern mit Brillen sogar für die *stitch-basierte Fusion* geringer. *HPN* mit einer *stitch-basierte Fusion* erreicht sogar einen um 0.5° besseren *BMAE* für Gesichter mit Brillen als ohne. Um dieses Merkmal genauer zu analysieren, werden im Folgenden die Winkelfehler der einzelnen Segmente abhängig von der Kopffrotation betrachtet.

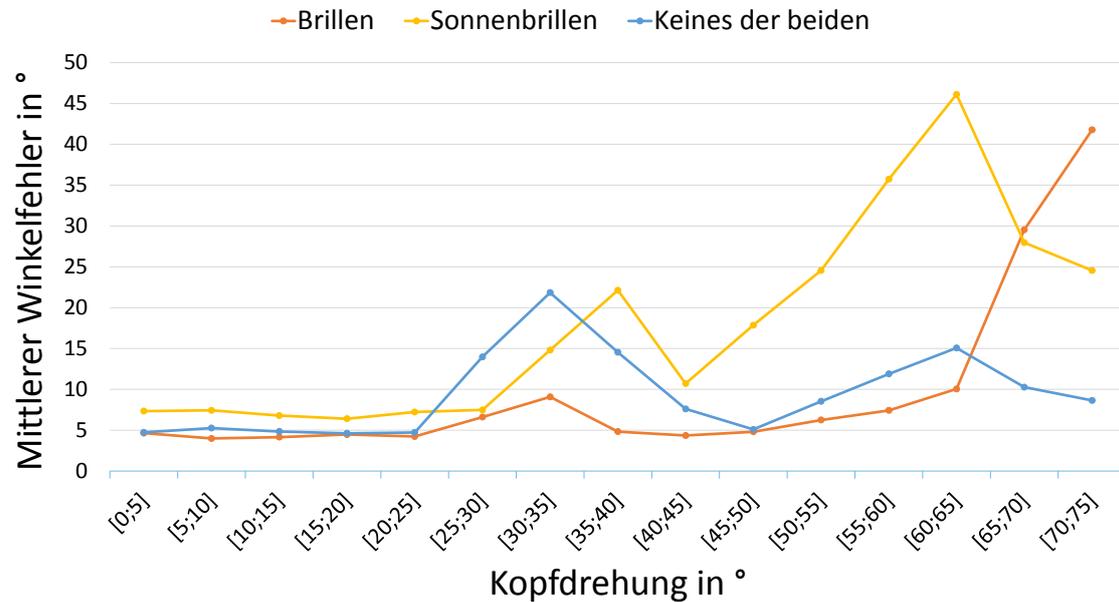


Abbildung 6.7.: Die Ergebnisse von HPN mit *später Fusion* abhängig von der Kopfdrehung für Fahrer mit Brille, Sonnenbrille und ohne Brille. Dabei ist die Kopfdrehung in Intervalle von 5° eingeteilt und für jedes Intervall ist der mittlere Winkelfehler dargestellt. Aus Schwarz et al. (2017) © 2017 IEEE.

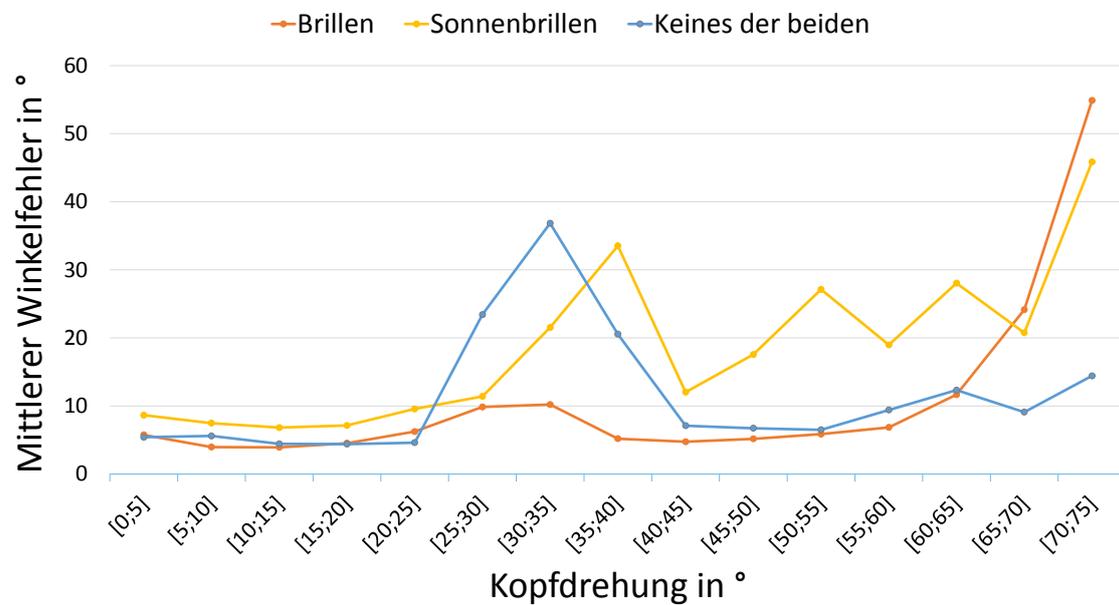


Abbildung 6.8.: Die Ergebnisse von HPN mit *stitch-Fusion* abhängig von der Kopfdrehung für Fahrer mit Brille, Sonnenbrille und ohne Brille. Dabei ist die Kopfdrehung in Intervalle von 5° eingeteilt und für jedes Intervall ist der mittlere Winkelfehler dargestellt.

Abbildung 6.7 und 6.8 zeigen die mittleren Winkelfehler des *HPN* mit der *späten Fusion* und der *stitch-Fusion* abhängig von der Rotation des Kopfes. Beide Grafiken zeigen den mittleren Winkelfehler der Gesichter mit Brillen, Sonnenbrillen und ohne Brille abhängig von der Kopfdrehung. Während bei der *späten Fusion* der Fehler bei den Gesichtern mit Sonnenbrillen für große Kopfdrehungen deutlich ansteigt, schwankt der mittlere Winkelfehler bei der *stitch-Fusion* unkorreliert. Dadurch wird deutlich, dass bei der *stitch-basierten Fusion* andere Merkmale innerhalb der Gesichtsregionen zum Ergebnis der Kopforientierung beitragen als bei den anderen Fusionsverfahren. Die *Heatmaps* in Abbildung 5.5 von Beispielen des *DriveA-Head*-Datensatzes visualisieren in rot welche Regionen des Gesichts den Fehler der Kopforientierung verringern. Die *Heatmaps* verdeutlichen, dass bei den verschiedenen Fusionsverfahren unterschiedliche Bereiche des Gesichts den Fehler verringern.

Bei beiden Verfahren sind für große Rotationen zwischen 40 und 65° die Winkelfehler für Fahrer mit Sonnenbrillen größer als für Fahrer ohne oder mit Brillen. Dieser Effekt zeigt, dass die Augenregion ein wichtiges Merkmal zur akkuraten Bestimmung von starken Kopfdrehungen ist. Die Winkelfehler von Gesichtern mit Brillen sind über fast den gesamten Drehbereich geringer als die der Personen ohne Brillen. Dieser interessante Effekt zeigt, dass sowohl die Rahmen der Brillen als auch die Augen als Merkmale für die tiefen neuronalen Netze agieren. Durch die Sichtbarkeit beider Merkmale wird die Genauigkeit der Verfahren erhöht.

6.5. Translation des Kopfes

Im Fahrzeuginnenraum ist die Position des Kopfes aus Sicherheitsgründen und zur Bestimmung der Blickrichtung von großem Interesse. Deshalb beinhaltet der *DriveAHead*-Datensatz, neben Referenzmessungen für die Orientierung des Kopfes, die Position des Kopfes. Mithilfe der Referenzmessung und den zusätzlichen Notationen für Verdeckungen wird die Bestimmung der Translation des Kopfes in diesem Abschnitt analysiert.

Die Genauigkeit der Positionsbestimmung des Kopfes von zwei Methoden wird im Folgenden untersucht. Hierfür werden die Ergebnisse von einem Verfahren basierend auf Infrarotdaten und einem basierend auf Tiefendaten evaluiert. Das Verfahren von Baltrušaitis et al. (2016) verwendet die Grauwerte der Infrarotbilder während das in Kapitel 4 vorgestellte Verfahren Tiefendaten verwendet. Beide Methoden schätzen neben der Kopforientierung die Translation des Kopfes. Bei der Untersuchung werden die einzelnen Translationsrichtungen x , y und z getrennt betrachtet. Abschnitt 6.5.1 zeigt die Ergebnisse der gesamten Testdaten, Abschnitt 6.5.2 diskutiert den Einfluss von Verdeckungen und Abschnitt 6.5.3 den Einfluss von Brillen und Sonnenbrillen auf das Ergebnis.

Methode	Modalität	Translationsfehler in mm		
		x	y	z
Prior	–	30.9	20.2	36.8
Openface ^a (Baltrušaitis et al., 2016)	Infrarotbild	6.4	7.6	27.8
HeHOP, Kapitel 4	Tiefenbild	4.1	3.6	5.3

^a vor-trainierte Version von Baltrušaitis et al. (2016).

Tabelle 6.5.: Einfluss der Modalitäten auf die Bestimmung der Translation. Euklidische Distanz in *mm* der geschätzten Translation und der Referenzmessung. Die Tabelle zeigt die Ergebnisse gemittelt über alle Testdaten für den Prior, Openface (Baltrušaitis et al., 2016) und HeHOP, siehe Kapitel 4.

6.5.1. Einfluss der Modalitäten

Auf den Testdaten wird die Position des Kopfes von einem Verfahren basierend auf Infrarotdaten (Baltrušaitis et al., 2016) und einem Verfahren, das Tiefendaten verwendet, *HeHOP* vorgestellt in Kapitel 4, evaluiert. Während die Methode ausgewertet auf Infrarotdaten auf zahlreichen RGB-Datensätzen trainiert ist, fand das Training der Parameter für das tiefen-basierte Verfahren auf den Trainingsdaten des *DriveAHead*-Datensatzes statt. Tabelle 6.6 zeigt, dass das 2D-basierte Verfahren von Baltrušaitis et al. (2016) in x- und y-Richtung vergleichbare Positionen zu dem tiefen-basierten Verfahren aus Kapitel 4 liefert. Für beide Verfahren ist die Bestimmung der z-Translation am schwierigsten. Diese Translation entspricht der Entfernung des Kopfes zum Sensor. Da Tiefendaten direkte Informationen über die 3D-Form des Kopfes liefern, erreicht das regressions-basierte Verfahren aus Kapitel 4, welches auf Tiefendaten angewendet wird, deutlich bessere Ergebnisse als die Methode auf Infrarotdaten.

6.5.2. Einfluss von Verdeckungen

Aufgrund der Verfügbarkeit von Notationen für Verdeckungen im *DriveAHead*-Datensatz ist es, wie für die Orientierung des Kopfes, auch für die Translation des Kopfes möglich, den Einfluss von Verdeckungen auf die Genauigkeit der Verfahren auszuwerten. Ein Gesicht wird als verdeckt bezeichnet, sobald es zu einer Verdeckung im Gesichtsbereich kommt, wie in Kapitel 3.4.3 beschrieben. Sonnenbrillen, Brillen und Selbstverdeckungen durch starke Kopffrotationen gelten nicht als Verdeckung.

Tabelle 6.6 zeigt die Genauigkeit der Translation abhängig von Verdeckungen. Die Fehlerwerte des Priors in der ersten Zeile zeigen deutlich höhere Werte für Gesichter mit Verdeckungen als ohne Verdeckungen in z-Richtung an. Daraus folgt, dass eine Verdeckung häufig mit einer Translation in z-Richtung verbunden ist. Das

Methode	Translationsfehler in mm					
	Ohne Verdeckung			Mit Verdeckung		
	x	y	z	x	y	z
Prior	27.8	19.7	30.2	37.8	21.4	51.7
Openface ^a (Baltrušaitis et al., 2016)	5.6	8.3	19.2	8.5	6.1	47.4
HeHOP (Kapitel 4)	3.3	3.4	4.7	5.9	4.2	6.7

^a vor-trainierte Version von Baltrušaitis et al. (2016).

Tabelle 6.6.: Einfluss von Verdeckungen auf die Bestimmung der Translation. Euklidische Distanz in Millimetern zwischen geschätzter Translation und der Referenzmessung. In der ersten Spalte ist der Fehler für Testdaten ohne Verdeckungen dargestellt. Die zweite Spalte zeigt den Fehler bei verdeckten Gesichtern.

2D-basierte Verfahren von Baltrušaitis et al. (2016) erreicht signifikant bessere Ergebnisse auf den Gesichtern ohne Verdeckungen mit einem Unterschied von mehr als 28 Millimetern in z-Richtung. Im Gegensatz dazu ist die in Kapitel 4 vorgestellte regressions-basierte Methode robuster gegenüber Verdeckungen, mit einem Unterschied von nur 2 Millimetern in z-Richtung gegenüber Gesichtern ohne Verdeckungen.

6.5.3. Einfluss von Brillen und Sonnenbrillen

Die Bestimmung der Blickrichtung von Personen mit Brillen oder Sonnenbrillen stellt eine große Herausforderung für Blickrichtungs-Algorithmen dar, da es zu Reflektionen in den Augenbereichen kommen kann. Deshalb ist insbesondere eine robuste Bestimmung der Kopfpose bei Brillen oder Sonnenbrillen von großem Interesse für eine grobe Abschätzung der Blickrichtung. Mit den Notationen des *DriveAHead*-Datensatzes, ob der Fahrer eine Brille oder Sonnenbrille trägt, wird im Folgenden der Einfluss von Brillen und Sonnenbrillen auf die Genauigkeit der Translation bestimmt.

In Tabelle 6.7 ist der Fehler der Translation des Kopfes für Gesichter mit Brille, Sonnenbrille und ohne Brille dargestellt. Die Ergebnisse des *Priors* zeigen, dass im Durchschnitt Fahrer mit Brille den Kopf in y- und z- Richtung weniger weit weg vom Mittelwert der Trainingsdaten bewegt haben als Fahrer mit Sonnenbrille oder ohne Brille. Allerdings ist die Bewegung in x-Richtung vergleichbar mit den Personen mit Sonnenbrillen. Das 2D-basierte Verfahren von Baltrušaitis et al. (2016) hat für Brillen und Sonnenbrillen in z-Richtung ähnliche Ergebnisse wie der *Prior*, allerdings werden die x- und y-Richtung besser erkannt. Ohne Brille und Sonnenbrille sind die Ergebnisse für diese Methode in z-Richtung signifikant besser als die des *Priors*. Das *HeHOP*-Verfahren, vorgestellt in Kapitel 4, ist gegenüber Sonnenbrillen und Brillen deutlich robuster. Der Verlust der Genauigkeit in z-Richtung für Gesichter

Method	Translationsfehler in mm								
	Brillen			Sonnenbrillen			Ohne		
	x	y	z	x	y	z	x	y	z
Prior	31.8	15.2	20.7	36.0	21.0	43.5	26.3	23.0	42.4
Openface ^a (Baltrušaitis et al., 2016)	4.6	10.2	21.2	7.9	8.3	37.3	6.6	5.5	24.9
HeHOP, Kapitel 4	3.1	3.7	3.9	5.5	4.3	6.8	3.7	3.1	5.1

^a vor-trainierte Version von Baltrušaitis et al. (2016).

Tabelle 6.7.: Einfluss von Brillen und Sonnenbrillen auf die Bestimmung der Translation. Euklidische Distanz in *mm* der geschätzten Translation und der Referenzmessung.

mit Sonnenbrille ist mit 2 Millimetern gegenüber der Genauigkeit der z-Richtung von Gesichtern ohne Brille oder Sonnenbrille sehr gering. Bei Fahrern mit Brille im Vergleich zu denen ohne Brille ist die Genauigkeit nahezu identisch. Daraus folgt, dass die *HeHOP*-Methode sehr robust gegenüber Brillen und Sonnenbrillen ist.

6.6. Fazit

In diesem Kapitel wurde eine ausführliche Evaluation verschiedener Verfahren zur Bestimmung der Kopfpose im Fahrzeug auf den Testdaten des *DriveAhead*-Datensatzes durchgeführt. Dabei wurden sowohl für die Orientierung des Kopfes als auch für die Translation des Kopfes vorhandene Verfahren nach dem Stand der Technik (Ahn et al., 2014; Baltrušaitis et al., 2016) mit den im Rahmen dieser Arbeit entstandenen Verfahren aus den Kapiteln 4 und 5 verglichen. Bei der Analyse wurden unterschiedliche Einflüsse auf die Ergebnisse der Verfahren im Detail untersucht und bewertet.

Für die *Orientierung des Kopfes* erzielt das in Kapitel 5 vorgestellte Head Pose Network (HPN) im Vergleich zu den Verfahren von Ahn et al. (2014) und Baltrušaitis et al. (2016) sowie dem in Kapitel 4 vorgestellten effizienten Verfahren die höchste Genauigkeit. Tiefendaten als Eingangswerte liefern bessere Ergebnisse mit einem *Balanced Mean Angular Error (BMAE)* von $14,2^\circ$ im Vergleich zu Infrarotdaten mit einem *BMAE* von $16,4^\circ$. Die in Kapitel 5.3 vorgestellten Fusionsverfahren, der *späten Fusion* und *stitch-Fusion*, verbessern den *BMAE* auf $13,4^\circ$ und $13,7^\circ$. Verdeckungen in der Gesichtsregion haben einen großen Einfluss auf die Genauigkeit der Bestimmung der Orientierung des Kopfes mit tiefen neuronalen Netzen. Bei der Bestimmung der Orientierung auf Gesichtern ohne Verdeckungen gegenüber denen mit Verdeckungen verschlechtert sich der *BMAE* um $7,8^\circ$ bei Anwendung des HPN mit *später Fusion*. Dagegen ist das HPN mit *stitch-Fusion* robuster gegenüber Verdeckungen mit einem Genauigkeitsverlust von $3,7^\circ$. Die Analyse der Gesichter mit Brille, Sonnenbrille gegenüber ohne Brille hat gezeigt, dass die neuronalen Netze die

Brillenrahmen als zusätzliche Merkmale verwenden und sich dadurch die Ergebnisse der Fahrer mit Brille gegenüber jenen ohne Brille verbessern können.

Bei der Bestimmung der *Translation des Kopfes* liefert das in Kapitel 4 vorgestellte Verfahren basierend auf Tiefendaten bessere Ergebnisse als das Verfahren von Baltrušaitis et al. (2016), welches auf Basis der Infrarotdaten evaluiert wurde. Die x- und y- Translation wurde mit ähnlicher Genauigkeit bestimmt, während bei der z-Translation das auf der Basis von Tiefenbildern ausgewertete Verfahren signifikant bessere Ergebnisse lieferte. Die in Kapitel 4 vorgestellte Methode wird durch Verdeckungen im Gesichtsbereich mit einem Unterschied von 2 Millimetern in alle Richtungen wenig beeinflusst während das Verfahren von Baltrušaitis et al. (2016) in z-Richtung einen um 28,2 Millimeter höheren Fehler liefert. In Bezug auf Brillen und Sonnenbrillen liefert das in Kapitel 4 vorgestellte Verfahren ebenfalls robuste Ergebnisse für die Translation des Kopfes. Insgesamt zeigt das in Kapitel 4 vorgestellte regressions-basierte Verfahren damit eine hohe Genauigkeit für die Translation des Kopfes im Fahrzeuginnenraum bei einer Robustheit gegenüber Verdeckungen im Gesichtsbereich sowie gegenüber Brillen und Sonnenbrillen.

Kapitel 7

Zusammenfassung

Die Kopfpose des Fahrers ist eine Schlüsselinformation für moderne Fahrerassistenz- und Sicherheitssysteme, da aus ihr die Haltung und das Verhalten des Fahrers abgeleitet werden können. In dieser Arbeit wurden ein Datensatz und Algorithmen zur Bestimmung der Kopfpose im Fahrzeuginnenraum entwickelt. Neben einer Bewertung der Güte der Kopfpose wurde der Einfluss verschiedener Verdeckungen des Gesichts durch Sonnenbrillen, Brillen und weitere Verdeckungen analysiert. Dabei kamen als Eingangsdaten Time-Of-Flight-Bilder (TOF), eines im Fahrzeuginnenraum verbauten Sensors zum Einsatz. Dieser lieferte sowohl Tiefen- als auch Infrarotbilder, wobei letztere Grauwertbildern entsprechen. Im folgenden Abschnitt werden die Beiträge der Arbeit zusammengefasst und ein Ausblick über mögliche zukünftige Forschungsfelder gegeben.

In der vorliegenden Arbeit wurde der *DriveAhead*-Datensatz während realer Fahrten im Fahrzeuginnenraum aufgenommen zur Evaluation von Algorithmen für die Bestimmung der Kopfpose aus Tiefen- und Infrarotdaten. Um eine Vergleichbarkeit zwischen unterschiedlichen Personen herzustellen, wurde ein anhand von Gesichtsländemarken eindeutig für jede Person definiertes Kopfkoordinatensystem erstellt. Die verwendete Referenz-Kopfpose gibt die Orientierung und Position dieses Koordinatensystems für jedes Bild an. Zur Bestimmung der Bewegung des Koordinatensystems wurde ein akkurates Motion-Capture-System verwendet. Der Datensatz stellt den ersten im Fahrzeug während realer Fahrten aufgenommenen Kopfposen-Datensatz mit Tiefendaten dar, der bisher in der Wissenschaft veröffentlicht wurde. Er beinhaltet sowohl Infrarot- als auch Tiefendaten und Referenzwerte des Kopfes für jedes Bild. Zusätzlich zur Referenz-Kopfpose beinhaltet jedes Bild des Datensatzes manuelle Notationen, die angeben, wann der Fahrer eine Brille oder Sonnenbrille getragen hat sowie, ob weitere Verdeckungen im Gesichtsbereich vorhanden waren. Der *DriveAhead*-Datensatz beinhaltet mit insgesamt mehr als einer Million Bildern von 20 unterschiedlichen Probanden eine weitaus größere Datenmenge als bisher vorhandene Kopfposen-Datensätze. Dadurch eröffnet sich ein neues Forschungsfeld der Kopfposen-Bestimmung aus Tiefen- und Infrarotdaten sowie der Kombination beider Modalitäten im Fahrzeuginnenraum.

Zur Bestimmung der Kopfpose aus Tiefendaten wurde ein regressions-basierter Algorithmus entwickelt. Das Verfahren berechnet die Orientierung und Position des Kopfes stufenweise, indem mit jeder Stufe die vorherige Orientierung und Position aktualisiert wird. Beginnend mit lokalen binären Eigenschaftsvektoren an Regionen bestimmt eine globale Regressionsmatrix die Aktualisierung der Orientierung und Position des Kopfes. Für die Erzeugung der binären Eigenschaftsvektoren wurden Entscheidungsbäume an lokalen Regionen im Gesicht trainiert, während die globale Regressionsmatrix mithilfe eines Minimierungsverfahrens in der Trainingsphase optimiert wurde. Bei der Anwendung des Verfahrens auf unbekannte Tiefenbilder erzeugen die Entscheidungsbäume dünnbesetzte binäre Vektoren, die anschließend mit einer Matrix multipliziert werden. Sowohl die Entscheidungsbäume als auch die Matrixmultiplikation benötigen nur wenige Rechenschritte, was insbesondere für die Anwendung im Fahrzeug bei limitierten Rechenkapazitäten ein enormer Vorteil ist. Damit stellt sich die Methode mit vergleichbaren Ergebnissen zum Stand der Forschung bei weniger Rechenzeit für die Anwendung im Automobil-Kontext als vorteilhaft dar.

Der Einfluss der Eingangsdaten auf die Bestimmung der Kopforientierung wurde untersucht, indem eine tiefe neuronale Netzstruktur auf der Basis von Tiefen- oder Infrarotbildern ausgewertet wurde und zur Kombination beider Modalitäten erweitert wurde. Hierfür wurde eine tiefe neuronale Netzstruktur, die auf dem aktuellen Stand der Forschung basiert, auf die Ermittlung der Kopforientierung angepasst. Mit dieser Architektur wurde die Kopforientierung sowohl aus einzelnen Infrarot- und Tiefenbildern bestimmt als auch auf der Kombination von beiden Modalitäten, indem neue Fusionsverfahren entwickelt wurden. Es wurde eine *frühe Fusion* vorgestellt, bei der bereits die Eingangsdaten kombiniert werden. Außerdem wurde eine *späte Fusion* präsentiert bei der als Ausgabe der Durchschnitt aus den tiefen neuronalen Netzen, angewendet auf das Infrarot- und auf das Tiefenbild, gebildet wird. Zusätzlich wurde eine *stitch-basierte Fusion* entwickelt bei der bereits während des Trainings ein Austausch zwischen den neuronalen Netzen basierend auf Tiefen- und Infrarotdaten ermöglicht wurde. Die Verwendung derselben Basis-Architektur in allen Verfahren ermöglicht eine direkte Vergleichbarkeit der unterschiedlichen Fusionsverfahren mit einem Netzwerk basierend entweder auf Infrarot- oder Tiefenbildern.

Die im Rahmen dieser Arbeit vorgestellten Algorithmen zur Bestimmung der Kopfpose wurden anhand des selbst aufgenommenen *DriveAhead*-Datensatzes während realer Fahrten im Fahrzeuginnenraum ausführlich analysiert. Das vorgestellte regressions-basierte Verfahren und die präsentierten tiefen neuronalen Netze wurden mit alternativen Algorithmen, die den Stand der Forschung repräsentieren, verglichen. Dabei erzielt das in der vorliegenden Arbeit vorgestellte tiefe neuronale Netz signifikant genauere Ergebnisse für die Kopforientierung als andere Verfahren. Während das tiefe neuronale Netz auf der Basis von Tiefendaten im Vergleich zur Anwendung auf Infrarotdaten eine höhere Genauigkeit erreicht, bieten die vorgestellten Fusionsverfahren weitaus genauere Ergebnisse. Insgesamt erzielte die Kombination der Modalitäten mit der *späten Fusion* die besten Ergebnisse. Allerdings zeigt sich die *stitch-basierte Fusion* robuster gegenüber Verdeckungen. Trägt der Fahrer eine

Sonnenbrille, beeinflusst dies das Verfahren, während der Rahmen von Brillen als zusätzliche Eigenschaft die Performance der Methoden sogar verbessert. Bei der Bestimmung der Translation des Kopfes zeigt das im Rahmen dieser Arbeit vorgestellte Verfahren, welches auf Tiefenbildern basiert, im Vergleich zu einem auf Grauwerten basierenden Verfahren bei der Abschätzung der Entfernung des Kopfes zum Sensor weitaus bessere Ergebnisse. Die Bestimmung der x- und y- Translation der beiden Verfahren ist vergleichbar. Brillen beeinflussen dieses regressions-basierte Verfahren nicht. Sonnenbrillen und Verdeckungen im Gesichtsbereich führen zu einem geringen Fehler von nur 2 Millimeter in z-Richtung.

Zusammenfassend analysiert diese Arbeit die Bestimmung der Kopfpose im Fahrzeuginnenraum, indem ein Datensatz und Algorithmen vorgestellt wurden. Für die Anwendung im Fahrzeuginnenraum wurde eine Referenzbestimmung der Orientierung und Position des Kopfes während realer Fahrten entwickelt. Mit den damit aufgenommenen Daten hat sich ergeben, dass Brillen im Gegensatz zu Sonnenbrillen und Verdeckungen im Gesichtsbereich für Algorithmen der Bildverarbeitung eine geringere Herausforderung darstellen. Mit dem im Rahmen dieser Arbeit vorgestellten regressions-basierten Verfahren kann die Position des Kopfes im Fahrzeuginnenraum mit wenig Rechenaufwand bis auf einen Tiefenfehler von durchschnittlich 5 Millimeter genau bestimmt werden. Die Definition einer neuen Metrik, der *Balanced Mean Angular Error (BMAE)*, zur Bewertung der Orientierung des Kopfes teilt den Winkelbereich in Segmente und bestimmt den Durchschnitt des Winkelfehlers dieser Segmente. Damit ermöglicht diese Metrik eine aussagekräftige Bewertung über den kompletten Orientierungsbereich des Kopfes während Autofahrten, indem die häufige frontale Kopfausrichtung berücksichtigt wird. Anhand dieser Metrik zeigt das in der vorliegenden Arbeit präsentierte tiefe neuronale Netz, trainiert und ausgewertet auf der Basis von Tiefenbildern geringfügig bessere Ergebnisse als auf der Basis von Infrarotdaten. Die Fusionsverfahren zur Kombination beider Modalitäten erreichen eine noch genauere Bestimmung der Orientierung des Kopfes mit einem *Balanced Mean Angular Error (BMAE)* von $13,4^\circ$. Damit wurden in dieser Arbeit erstmals tiefen-basierte sowie Fusionsverfahren auf einem herausfordernden, der Wissenschaft zur Verfügung stehenden Datensatz während realer Fahrten ausgewertet. Mit der hohen Genauigkeit und der Robustheit gegenüber Verdeckungen stellen die in der vorliegenden Arbeit präsentierten Verfahren auf diesem *DriveAhead*-Datensatz den aktuellen Stand der Forschung für die Bestimmung der Kopfpose aus Tiefendaten im Fahrzeuginnenraum dar.

7.1. Einschränkungen und Ausblick.

Mit dieser Arbeit wurden Verfahren mit sehr guten Ergebnissen und ein neuer Datensatz entwickelt zur Bestimmung der Kopfpose im Fahrzeuginnenraum. Trotzdem bleibt noch Raum für Verbesserung und damit eröffnen sich neue Forschungsfelder.

Lichtverhältnisse im Fahrzeuginnenraum. Im Fahrzeuginnenraum können Lichtverhältnisse die Genauigkeit des Sensorsystems und der Algorithmen zur Bestimmung der Kopfpose beeinflussen. In dieser Arbeit wurde ein Sensor gewählt, der gegenüber Sonnenlicht robuste Tiefendaten liefert. Zusätzlich wurden die Algorithmen auf dem *DriveAHead*-Datensatz analysiert, der reale Autofahrten bei unterschiedlichen Wetterbedingungen und damit variierenden Lichtverhältnissen beinhaltet. Eine weitere Möglichkeit ist, die Algorithmen abhängig von bekannten Lichtverhältnissen zu untersuchen. Hierfür könnte ähnlich zu der hier durchgeführten Analyse des Sensorsystems, ein Datensatz im Labor und im Fahrzeug mit der gemessenen Fremdleistung aufgenommen werden.

Vibrationen im Fahrzeuginnenraum. In dieser Arbeit wurden Vibrationen des Sensor- und Referenzsystems nicht berücksichtigt. Die Vibrationen der Systeme wurden limitiert, indem diese mit individuell angefertigten Halterungen befestigt wurden.

Tracking-basierte Algorithmen. Der hier vorgestellte *DriveAHead*-Datensatz eröffnet die Möglichkeit neue Algorithmen für die Anwendung im Fahrzeuginnenraum zu entwickeln. Die vorliegende Arbeit beschränkt sich auf *frame-by-frame* Verfahren, die unabhängig vom vorherigen Bild die Orientierung und Position des Kopfes bestimmen. Eine Möglichkeit zur Verbesserung der Ergebnisse ist die Kombination der *frame*-basierten Verfahren mit *tracking*-Methoden. Die vorgestellten *frame*-basierten Verfahren können als Initialisierung für diese Verfahren dienen und damit die Anwendung der *tracking*-basierten Methoden nach Verdeckungen ermöglichen.

Tiefe neuronale Netze zur Bestimmung der Position und Orientierung des Kopfes. Durch die große Anzahl von Trainingsdaten ermöglicht der Datensatz die Entwicklung von neuen *Convolutional Neural Networks*-Architekturen zur Bestimmung der Orientierung und Position des Kopfes im Fahrzeuginnenraum. Das in der vorliegenden Arbeit vorgestellte tiefe neuronale Netz bestimmt die Orientierung des Kopfes. Diese Architektur oder eine andere Struktur kann zur Bestimmung der Translation des Kopfes weiterentwickelt werden.

Anhang A

DriveAHead Probanden

Sequenz	ID	Geschlecht		Daten vorhanden		
		männlich	weiblich	mit Brille	mit Sonnenbrille	ohne etwas
01	1		x		x	x
02	2		x		x	x
03	3	x			x	x
04	4	x		x	x	
05	5	x		x		
06	6	x			x	x
07	7		x		x	x
08	8	x		x		x
09	9	x			x	x
10	10	x			x	x
11	11		x		x	x
12	12	x			x	x
13	13	x			x	x
14	14	x		x		
15	15	x		x		
16	16	x			x	
17	17	x			x	x
18	18	x			x	x
19	19	x			x	x
20	20	x		x	x	
21	10	x				x

Tabelle A.1.: Übersicht der Probanden des DriveAHead Datensatzes. Für jede Sequenz ist die Probanden ID angegeben, das Geschlecht und ob Daten mit Brille, Sonnenbrille oder ohne Brille vorhanden sind.

Anhang B

DriveAhead Beispielbilder

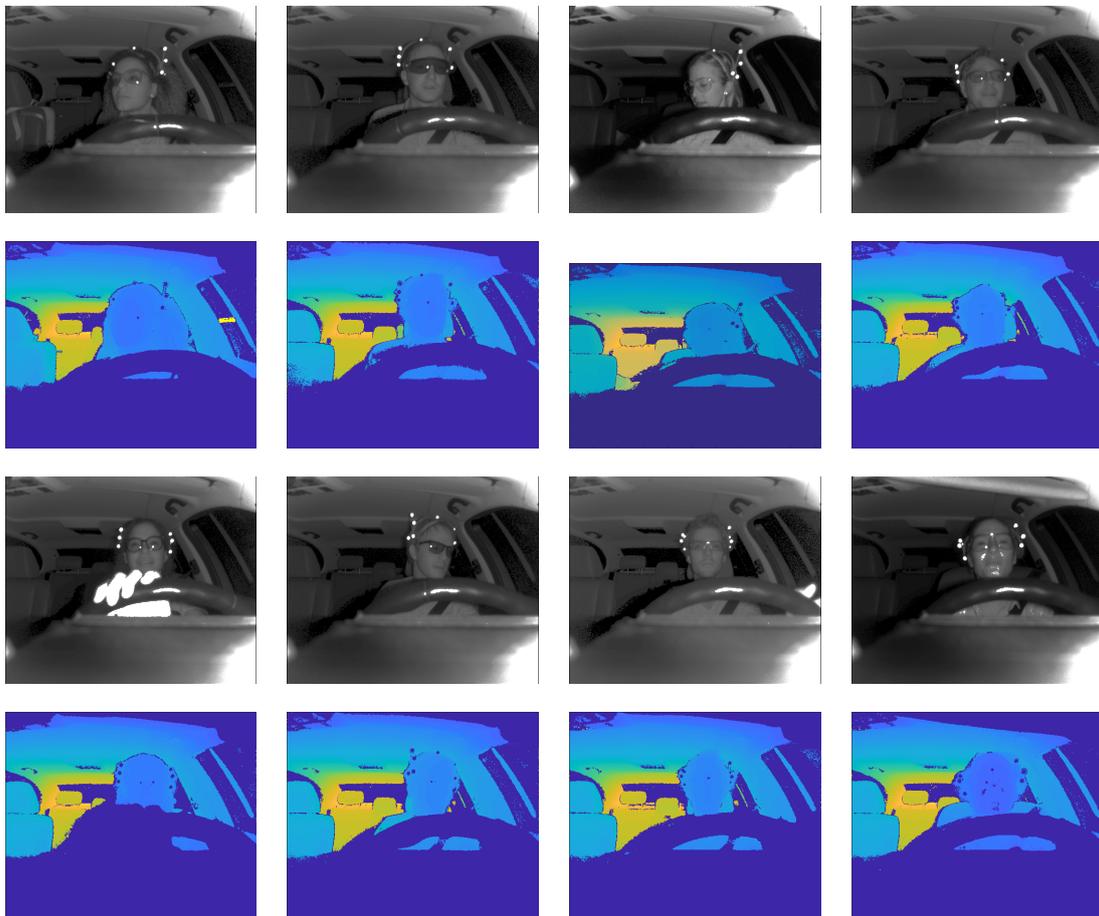


Abbildung B.1.: Beispielbilder von Probanden mit Sonnenbrillen. Die erste und dritte Zeile zeigen die Infrarotbilder, während die zweite und vierte Zeile die Tiefenbilder zeigen.

Eigene Veröffentlichungen

Anke Schwarz, Esther-Sabrina Wacker, Manuel Martin, M. Saquib Sarfraz, and Rainer Stiefelhagen. 3D Facial Landmark Detection: How to Deal with Head Rotations? In *German Conference on Pattern Recognition (GCPR)*, pages 424–434. Springer, 2015.

Anke Schwarz, Zhuang Lin, and Rainer Stiefelhagen. HeHOP: Highly efficient Head Orientation and Position estimation. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. [6](#), [59](#), [65](#), [66](#), [67](#)

Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen. DriveAHead - A Large-Scale Driver Head Pose Dataset. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1165–1174. IEEE, 2017. [3](#), [6](#), [9](#), [23](#), [42](#), [55](#), [56](#), [71](#), [81](#), [86](#), [93](#), [96](#)

Literaturverzeichnis

- Advanced Realtime Tracking (ART). URL www.ar-tracking.com. 25
- B. Ahn, J. Park, und I. S. Kweon. Real-time head orientation from a monocular camera using deep neural network. In *Asian Conference on Computer Vision (ACCV)*, S. 82–96. Springer, 2014. 15, 16, 19, 73, 81, 88, 90, 92, 93, 100
- N. Alioua, A. Amine, A. Rogozan, A. Bensrhair, und M. Rziza. Driver head pose estimation using efficient descriptor fusion. In *EURASIP Journal on Image and Video Processing*. Springer, 2016. 15, 16
- M. Ariz, J. J. Bengoechea, A. Villanueva, und R. Cabeza. A novel 2d/3d database with automatic face annotation for head tracking and pose estimation. In *Computer Vision and Image Understanding*, volume 148, S. 201–210. Elsevier, 2016. 9, 10
- T. Baltrušaitis, P. Robinson, und L.-P. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2610–2617. IEEE, 2012. 9, 10, 88
- T. Baltrušaitis, P. Robinson, und L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Winter Conference on Applications of Computer Vision (WACV)*, S. 1–10. IEEE, 2016. 11, 12, 14, 17, 42, 81, 88, 90, 97, 98, 99, 100, 101
- T. Bär, J. F. Reuter, und J. M. Zöllner. Driver head pose and gaze estimation based on multi-template ICP 3-D point cloud alignment. In *International Conference on Intelligent Transportation Systems (ITSC)*, S. 1797–1802. IEEE, 2012. 18, 19
- P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, und N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 35, S. 2930–2940. IEEE, 2013. 88
- P. J. Besl und N. D. McKay. A method for registration of 3-D shapes. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 14, S. 239–256. IEEE, 1992. 37

- L. Beyer, A. Hermans, und B. Leibe. Biternion nets: Continuous head pose regression from discrete training labels. In *German Conference on Pattern Recognition (GCPR)*, 2015. 22
- G. Borghi, M. Venturelli, R. Vezzani, und R. Cucchiara. POSEidon: Face-From-Depth for Driver Pose Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 19, 23
- M. Breidt, H. H. Bülthoff, und C. Curio. Accurate 3d head pose estimation under real-world driving conditions: A pilot study. In *International Conference on Intelligent Transportation Systems (ITSC)*, S. 1261–1268. IEEE, 2016. 19
- M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, und H. Pfister. Real-Time Face Pose Estimation from Single Range Images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1–8. IEEE, 2008. 17
- I. N. Bronstein, J. Hromkovic, B. Luderer, H.-R. Schwarz, J. Blath, A. Schied, S. Dempe, G. Wanka, und S. Gottwald. *Taschenbuch der Mathematik*, volume 1. Springer, 2012. 43
- Q. Cai, D. Gallup, C. Zhang, und Z. Zhang. 3D Deformable Face Tracking with a Commodity Depth Camera. In *European Conference on Computer Vision (ECCV)*, S. 229–242. Springer, 2010. 18
- A. Collet, M. Martinez, und S. S. Srinivasa. The MOPED framework: Object recognition and pose estimation for manipulation. In *The International Journal of Robotics Research (IJRR)*, volume 30, S. 1284–1306. SAGE Publications Sage UK: London, England, 2011. 22, 51, 83
- T. F. Cootes, G. J. Edwards, und C. J. Taylor. Active Appearance Models. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 23, S. 681–685. IEEE, 2001. 14, 15
- E. B. Dam, M. Koch, und M. Lillholm. *Quaternions, interpolation and animation*, volume 2. Datalogisk Institut, Københavns Universitet, 1998. 22, 63
- M. Dantone, J. Gall, G. Fanelli, und L. Van Gool. Real-time Facial Feature Detection using Conditional Regression Forests. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2578–2585. IEEE, 2012. 15
- D. F. Dementhon und L. S. Davis. Model-based object pose in 25 lines of code. In *International Journal of Computer Vision (IJCV)*, volume 15, S. 123–141. Springer, 1995. 12, 16, 88
- Faceshift AG. URL www.faceshift.com. 9, 10, 12
- L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot, und G. Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer, 2016. 49

- G. Fanelli, J. Gall, und L. Van Gool. Real Time Head Pose Estimation with Random Regression Forests. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 617–624. IEEE, 2011a. 9, 10, 17, 21, 65, 68
- G. Fanelli, T. Weise, J. Gall, und L. van Gool. Real Time Head Pose Estimation from Consumer Depth Cameras. In *German Conference on Pattern Recognition (GCPR)*, S. 101–110. Springer, 2011b. 17
- G. Fanelli, M. Dantone, J. Gall, A. Fossati, und L. Van Gool. Random Forests for Real Time 3D Face Analysis. In *International Journal of Computer Vision (IJCV)*, volume 101, S. 437–458. Springer, 2013. 5, 9, 10, 12, 15, 16, 18, 19, 20, 21, 42, 59, 63, 64, 65, 67, 68, 69, 82, 88
- S. Foix, G. Alenya, und C. Torras. Lock-in time-of-flight (ToF) cameras: A survey. In *Sensors Journal*, volume 11, S. 1917–1926. IEEE, 2011. 30
- L. Fridman, J. Lee, B. Reimer, und T. Victor. ‘Owl’and ‘Lizard’: patterns of head pose and eye pose in driver gaze classification. In *IET Computer Vision*, volume 10, S. 308–313, 2016. 1
- A. Gee und R. Cipolla. Determining the gaze of faces in images. In *Image and Vision Computing*, volume 12, S. 639–647. Elsevier, 1994. 14
- D. Gilbarg und N. S. Trudinger. *Elliptic partial differential equations of second order*. Springer, 2015. 86
- N. Gourier, D. Hall, und J. L. Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, volume 6, 2004b. 8, 9, 11, 15, 16
- R. Gross, I. Matthews, J. Cohn, T. Kanade, und S. Baker. Multi-pie. In *Image and Vision Computing*, volume 28, S. 807–813. Elsevier, 2010. 88
- W. R. Hamilton. Ii. on quaternions; or on a new system of imaginaries in algebra. *Philosophical Magazine Series 3*, 25(163):10–13, 1844. 22
- K. He, X. Zhang, S. Ren, und J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, S. 630–645. Springer, 2016. 71
- M. Höffken, E. Tarayan, U. Kresel, und K. Dietmayer. Stereo vision-based driver head pose estimation. In *Intelligent Vehicles Symposium (IV)*, S. 253–260. IEEE, 2014. 19
- T. Horprasert, Y. Yacoob, und L. S. Davis. Computing 3-d head orientation from a monocular image sequence. In *International Conference on Automatic Face and Gesture Recognition (FG)*, S. 242–247, 1996. 14
- Image Processing Toolbox MATLAB, R2016a. 86

- S. Kaymak und I. Patras. Exploiting Depth and Intensity Information for Head Pose Estimation with Random Forests and Tensor Models. In *Asian Conference on Computer Vision (ACCV)*, S. 160–170. Springer, 2012. 21
- A. Kendall, M. Grimes, und R. Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2938–2946. IEEE, 2015. 73
- D. E. King. Dlib-ml: A machine learning toolkit. In *Journal of Machine Learning Research*, volume 10, S. 1755–1758, 2009. 73, 87
- D. Kingma und J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014a. 15, 74
- H. M. Kjer und J. Wilm. *Evaluation of surface registration algorithms for PET motion correction*. Bachelorarbeit, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2010. 37
- A. Krizhevsky, I. Sutskever, und G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, S. 1097–1105, 2012. 71
- N. Krüger, M. Pöttsch, und C. von der Malsburg. Determination of face position and pose with a learned representation based on labelled graphs. In *Image and Vision Computing*, volume 15, S. 665–673. Elsevier, 1997. 14
- M. La Cascia, S. Sclaroff, und V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 22, S. 322–336. IEEE, 2000. 8, 9
- B. Langmann, K. Hartmann, und O. Loffeld. Depth Camera Technology Comparison and Performance Evaluation. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, S. 438–444, 2012. 34
- A. Lanitis, C. J. Taylor, und T. F. Cootes. Automatic interpretation and coding of face images using flexible models. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 19, S. 743–756. IEEE, 1997. 14
- V. Le, J. Brandt, Z. Lin, L. Bourdev, und T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision (ECCV)*, S. 679–692. Springer, 2012. 88
- L. Li. Time-of-flight camera—an introduction. *Technical white paper*, (SLOA190B), 2014. 31
- Y. Li, S. Gong, J. Sherrah, und H. Liddell. Support vector machine based multi-view face detection and recognition. In *Image and Vision Computing*, volume 22, S. 413–427. Elsevier, 2004. 15

- Y. Ma, Y. Konishi, K. Kinoshita, S. Lao, und M. Kawade. Sparse bayesian regression for head pose estimation. In *International Conference on Pattern Recognition (ICPR)*, S. 507–510. IEEE, 2006. 15
- F. L. Markley, Y. Cheng, J. L. Crassidis, und Y. Oshman. Averaging quaternions. *Journal of Guidance Control and Dynamics*, 30(4):1193, 2007. 49
- M. Martin, F. van de Camp, und R. Stiefelhagen. Real Time Head Model Creation and Head Pose Estimation on Consumer Depth Cameras. In *International Conference on 3D Vision (3DV)*, S. 641–648. IEEE, 2014. 18
- S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, und M. Trivedi. On the design and evaluation of robust head pose for visual user interfaces: Algorithms, databases, and comparisons. In *International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, S. 149–154. ACM, 2012. 9, 10, 15, 16, 23
- G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, und J. Kautz. Robust Model-Based 3D Head Pose Estimation. In *International Conference on Computer Vision (ICCV)*, S. 3649–3657. IEEE, 2015. 18
- Microsoft. Kinect One, 2017. URL <https://developer.microsoft.com/de-de/windows/kinect/hardware>. 05.09.2017. 11, 31, 33, 34
- I. Misra, A. Shrivastava, A. Gupta, und M. Hebert. Cross-stitch networks for multi-task learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 3994–4003. IEEE, 2016. 21, 71, 76
- H. Moon und M. L. Miller. Estimating facial pose from a sparse representation [face recognition applications]. In *International Conference on Image Processing (ICIP)*, S. 75–78, 2004. 15
- E. Murphy-Chutorian und M. M. Trivedi. Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking. In *Intelligent Vehicles Symposium (IV)*, S. 512–517. IEEE, 2008. 16
- E. Murphy-Chutorian und M. M. Trivedi. Head Pose Estimation in Computer Vision: A Survey. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, S. 607–626. IEEE, 2009. 14
- E. Murphy-Chutorian, A. Doshi, und M. M. Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *International Conference on Intelligent Transportation Systems (ITSC)*, S. 709–714. IEEE, 2007. 15, 16
- V. Nair und G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*, S. 807–814, 2010. 73

- K. Nickel und R. Stiefelhagen. Pointing gesture recognition based on 3d-tracking of face, hands and head orientation. In *International Conference on Multimodal Interfaces (ICMI)*, S. 140–146. ACM, 2003. 17
- OpenStreetMap-Mitwirkende, 2017. URL www.openstreetmap.org/copyright. 03.04.2017. 27
- M. Osadchy, Y. L. Cun, und M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research (JMLR)*, S. 1197–1215, 2007. 15
- P. Paderis, X. Zabulis, und A. A. Argyros. Head pose estimation on depth data based on Particle Swarm Optimization. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, S. 42–49. IEEE, 2012. 18
- C. Papazov, T. K. Marks, und M. Jones. Real-Time 3D Head Pose and Facial Landmark Estimation From Depth Images Using Triangular Surface Patch Features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 4722–4730. IEEE, 2015. 18, 20, 64, 68
- G. A. Peláez C., F. García, A. de La Escalera, und J. M. Armingol. Driver monitoring based on low-cost 3-d sensors. In *Transactions on Intelligent Transportation Systems*, S. 1855–1860. IEEE, 2014. 19
- P. M. Photonics. 19k-s3, 2014. URL http://www.pmdtec.com/html/pdf/pmdPhotonICs_19k_S3.pdf. 01.09.2017. 31, 32, 33, 34
- R. Rae und H. J. Ritter. Recognition of human head orientation based on artificial neural networks. In *Transactions on neural networks*, S. 257–265. IEEE, 1998. 15
- K. Ramnath, S. Koterba, J. Xiao, C. Hu, I. Matthews, S. Baker, J. Cohn, und T. Kanade. Multi-view AAM fitting and construction. In *International Journal of Computer Vision (IJCV)*, S. 183–204. Springer, 2008. 15
- C. Redondo-Cabrera, R. López-Sastre, und T. Tuytelaars. All together now: Simultaneous object detection and continuous pose estimation using a hough forest with probabilistic locally enhanced voting. In *British Machine Vision Conference (BMVC)*, S. 1–12, 2014. 18, 21
- S. Ren, X. Cao, Y. Wei, und J. Sun. Face Alignment at 3000 FPS via Regressing Local Binary Features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 1685–1692. IEEE, 2014. 20, 59, 61
- M. Rezaei und R. Klette. Look at the driver, look at the road: No distraction! no accident! In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 129–136. IEEE, 2014. 16
- G. Riegler, D. Ferstl, M. Rütger, und H. Bischof. Hough Networks for Head Pose Estimation and Facial Feature Localization. In *British Machine Vision Conference (BMVC)*, S. 437–458, 2014. 18, 20, 68

- T. Rueda-Domingo, P. Lardelli-Claret, J. de Dios Luna-del-Castillo, J. J. Jiménez-Moleón, M. García-Martín, und A. Bueno-Cavanillas. The influence of passengers on the risk of the driver causing a car collision in Spain: Analysis of collisions from 1990 to 1999. In *Accident Analysis and Prevention*, S. 481–489. Elsevier, 2004. [1](#)
- O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, . Imagenet large scale visual recognition challenge. In *International Journal of Computer Vision (IJCV)*. Springer, 2015. [71](#)
- C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, und M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *International Conference on Computer Vision Workshops (ICCVW)*, S. 397–403. IEEE, 2013. [46](#)
- A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, und L. Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, S. 47–56. Springer, 2008. [8](#), [9](#), [11](#)
- S. Schuster, C. Leistner, P. Wohlhart, P. M. Roth, und H. Bischof. Alternating regression forests for object detection and pose estimation. In *International Conference on Computer Vision (ICCV)*, S. 417–424. IEEE, 2013. [18](#)
- A. Schwarz, Z. Lin, und R. Stiefelhagen. HeHOP: Highly efficient Head Orientation and Position estimation. In *Winter Conference on Applications of Computer Vision (WACV)*, S. 1–8. IEEE, 2016. [6](#), [59](#), [65](#), [66](#), [67](#)
- A. Schwarz, M. Haurilet, M. Martinez, und R. Stiefelhagen. DriveAHead - A Large-Scale Driver Head Pose Dataset. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, S. 1165–1174. IEEE, 2017. [3](#), [6](#), [9](#), [23](#), [42](#), [55](#), [56](#), [71](#), [81](#), [86](#), [93](#), [96](#)
- E. Seemann, K. Nickel, und R. Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *International Conference on Automatic Face and Gesture Recognition (FG)*, S. 626–631. IEEE, 2004. [15](#), [17](#), [20](#)
- K. Simonyan und A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. [20](#), [21](#), [71](#), [72](#), [73](#), [79](#), [89](#)
- Softkinetic. Ds325, 2014. URL http://www.softkinetic.com/Portals/0/Download/WEB_20120907_SK_DS325_Datasheet_V2.1.pdf. 01.09.2017. [31](#), [32](#), [33](#)
- P. Soille. *Morphological image analysis: Principles and applications*. Springer Science & Business Media, 2013. [85](#)
- R. Stiefelhagen, J. Yang, und A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. In *Transactions on Neural Networks*, S. 928–938. IEEE, 2002. [15](#)

- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, und A. Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. 71
- D. J. Tan, F. Tombari, und N. Navab. Real-Time Accurate 3D Head Tracking and Pose Estimation with Consumer RGB-D Cameras. In *International Journal of Computer Vision (IJCV)*, S. 1–26. Springer, 2017. 18
- A. Tawari, S. Martin, und M. M. Trivedi. Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations. In *Transactions on Intelligent Transportation Systems (ITSC)*, S. 818–830. IEEE, 2014. 9, 11, 12, 16, 23
- Y. Tessema, M. Höffken, und U. Kreßel. Driver Head Pose Estimation by Regression. In *Advanced Microsystems for Automotive Applications*, S. 55–67. Springer, 2016. 19
- B. Trefflich. *Videogestützte Überwachung der Fahreraufmerksamkeit und Adaption von Fahrerassistenzsystemen*. Dissertation, Technische Universität Ilmenau, 2010. 1
- S. Tulyakov, R.-L. Vieri, S. Semeniuta, und N. Sebe. Robust real-time extreme head pose estimation. In *International Conference on Pattern Recognition (ICPR)*, S. 2263–2268. IEEE, 2014. 18
- U. S. Department of Transportation. Bureau of transportation statistics. *NHTS 2001 Highlights Report*, BTS03-05, 2003. 1
- M. Venturelli, G. Borghi, R. Vezzani, und R. Cucchiara. Deep Head Pose Estimation from Depth Data for In-car Automotive Applications. *arXiv preprint arXiv:1703.01883*, 2017. 19
- P. Viola und M. J. Jones. Robust Real-Time Face Detection. In *International Journal of Computer Vision (IJCV)*, S. 137–154. Springer, 2004. 64
- M. Voit, K. Nickel, und R. Stiefelhagen. Head Pose Estimation in Single- and Multi-view Environments – Results on the CLEAR’07 Benchmarks. In *Multimodal Technologies for Perception of Humans*, S. 307–316. Springer, 2008. 15
- J.-G. Wang und E. Sung. Em enhancement of 3d head pose estimated by point at infinity. In *Image and Vision Computing*, S. 1864–1874. Elsevier, 2007. 14
- R. Wolman. *Untersuchung von TOF basierten Tiefendaten bezüglich der Fahrerzustandsbestimmung*. Bachelorarbeit, Hochschule Heilbronn, 2014. 23, 30, 32, 33, 35, 36, 37, 38, 39, 40
- X. Xiong und F. De la Torre. Supervised descent method and its applications to face alignment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 532–539. IEEE, 2013. 14

- Y. Yun, M. H. Changrampadi, und I. Y. Gu. Head pose classification by multi-class AdaBoost with fusion of RGB and depth images. In *International Conference on Signal Processing and Integrated Networks (SPIN)*, S. 174–177. IEEE, 2014. 21
- M. D. Zeiler und R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, S. 818–833, 2014. 77
- Z. Zhang. A flexible new technique for camera calibration. In *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, S. 1330–1334. IEEE, 2000. 44
- X. Zhu und D. Ramanan. Face Detection, Pose Estimation, and Landmark Localization in the Wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 2879–2886. IEEE, 2012. 15, 19