Gottfried Wilhelm Leibniz Universität Hannover
Fachrichtung Geodäsie und Geoinformatik
Institut für Photogrammetrie und GeoInformation

**Bachelorarbeit**

# Untersuchung eines Mehr-Personen-Klassifikators zur Wiedererkennung von Personen in Bildsequenzen

Dominic Grüning

Matrikel-Nr.: 3022890

20. September 2016

1. Prüfer: Prof. Dr.-habil. Christian Heipke
2. Prüfer: Dipl.-Ing. Tobias Klinger

# Inhaltsverzeichnis

## Erklärung der Urheberschaft

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum Unterschrift

# 1 Motivation und Zielsetzung

## 1.1 Motivation

Bei immer mehr Aufgaben des täglichen Lebens werden Menschen von Computern unterstützt, zum Beispiel beim Fahren durch Fahrassistenz-Systeme oder bei der Hausarbeit durch autonome Rasenmäher. Mit dieser zunehmenden Automatisierung der Welt wird es für Computer immer wichtiger, nicht nur für, sondern vor allem mit Menschen zu arbeiten. Für ein Computersystem ist es dabei sinnvoll, die Informationen über seine Umwelt schnell und berührungsfrei zu erhalten: Es muss in seiner Umwelt sehen können.
Die Fähigkeit eines Computers, Bildinformationen aus Kameras zu verarbeiten und zu deuten (Computer Vision), setzt im Zusammenhang mit der Aufgabe, mit Menschen zu arbeiten, voraus, dass Menschen im Allgemeinen nicht nur als solche erkannt, sondern auch als Individuen wiedererkannt werden. Dies muss unter anderem im Bereich des autonomen Fahrens in Echtzeit geschehen, damit Menschen nicht behindert oder gefährdet werden. In Bezug auf das Arbeiten mit Menschen reichen die Anwendungsgebiete der Computer Vision von autonomen Fahrzeugen über Unterhaltungstechniken und Mensch-Maschine-Kommunikation bis hin zu automatisierten Überwachungssystemen. In diese Bereiche fallen zwei wesentliche Teilaspekte der Computer Vision in Bezug auf die Erkennung von Menschen: Das Tracking und die Wiedererkennung.
Das Tracking beschreibt für sich genommen die Aufzeichnung der Bewegungstrajektorien von Objekten im Raum oder im Bild über die Zeit, wobei die Erkennung dieser Objekte (Detektion) einen wesentlichen Zwischenschritt ausmacht. An das Tracking angelehnt ist die rekursive Zustands- bzw. Positionsschätzung: Zum einen kann die Position eines Objekts im vorhergegangenen Zeitschritt durch Informationen bzw. Messungen aus dem aktuellen Zeitschritt verbessert bzw. geupdated werden, zum anderen kann aus Messungen die Position des Objekts im kommenden Zeitschritt geschätzt bzw. prädiziert werden. Um die Positionen von Objekten korrekt zu prädizieren, ist es sinnvoll, Annahmen über die Bewegung der Objekte zu treffen, z.B. dass diese sich mit konstanter Geschwindigkeit fortbewegen: Ein einzelnes Objekt wird sich von Bild zu Bild nur wenig im Raum bewegen, die prädizierte Position liegt damit nahe an der letzten bekannten Position. Dies ist unproblematisch, wenn nur ein Objekt auf einmal getrackt wird, da es pro Bild nur eine korrekte Detektion und eine Trajektorie gibt. Sollen allerdings mehrere Objekte gleichzeitig in einer Bildsequenz getrackt werden, müssen mehrere Trajektorien und Detektionen verarbeitet werden. Das Problem dabei ist, dass nicht von vornherein bekannt ist, welche Detektion zu welcher Trajektorie gehört; die Detektionen müssen den Trajektorien zugeordnet werden. Dieses Zuordnungsproblem kann durch die Klassifikation als Mittel zur Wiedererkennung angegangen werden: Das Ziel dabei ist es, für jedes zu trackende Objekt Merkmale aus den Farbinformationen zu extrahieren, mit denen sich das Objekt

beschreiben lässt. Ein weiteres Problem, welches beim Tracking auftreten kann, ist die Verdeckung von Objekten: Ist ein zu trackendes Objekt von einem anderen Objekt aus der Sicht der Kamera verdeckt, so wird es gegebenenfalls nicht detektiert und die Trajektorie nicht geupdated. Ist das Objekt einige Bilder später wieder im Bild zu sehen, kann es vom Detektor erkannt und durch die Klassifikation als eindeutiges Objekt wiedererkannt werden. Durch die Wiedererkennung ist es möglich, die Trajektorie des Objekts nach der Verdeckung fortzuführen.

## 1.2 Stand der Forschung

Im Folgenden werden bestehende Ansätze bezüglich des Trackings und der Klassifikation vorgestellt. Ebenso werden Möglichkeiten aufgezeigt, wie diese Ansätze genutzt werden können, um ein bestehendes Verfahren hinsichtlich der Tracking-Genauigkeit zu verbessern.

In Shu et al. [2012] werden die Bildinformationen bzw. Merkmale eines Menschen im Bild genutzt, um diesen individuell zu beschreiben bzw. zu klassifizieren. Dabei werden insbesondere die Farbinformationen einzelner Körperteile eines Menschen im Bild verwendet. Dies bietet bei der Verdeckung bestimmter Körperteile die Möglichkeit, diese nicht mit in die Klassifizierung einzubeziehen.
Ein Problem bei der Klassifikation zur Unterstützung des Personentrackings kommt durch die Tatsache zustande, dass sich die Personen im Bild bewegen. Die Klassifikation baut darauf auf, dass die Merkmale des zu klassifizierenden Objekts ähnlich bleiben. Bewegt sich eine Person, ändern sich ihre Haltung und damit auch die Farbinformationen im Bild, aus denen die Merkmale der Person extrahiert werden; die Wiedererkennung einer Person durch die Klassifikation wird damit erschwert. In Farenzena et al. [2010] wurde dafür ein Ansatz vorgestellt, der die Merkmalsveränderung (aufgrund der Variationen der Posen) durch Bestimmung von Symmetrieachsen am Menschen abfängt. Dieser Ansatz zur Merkmalsextraktion vereinfacht die Wiedererkennung von Personen und kann dazu verwendet werden, bestehende Tracking-Verfahren hinsichtlich der Wiedererkennung von Personen nach einer Verdeckung zu verbessern.
Die vorliegende Arbeit baut auf dem in Klinger and Muhle [2012] vorgestellten Ansatz auf: Das Tracking von Personen wird dabei primär durch die Detektion umgesetzt. Der Detektor liefert als Ergebnis eine Bounding Box bzw. ein Rechteck, welches eine Person vollständig beinhaltet. Die Zuordnung zwischen Detektion und Trajektorien wird durch die Klassifikation von Personen vereinfacht. Die Prädiktion der Bewegungen der Personen wird durch einen Kalmanfilter mit linearem Bewegungsansatz realisiert. Die Merkmale von Personen werden durch einen *Random Forest* klassifiziert, welcher sich aus einer Menge von Entscheidungsbäumen zusammensetzt. Ein Nachteil dieses Ansatzes liegt in der Umsetzung des Klassifikators, indem ein ellipsenförmiger Bildausschnitt von einer Person

alle Farbinformationen als Merkmale dieses Menschen enthält. Die Ellipse ist eine eher schlechte Approximation der Kontur eines Menschen, da im Ausschnitt der Hintergrund enthalten ist, welcher nicht mit zur Klassifikation genutzt werden sollte. Zudem wird der Merkmalsraum durch das Verwenden aller Farbinformationen in der Ellipse sehr hochdimensional, wodurch mehr Trainingsdaten für den Klassifikator erforderlich sind.
Wesentliche Probleme, die im Ansatz von Klinger and Muhle [2012] wie auch in anderen Tracking-Verfahren auftreten, sind *IdentitySwitches* und *Fragmentations*. Ersteres beschreibt den Fall, dass die Detektion der einen Person der Trajektorie einer anderen Person zugewiesen wird. Letzteres tritt vor Allem nach einer Verdeckung auf: Kann eine Person aufgrund von Verdeckungen, z.B. durch Laternen oder andere Menschen, nicht detektiert werden, so wird die Trajektorie unterbrochen bzw. fragmentiert.

## 1.3 Zielsetzung und Aufbau der Arbeit

Das Ziel der vorliegenden Arbeit ist es, den von Klinger and Muhle [2012] entwickelten Ansatz hinsichtlich der Klassifikation und der Überbrückung von Verdeckungen zu verbessern. Dazu wird unter anderem der in Farenzena et al. [2010] vorgestellte Ansatz der Merkmalsextraktion, aufgrund der komplexen Programmierung allerdings leicht verändert, umgesetzt. Insbesondere wird untersucht, wie die Merkmale, die einen Menschen beschreiben, bestmöglich aus den Bildinformationen extrahiert werden können und wie die Zuordnung von Detektionen zu Trajektorien verbessert werden kann.

Die vorliegende Arbeit gliedert sich wie folgt: In Kapitel 2 werden die Grundlagen dieser Arbeit erklärt, um einen thematischen Einstieg in das behandelte Thema zu geben. Kapitel 3 beschäftigt sich mit der methodischen Untersuchung in Bezug auf die Merkmale von Menschen sowie dem Zuordnungsproblem. Die entwickelten Methoden werden in Kapitel 4 experimentell untersucht. In Kapitel 5 werden die erzielten Ergebnisse miteinander verglichen und bezüglich ihrer Signifikanz bewertet. Kapitel 6 fasst abschließend die entwickelten Methoden und deren Ergebnisse zusammen, vergleicht sie mit Ergebnissen anderer Arbeiten und stellt einen Ausblick auf weiterführende Untersuchungen.

# 2 Grundlagen

Das grundlegende Ziel dieser Arbeit ist es, die Trajektorien von Personen in der Bildsequenz nach einer Verdeckung korrekt fortzuführen. Dazu ist es nötig, detektierte Personen so zu klassifizieren, dass diese eindeutig wiedererkannt werden, auch wenn sie für einige Bilder nicht detektiert wurden. Die beiden wichtigsten Werkzeuge für diese Aufgabe, der *Online Random Forest* und das detektionsbasierte Tracking, werden im folgenden erläutert.

## 2.1 Random Forest als überwachtes Klassifikationsverfahren

Die Aufgabe eines Klassifikators ist im Allgemeinen, ein Objekt einer Klasse zuzuordnen. Bei der überwachten Klassifikation muss der Klassifikator allerdings erst anhand von Trainingsdaten lernen, welche Klassen er im Merkmalsraum voneinander unterscheiden muss. Diese Trainingsdaten repräsentieren Objekte, deren Merkmale und Klassen bekannt sind. Durch das Trainieren ist es dem Klassifikator möglich, Grenzen im Merkmalsraum zu finden, die die Klassen voneinander trennen. Die Zuordnung von Objekten zu ihren Klassen erfolgt dementsprechend anhand der Lage der Objektmerkmale im Merkmalsraum. Die Grundlage des hier verwendeten Klassifikators ist ein binärer Entscheidungsbaum: An jedem Knoten eines Baumes wird anhand der Merkmale des zu klassifizierenden Objektes eine Binär-Entscheidung getroffen. Am Ende von mehreren Knoten steht ein Blatt, welches dem Klassifikatorergebnis entspricht. Das Prinzip ist in Abbildung 1 dargestellt.

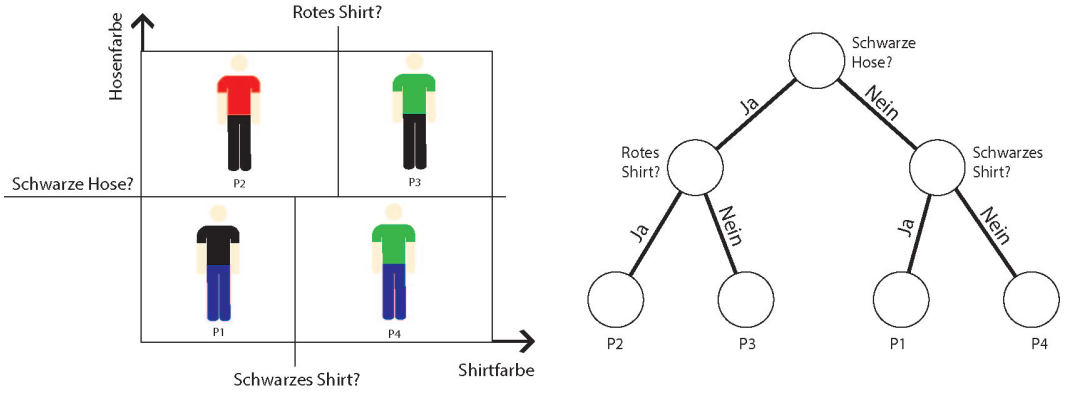



Abbildung 1: Merkmalsraum (links) mit schematischer Darstellung des entsprechenden Binärbaums (rechts)

Zum Anlernen des Baumes wird die CART-Methode (*Classification and Regression Trees*) verwendet. Dabei wird der Merkmalsraum durch lineare Entscheidungsgrenzen aufgeteilt. Die Grenzen müssen nicht parallel zu den Achsen des Merkmalsraumes sein, wodurch mehrere Merkmale gleichzeitig berücksichtigt werden können. Für jeden Knoten im Baum wird eine gewisse Anzahl an Grenzen zufällig generiert, für welche überprüft

wird, wie gut sie die Daten voneinander trennen. Die Grenze, welche die Daten am besten trennt, wird für den jeweiligen Knoten gewählt. Das Anlernen dieser Grenzen erfolgt nur mit einem Teil (z.B. 1/3) der Trainingsdaten. Durch das Anwenden der Entscheidungen auf die restlichen Daten wird für jedes Blatt ein normiertes Histogramm erstellt. Die relative Häufigkeit, mit der ein Datensatz einer bestimmten Klasse einem Blatt zugeordnet wird, kann als a posteriori-Wahrscheinlichkeit für die Zugehörigkeit des Datensatzes zu der entsprechenden Klasse interpretiert werden.
Um Unsicherheitsmaße für den Klassifikator zu bestimmen, wird das *Bootstrapping*-Verfahren angewendet. Dabei werden Datenpunkte eines Eingangs-Datensatzes ausgewählt und dem Bootstrap-Datensatz hinzugefügt, wobei einzelne Datenpunkte mehrmals dem Bootstrap-Datensatz hinzugefügt werden können. Es werden mehrere Bootstrap-Datensätze erzeugt, auf welchen jeweils der Klassifikator trainiert wird. Durch die zufällige Auswahl der Datenpunkte in den Bootstrap-Datensätzen werden für jeden Knoten eines jeden Baumes unterschiedliche Entscheidungsgrenzen ausgewählt, was zu unterschiedlichen Verteilungsergebnissen in den Blättern der Bäumen führt. Das Unsicherheitsmaß des Klassifikators entspricht der Variation der Ergebnisse von jedem der Bäume und wird dazu verwendet, die Klassifikationsgenauigkeit des Klassifikators zu bestimmen. Werden die Ergebnisse der unterschiedlich trainierten Bäume zusammengefasst, spricht man von *Bootstrap Aggregating* bzw. *Bagging*. Dabei wird ein zu klassifizierendes Objekt bzw. dessen Merkmalsdatensatz durch die Bäume klassifiziert. Aufgrund der unterschiedlichen Bootstrap-Daten, mit denen die Bäume trainiert wurden, können sich die Klassifikationsergebnisse der Bäume voneinander unterscheiden. Die Ergebnisse können als Häufigkeitsverteilung bzw. Histogramm dargestellt werden, in welchem für jede Klasse festgehalten wird, mit welcher relativen Häufigkeit diese Klasse das Ergebnis der Bäume ist. Das Ergebnis des Klassifikators ist die Klasse, die aus den Ergebnissen der einzelnen Bäumen den größten Anteil in der Häufigkeitsverteilung ausmacht. Ein Random Forest stellt die Anwendung des Baggings auf die CART-Methode dar. Zusammengefasst werden bei Random Forests mehrere Entscheidungsbäume aus zufällig generierten Entscheidungsfunktionen aufgebaut. Die zu klassifizierenden Merkmalsvektoren werden durch alle Entscheidungsbäume geschleust und die Ergebnisse über alle Bäume gemittelt, so dass das Ergebnis der Klasse mit der größten Häufigkeit im Histogramm des Random Forests entspricht.
Diese Methode des überwachten Lernens setzt voraus, dass zum Trainieren des Klassifikators Trainingsdaten existieren. Dies ist für Echtzeit-Anwendungen, in denen die Daten nacheinander zur Verfügung stehen, nicht gegeben. Auch kann hierbei nicht zwischen Trainings- und Testphase unterschieden werden; die generierten Bäume müssen zur Laufzeit auf ihre Genauigkeit hin überprüft werden. Einen Ansatz zur Lösung dieses Problems liefert Saffari et al. [2009] durch Anwendung von Online Random Forests (ORFs). Die Grundidee dabei ist, die Entscheidungsbäume zur Laufzeit wachsen zu lassen. In den Knoten der Bäume werden zufällige Entscheidungsfunktionen (Grenzen im Merkmals-

raum) generiert, welche durch ihren Informationsgewinn bewertet werden. Ein Knoten wird aufgeteilt (bzw. der Baum wächst), wenn sowohl für eine robuste Statistik hinreichend viele Trainingsdaten in den Knoten eingegangen sind und der Informationsgewinn einer Entscheidungsfunktion im Knoten einen Grenzwert übersteigt. Die Entscheidungsfunktion mit dem größten Informationsgewinn wird als Funktion für den Knoten festgelegt. Außerdem werden die gesammelten statistischen Informationen an die neu generierten Kind-Knoten weitergegeben. Ein Problem beim Tracking in Echtzeit ist, dass sich die zu trackenden Objekte im Raum bewegen können, wodurch sich die Position einer entsprechenden Klasse im Merkmalsraum ändern kann. Ein Ansatz zur Lösung des Problems ist, dass einzelne Entscheidungsbäume des Online Random Forests, in Abhängigkeit von der Anzahl der Daten, mit denen der Baum generiert wurde, gelöscht werden. Dies wirkt sich nicht wesentlich auf die Klassifikationsgenauigkeit des ORFs aus, da ein einzelner Baum wenig Gewicht im Vergleich zum ganzen ORF hat. Durch das Löschen wird sichergestellt, dass neue Verteilungen im Merkmalsraum berücksichtigt werden und der ORF stets an die aktuellen Daten angepasst ist.

## 2.2 Detektionsbasiertes Tracking

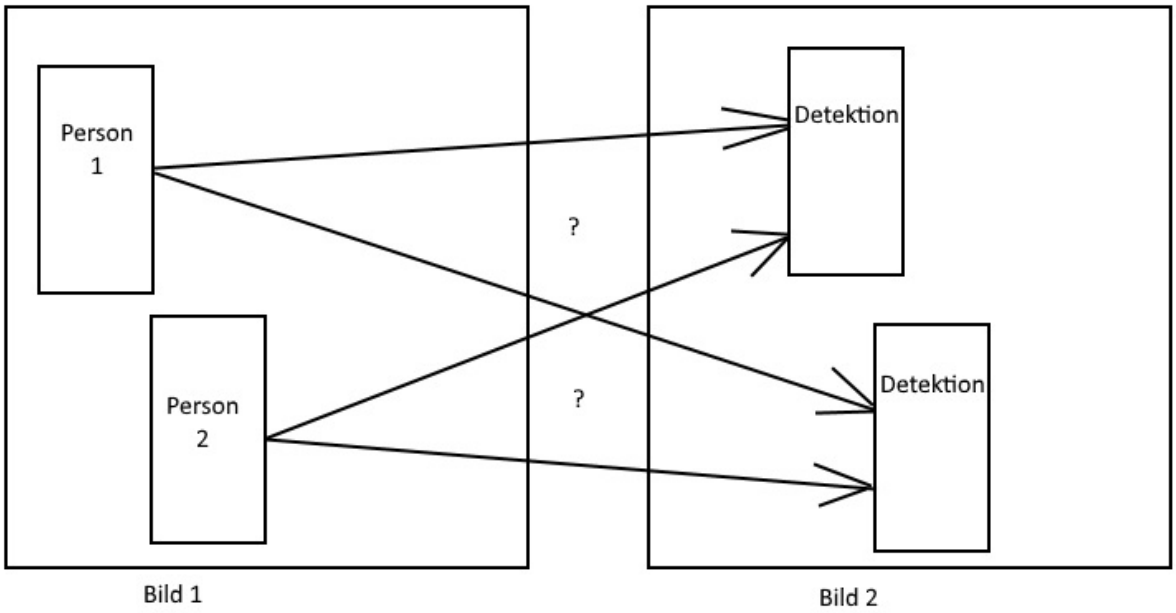



Abbildung 2: Skizzierung des Zuordnungsproblems. In der Zeit zwischen Bild 1 und Bild 2 haben sich die als Bounding Boxen dargestellten Personen bewegt. In Bild 2 sind nur die Positionen der neuen Detektionen bekannt, nicht aber, welche Person zu welcher Detektion gehört.

Detektionsbasiertes Tracking (engl.: Tracking-by-Detection) bezeichnet die Herstellung zeitlicher Korrespondenzen zwischen Detektionen von Personen in Bildern in der Form von Trajektorien und besteht im Wesentlichen aus der Detektion der zu trackenden Objekte

und der Zuordnung von Detektionen zu Trajektorien. Um Tracking in Echtzeit durchführen zu können, wird dies häufig mit einem rekursiven Zustandsfilter (z.B. Kalman-Filter) kombiniert, wodurch Trajektorien geglättet und Positionen von Personen prädiziert werden können. Die Umsetzung dieser Teilaufgaben kann durch unterschiedliche Ansätze erfolgen. Der im Folgenden beschriebene Ansatz entspricht dem in Klinger and Muhle [2012] verwendeten Ansatz, da dieser die Grundlage für die vorliegende Arbeit bildet.
Die Detektion wird sowohl durch einen Personendetektor als auch durch die Klassifikation mit einem Online Random Forest umgesetzt. Genutzt wird der von Dalal and Triggs [2005] entwickelte Personendetektor, der im Open-Source-Projekt *OpenCV* von Bradski [2000] zur Verfügung steht.

Da der Detektor allerdings nicht immer die pixelgenaue Position einer Person liefert, wird der Klassifikator im Bereich der prädizierten Position einer Person angewendet, um für die umliegenden Pixel die jeweilige Wahrscheinlichkeit zu bestimmen, mit der sich die Person an der Stelle des Pixels befindet. Die aus Detektor und Klassifikator ermittelte Position geht als Messung in einen Kalmanfilter ein, welcher die Prädiktion zukünftiger Positionen realisiert. Für die Prädiktion wird angenommen, dass sich die Personen von Bild zu Bild, also über einen kurzen Zeitraum, geradlinig und mit konstanter Geschwindigkeit im Raum bewegen. Die prädizierte Position setzt sich aus der zuletzt prädizierten Position und der Positionsmessung durch die Detektion zusammen. Die Trajektorien ergeben sich aus dem Fußpunkt, d.h. der Mitte der unteren Seite der Bounding Box einer detektierten Person, da angenommen wird, dass sich die Füße einer Person dort befinden. Die Bounding Box ist das Ergebnis der Detektion. Die Zuordnung zwischen Detektionen und Trajektorien wird durch den Klassifikator und der Prädiktion umgesetzt: Abhängig vom Klassifikationsergebnis einer detektierten Person und äbhängig vom Abstand zwischen Detektion und prädizierter Position einer Trajektorie werden Kennzahlen berechnet, die jede mögliche Kombination der Detektionen und Trajektorien bewertet. Das Zuordnungsproblem ist in Abbildung 2 skizziert.

# 3 Methodik zur Untersuchung der Merkmalskonstellation und der Überbrückung von Verdeckungen

Die Idee hinter den im Folgenden vorgestellten Methoden ist, dass ein detektionsbasiertes Trackingverfahren durch Nutzen eines geeigneten Klassifikators hinsichtlich der Wiedererkennung von Personen nach einer Verdeckung verbessert werden kann. Dazu werden die zu trackenden Personen durch Klassen in einem ORF repräsentiert. Im Folgenden werden Ansätze vorgestellt, wie sich Personen im Merkmalsraum darstellen lassen, um diese durch Anwendung des Klassifikators eindeutig wiederzuerkennen. Außerdem werden hinsichtlich der Überbrückung von Verdeckungen verschiedene Ansätze untersucht, wie der Klassifikator für die Zuordnung von Detektionen zu Trajektorien genutzt werden kann.

## 3.1 Untersuchung möglicher Merkmalskonstellationen für den Klassifikator

Für die Merkmalsextraktion ist es das Ziel, ein Modell zur Unterteilung der Bounding Box zu finden, welches die Kontur eines Menschen möglichst gut beschreibt. Dabei ist auch die Größe des Merkmalsvektors zu beachten, also die Anzahl der Kennzahlen, mit denen eine Person beschrieben wird: Werden die Farbinformationen eines Menschen im Bild nicht zusammengefasst, sondern direkt als Kennzahlen zur Klassifikation verwendet, so nimmt der Merkmalsvektor große Dimensionen an, was das Trainieren des ORF erschwert, da mehr Trainingsdaten benötigt werden. Werden die Informationen zu sehr zusammengefasst, sind die Kennzahlen möglicherweise zu generalisiert, wodurch die Personen nicht mehr korrekt klassifiziert werden können. Da die Bounding Boxen als Ergebnis des Detektors nicht immer die gleiche Größe haben, werden diese für jede Merkmalskonstellation auf einheitliche Größen, im Rahmen dieser Arbeit meistens 24x48 Pixel, gestreckt bzw. gestaucht. Dies wird vor allem dafür benötigt, dass die Merkmalsvektoren, die die Personen beschreiben, für eine Konstellation gleiche Dimensionen haben und damit überhaupt durch einen ORF zu verarbeiten sind.

### 3.1.1 Nutzung verschiedener Farbräume

Da alle Berechnungen für die Merkmale primär auf den Farbwerten der Pixel beruhen, stellt sich die Frage, welcher Farbraum verwendet wird. Das Problem bei der Verwendung des RGB-Systems ist, dass sich die Positionen der Merkmale eines Objektes im RGB-Farbraum bzw. Merkmalsraum stark ändert, wenn sich die Belichtung des Objektes ändert. Da sich die Menschen in den Bildsequenzen bewegen, ist davon auszugehen, dass sich die Farbwerte der Menschen ebenso ändert. Eine Alternative zum RGB-System

ist das HSV-System, welches einen Pixel durch den Farbwert (Hue), die Sättigung (Saturation) und die Helligkeit (Value) beschreibt. Die Annahme ist, dass sich bei anderer Beleuchtung nur der Helligkeitsanteil des Pixels ändert, nicht aber der Farbwert und die Sättigung, wodurch die Veränderung der Position im Merkmalsraum geringer ist, als bei der Verwendung des RGB-Systems.

### 3.1.2 Ellipse

Die Idee hinter der Ellipsen-Merkmalskonstellation aus Klinger and Muhle [2012] ist, dass sich der Bereich im Bild, in dem sich eine Person befindet, von der Bounding Box auf eine Ellipse reduzieren lässt. Die Person ist damit im Ausschnitt enthalten, allerdings entfallen hierbei einige Pixel, die nicht die Person, sondern den Hintergrund darstellen. Das Reduzieren der Hintergrundpixel im verarbeiteten Ausschnitt hat den Vorteil, dass die Informationen, die die Person selbst nicht beschreiben, nicht zur Klassifizierung verwendet werden. Die große bzw. kleine Achse der Ellipse haben dabei die Länge der langen bzw. kurzen Seite der Bounding Box. Der Merkmalsvektor setzt sich aus den Pixeln innerhalb der Ellipse zusammen, wobei ein Pixel drei Werte beschreibt, da ein dreidimensionaler Farbraum verwendet wird. Für eine Bounding Box der Größe 24x48px hat der Merkmalsvektor 2847 Dimensionen. Dieses Modell für die Merkmalskonstellation ist in Abbildung 3 skizziert.

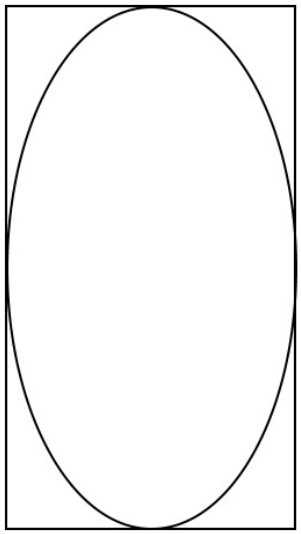


Abbildung 3: Skizzierung der Ellipsen-Merkmalskonstellation. Die Farbwerte der Pixel innerhalb der Ellipse stellen die Merkmale dar.

### 3.1.3 Querstreifen

Der Querstreifen-Konstellation liegt neben der Annahme, dass sich eine Person hauptsächlich in der Mitte der Bounding Box befindet, vor allem die Intuition zugrunde, dass

sich Menschen gut in der Vertikalen beschreiben lassen. Dadurch können Merkmale wie Bekleidungsfarbe für Oberteil und Hose sowie die Haarfarbe in den Merkmalsvektor zusammengefasst eingearbeitet werden. Realisiert wird dies zunächst durch die Betrachtung einer einzelnen Zeile der Bounding Box für eine Dimension des Farbraums. Aus den Pixeln der Zeile werden der Mittelwert und die Standardabweichung unter der Annahme berechnet, dass diese beiden Kennzahlen die Eigenschaften der Zeile gut wiedergeben, ohne alle Farbwerte der Zeile in den Merkmalsvektor aufzunehmen. Ein weiterer Vorteil ist die Reduzierung der Dimensionen des Merkmalsraums, wodurch weniger Trainingsdaten für den Klassifikator benötigt werden. Der Merkmalsvektor setzt sich hier aus Mittelwert und Standardabweichung für jede Zeile und Farbdimension zusammen, so dass sich für eine Bounding Box der Größe 24x48px ein Merkmalsvektor mit 288 Dimensionen ergibt (48 Zeilen á 3 Farbdimensionen á 2 Kenngrößen).
Die Annahme, dass sich eine Person hauptsächlich in der Mitte der Bounding Box befindet, wird dadurch berücksichtigt, dass die Farbwerte zur Berechnung des Mittelwertes durch eine Gauß-Verteilung horizontal gewichtet werden; die Gewichtung g ergibt sich dabei durch die Formel

$$g = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{(c-\mu)^2}{2 \cdot \sigma^2}} \tag{1}$$

wobei $c \in [0, ..., 23]$ der Index des jeweiligen Pixels der Zeile, $\mu = 11.5$ der mittlere Index einer Zeile und $\sigma$ die zu wählende Standardabweichung der Verteilung ist. Die Gewichtung hat zur Folge, dass die Farbwerte am linken und rechten Rand der Bounding Box, wo sich eine Person der Annahme nach nicht befindet, weniger in die Berechnung der Merkmale eingehen, als die Farbwerte in der Mitte der Bounding Box, wo sich eine Person der Annahme nach befindet. Die Form dieser Merkmalskonstellation ist in Abbildung 4 skizziert.

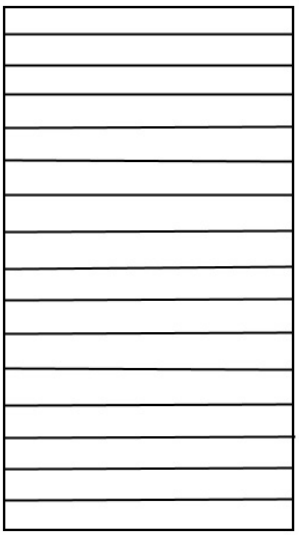


Abbildung 4: Skizze der Querstreifen-Merkmalskonstellation. Die Bounding Box wird vertikal in Streifen gegliedert.

### 3.1.4 Einteilung in Regionen

Die Einteilung der Bounding Box in Regionen stellt ein anderes Modell der Annahme, dass sich Personen in der Mitte der Bounding Box befinden und damit die Pixel am Rand der Bounding Box weniger gewichtet werden als in der Mitte, dar. Zusätzlich werden die Annahmen nach Bammes [1990] über die Figur des Menschen verwendet, um die Bounding Box in Regionen einzuteilen, die den Oberkörper und die Beine darstellen (s. Abb. Abbildung 5). Die Einteilung erfolgt dabei nur entlang der Vertikalen; in der Horizontalen sind die Regionen 12 Pixel lang, was der Hälfte der Bounding Box entspricht, und sind in der Mitte der Bounding Box gelagert. Der Bereich des Kopfes wird bei dieser Konstellation außer Acht gelassen, da es mit der Auflösung der Bilder nicht möglich ist, genügend charakteristische Merkmale aus einem Gesicht oder der Haarfarbe zu entnehmen, wie in Farenzena et al. [2010] gezeigt wurde.
Die Merkmale für diese Konstellation setzen sich aus Mittelwert und Standardabweichung der beiden Regionen in jeder Farbdimension zusammen, so dass der Merkmalsvektor hier 12 Dimensionen (2 Kenngrößen á 2 Regionen á 3 Farbdimensionen) hat. Die Form dieser Konstellation ist in Abbildung 6 skizziert.

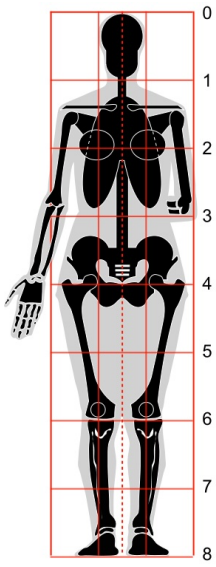



Abbildung 5: Einteilung des Menschen[1]

[1] Bildquelle: `upload.wikimedia.org/wikipedia/commons/4/4b/Human_body_proportions2_svg.svg`, abgerufen am 6. September 2016

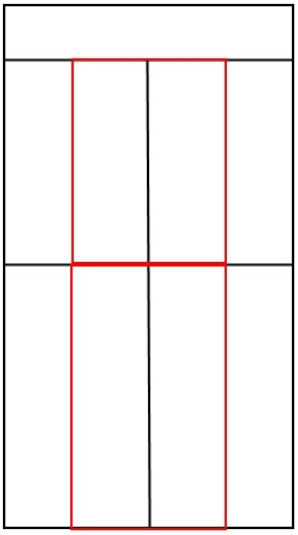


Abbildung 6: Skizze der Einteilung in Regionen. Aus den Pixeln innerhalb der roten Rechtecke werden jeweils Mittelwert und Standardabweichung für jede Farbdimension als Merkmale verwendet.

### 3.1.5 Einteilung in symmetrische Regionen

Die Einteilung in symmetrische Regionen stellt eine Erweiterung der Einteilung in Regionen dar und wurde ähnlich wie in Farenzena et al. [2010] umgesetzt. Die vertikale Einteilung der Bounding Box bleibt wie bei der Einteilung in Regionen unter der Annahme erhalten, dass die Variationen der Pose eines Menschen vor allem in der Horizontalen auftreten, weshalb in der Horizontalen für beide Regionen (Oberkörper und Beine) Symmetrieachsen gesucht werden. Die Achsen liegen rechnerisch zwischen den Pixeln.

Die Berechnung der Symmetrieachsen unterliegt der Annahme, dass die Differenz der Farbwerte aus zwei Teilregionen, die sich an einer Symmetrieachse gegenüberliegen, minimal wird. Die Differenz der Farbwerte wird dabei zeilenweise berechnet und über alle Zeilen der Regionen und allen Farbdimensionen aufsummiert (s. Abb. 7). Die Differenzsumme wird für jede mögliche Lage der Symmetrieachse berechnet. Bei einer Größe der Bounding Box von 24x48px ist die Box in der horizontalen 24px lang, die horizontale Größe der genutzten Regionen beträgt 12px. Werden die Pixel von links nach rechts bei 0 beginnend gezählt, so beschränkt sich die horizontale Lage der vertikal verlaufenden Symmetrieachse auf die Position zwischen dem 5. und 6. Pixel, wenn sie am weitesten links liegt, und auf die Position zwischen dem 17. und 18. Pixel, wenn sie am weitesten rechts liegt. Die Symmetrieachsen werden an den Stellen gewählt, an denen die Differenzsummen minimal sind.

Die Extraktion der Merkmale aus den Regionen erfolgt wie bei der Einteilung in Regionen. Die Einteilung in symmetrische Regionen ist in Abbildung 8 skizziert.

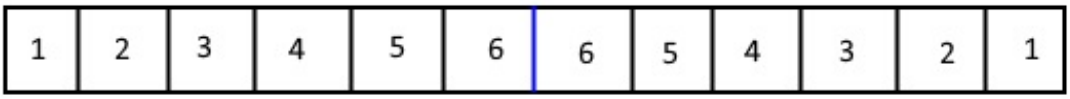



Abbildung 7: Skizze der Berechnung der Symmetrieachse für eine Zeile. Aus den Pixeln mit gleicher Nummer wird in allen Farbdimensionen die Differenz berechnet und für alle Pixelpaare aufsummiert. In blau dargestellt ist die angenommene Symmetrieachse.

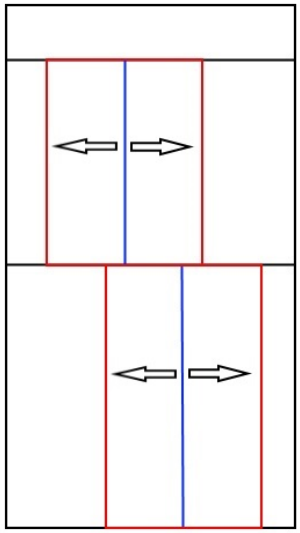


Abbildung 8: Skizze der Einteilung in symmetrische Regionen. Aus den Pixeln innerhalb der roten Rechtecke werden jeweils Mittelwert und Standardabweichung für jede Farbdimension als Merkmale verwendet. Die blauen Linien stellen die Symmetrieachsen im Ober- bzw. Unterkörperbereich dar. Die Pfeile stellen dar, dass die Symmetrieachsen verschieden liegen können.

## 3.2 Das Zuordnungsproblem und die Überbrückung von Verdeckungen

Ein Problem bei Tracking-by-Detection-Verfahren stellt die Unterbrechung von Trajektorien (Fragmentations) dar: Wird eine Person aufgrund einer Verdeckung nicht detektiert, so kann ihre Trajektorie für eine bestimmte Anzahl an Bildern ohne Detektion durch die Prädiktionen weitergeführt werden, wie es in Klinger and Muhle [2012] umgesetzt ist. Dass die Trajektorie nicht verworfen wird, ermöglicht das Weiterführen der Trajektorie nach der Verdeckung, wenn die Person wiedererkannt wird. Da die Annahme, dass sich Personen über einen kleinen Zeitraum mit konstanter Geschwindigkeit geradlinig fortbewegen, für längere Zeiträume nicht mehr zutrifft, werden die Prädiktionen mit der Zeit immer ungenauer, wodurch ein Wiedererkennen der Person nur durch die Nähe ihrer Detektion zur Prädiktion unwahrscheinlich ist. Wird die Person bei erneuter Detektion nicht wiedererkannt, wird sie durch eine neue Trajektorie fortgeführt. Die Trajektorien, die diese Person beschreiben, sind damit fragmentiert. Die im Folgenden vorgestellten Ansätze

werden vor dem Hintergrund entwickelt, Trajektorien auch nach einer Verdeckung korrekt fortzuführen, indem der Klassifikator und die Prädiktion zur Lösung des Zuordnungsproblems unterschiedlich gewichtet werden. Die Zuordnung von Detektionen zu Trajektorien muss unter bestimmten Bedingungen erfolgen: Da jede Trajektorie nur genau einen Menschen beschreibt und jeder Mensch nur durch genau eine Trajektorie beschrieben wird, darf jeder Detektion nur eine Trajektorie und jeder Trajektorie nur eine Detektion zugeordnet werden. Eine Ausnahme stellt die Verdeckung einer Person dar: Wird eine Person verdeckt und damit nicht detektiert, wird die Trajektorie der Person mit den Prädiktionen weitergeführt werden, sofern der Trajektorie nicht (fälschlicherweise) eine andere Detektion zugeordnet wird. Das Zuordnungsproblem ist in Abbildung 9 skizziert.

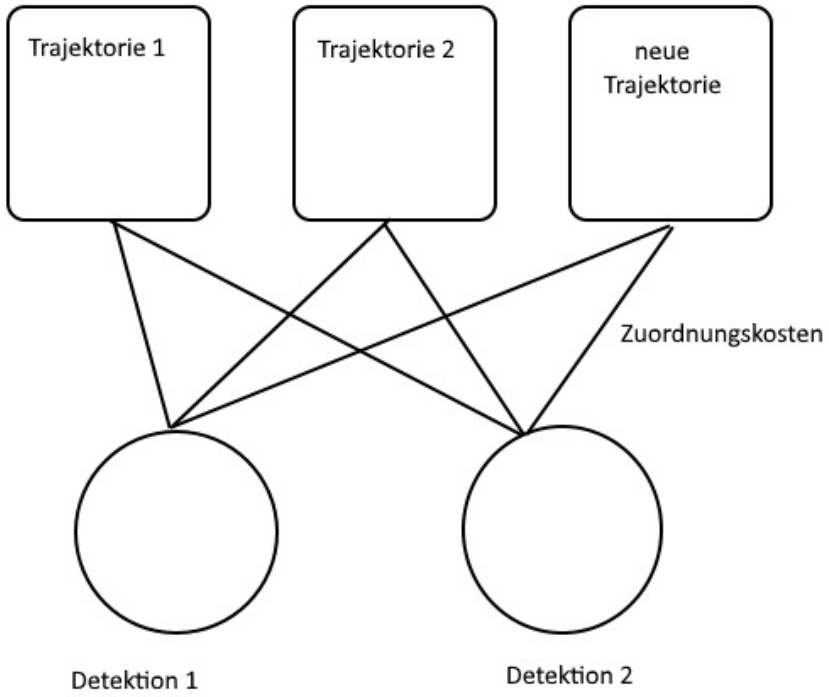



Abbildung 9: Skizzierung des Zuordnungsproblems. Für jede Kombination aus Trajektorie und Detektion werden aus der Prädiktion und der Klassifikation Zuordnungskosten berechnet. Einer Detektion wird die Trajektorie zugeordnet, die in Kombination die größten Kosten haben. Für jede Detektion muss auch überprüft werden, ob eine neue Trajektorie erstellt werden muss. Dieser Fall tritt z.B. dann auf, wenn eine neue Person ins Bild tritt.

Zum Treffen der Zuordnungen besteht der Ansatz, über alle bestehenden Trajektorien und alle aktuellen Detektionen zu iterieren. Für jede Kombination werden dabei zwei Ähnlichkeitsmaße berechnet: Da angenommen wird, dass sich Menschen nicht zufällig durch den Raum bewegen, sondern sich tendenziell mit konstanter Geschwindigkeit fortbewegen, kann die zukünftige Position einer Person durch den Kalman-Filter prädiziert werden. Diese Prädiktion bezieht sich also auf eine bereits vorhandene Trajektorie. Für jede Kombination aus Trajektorie und Detektion wird die räumliche Distanz zwischen

Prädiktion und Detektion berechnet und als Ähnlichkeitsmaß angesehen; je näher die Detektion an der Prädiktion liegt, desto größer ist die Wahrscheinlichkeit, dass diese beiden die selbe Person beschreiben. Dies wird problematisch, sobald Menschen nahe beieinander laufen, da es hierbei zu einem Identity Switch kommen kann. Dies beschreibt den Fall, dass die Trajektorie einer Person der Detektion einer anderen Person zugeordnet wird und stellt somit ein wesentliches Problem des Trackings dar. Ein Beispiel für einen Identity Switch ist in Abbildung 10 dargestellt. Um dieses Problem zu umgehen, wird ein zweites Ähnlichkeitsmaß berechnet.
Da jede Trajektorie Merkmals-Informationen über ihre jeweilige Person enthält, kann durch das Klassifizieren der Detektion überprüft werden, mit welcher Wahrscheinlichkeit der detektierte Mensch als der Mensch, den die Trajektorie repräsentiert, klassifiziert wird. Das Ähnlichkeitsmaß entspricht dem relativen Klassenanteil im Histogramm als Ergebnis des ORF.
Die Ähnlichkeitsmaße werden als Sicherheiten der Prädiktion $S_P$ und der Klassifikation $S_K$ für jede Kombination aus Trajektorie und Detektion berechnet und zu einer Gesamtsicherheit $S_G$ zusammengefasst. Es wird jene Kombination für die Zuordnung gewählt, welche die höchste Gesamtsicherheit aufweist. Im Folgenden werden vier Möglichkeiten untersucht, wie die Sicherheiten zusammengefasst werden können.

(a) (b) (c)

Abbildung 10: Beispiel für einen Identity Switch in der verarbeiteten Bildsequenz. Bild (a) zeigt zwei aufeinander zulaufende Personen (türkise und braune Bounding Box). Bild (b) zeigt die Verdeckung der Person mit der braunen Bounding Box durch die Person mit der türkisen Bounding Box. In Bild (c) wurde die türkise Trajektorie fälschlicherweise mit der Person fortgesetzt, die vorher durch die braune Trajektorie beschrieben wurde.

### 3.2.1 Multiplikativer Ansatz

Dieser Ansatz entspricht dem in Klinger and Muhle [2012] vorgestellten Ansatz zur Berechnung der Zuordnungskosten bzw. Gesamtsicherheiten der Zuordnungskombinationen.

Hierbei berechnet sich die Gesamtsicherheit durch

$$S_G = S_P \cdot S_K. \tag{2}$$

Der Vorteil dieses Ansatzes ist, dass die Gesamtsicherheit klein ist, wenn $S_P$ oder $S_K$ klein ist. Sind z.B. zwei Personen im Bild, die nebeneinander herlaufen, so wird die Prädiktionssicherheit für die Zuordnung beider Trajektorien zu beiden Detektionen ähnlich groß. Die Klassifikatorsicherheit für die falsche Trajektorie-Detektion-Kombination wird kleiner sein, als die Klassifikatorsicherheit für die korrekte Kombination, wodurch die Gesamtsicherheit der richtigen Kombination größer ist als die Gesamtsicherheit der falschen Kombination.

### 3.2.2 Additiver Ansatz

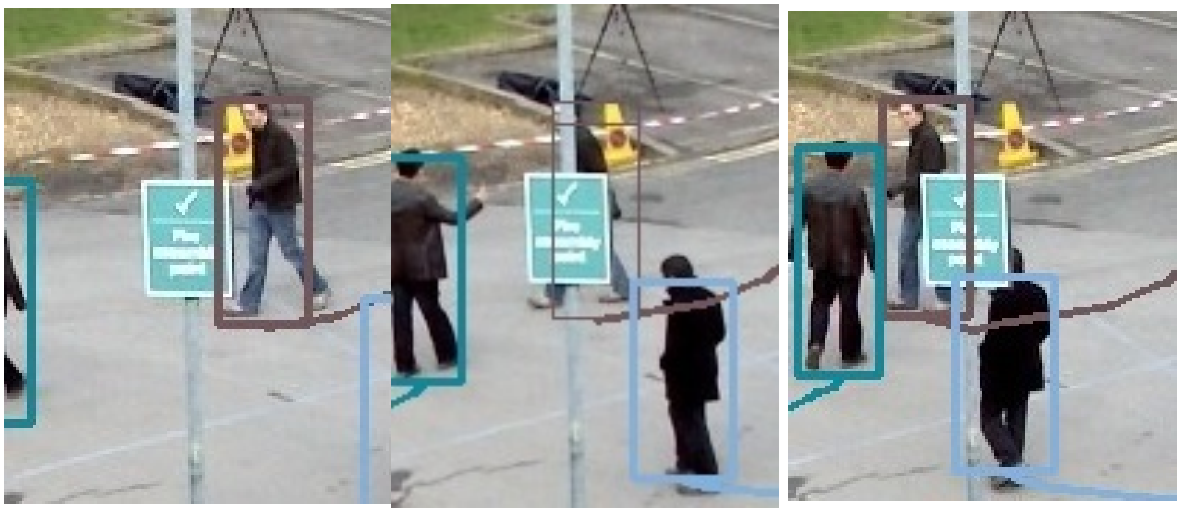
In diesem Ansatz berechnet sich die Gesamtsicherheit wie folgt:

$$S_G = S_P + S_K \tag{3}$$

Der Vorteil dieses Ansatzes ist der Nachteil des multiplikativen Ansatzes: Sollte sich eine Person nicht linear bewegen, so bleibt die Gesamtsicherheit groß, wenn die Klassifikatorsicherheit groß ist. Sie wird nicht negativ durch die Prädiktionssicherheit beeinflusst. Dieser Ansatz stellt eine Kontrolle der Addition der Sicherheiten für die folgenden Ansätze dar, da diese ähnlich durchgeführt werden.

### 3.2.3 Gewichteter Ansatz

Für diesen Ansatz wird die Eigenschaft der Trajektorien genutzt, dass diese nicht nur die prädizierte Position und die Merkmale ihrer entsprechenden Person speichert, sondern auch die Anzahl der Bilder, seit dem die Person nicht detektiert wird. Dass eine Person nicht detektiert wird, kann z.B. daran liegen, dass sie durch einen Gegenstand oder eine andere Person verdeckt ist, wie es in Abbildung 11 gezeigt ist. In diesem Fall wird die Trajektorie nicht sofort verworfen, sondern noch eine bestimmte Anzahl an Bildern lang mit den prädizierten Positionen fortgeführt. Die Anzahl der Bilder wird dabei als Parameter *maximal absence-count (maxabsc)* definiert, die Anzahl der Bilder, in denen eine Person nicht detektiert wird, wird für jede Trajektorie als *absc* bezeichnet. In den folgenden beiden Ansätzen werden die Prädiktionssicherheit und die Klassifikatorsicherheit in Abhängigkeit von der Anzahl der Bilder, in denen eine Person nicht detektiert wurde, gewichtet. Die Idee dabei ist, dass die Prädiktion stärker gewichtet wird, je kleiner *absc* ist, da sich die Person dann tendenziell näher an der Prädiktion befindet. Je größer *absc* wird, desto mehr wird der Klassifikator gewichtet, da dieser unabhängig von der Position der Person eine Zuordnung treffen kann.



(a) (b) (c)

Abbildung 11: Sequenz einer Verdeckung mit Wiedererkennung der Person. Bild (a) und (c) zeigen die detektierte Person (braune Bounding Box) vor und nach der Verdeckung. In Bild (b) wurde die Person nicht detektiert. Die für diesen Fall dünner dargestellte Bounding Box wird an der prädizierten Position weitergeführt.

**Teilgewichtung** Für die Teilgewichtung wird angenommen, dass es eine bestimmte Anzahl an möglichen Gewichtungsstufen gibt. Die Gewichtungen setzen sich aus den $maxabsc$ Fällen der nicht-Detektion und dem Fall der Detektion zusammen, so dass es $maxabsc+1$ mögliche Gewichtungsstufen gibt. Die Gesamtsicherheit $S_G$ berechnet sich wie folgt:

$$S_G = S_P \cdot \frac{(maxabsc - absc + 1)}{maxabsc + 1} + S_K \cdot \frac{(absc + 1)}{maxabsc + 1} = S_P \cdot G_P + S_K \cdot G_K \tag{4}$$

Dabei sind $G_P, G_K$ die Gewichtungen der Prädiktion respektive der Klassifikation. Zur Veranschaulichung sind die Gewichtungen in Abbildung 12 dargestellt.

**Sigmoid-Gewichtung** Für diese Gewichtungsart wird die Gewichtung der Prädiktion $G_P$ aus der Sigmoid-Funktion berechnet:

$$G_P = \frac{1}{1 + e^{b \cdot absc}} \tag{5}$$

Dabei ist $b$ ein Parameter der Funktion, der bestimmt, wie schnell sich der Funktionswert ändert. Die Gewichtung der Klassifikation $G_K$ berechnet sich wie folgt:

$$G_K = 1 - G_P \tag{6}$$

Zur Veranschaulichung sind die Gewichtungen in Abbildung 13 dargestellt.

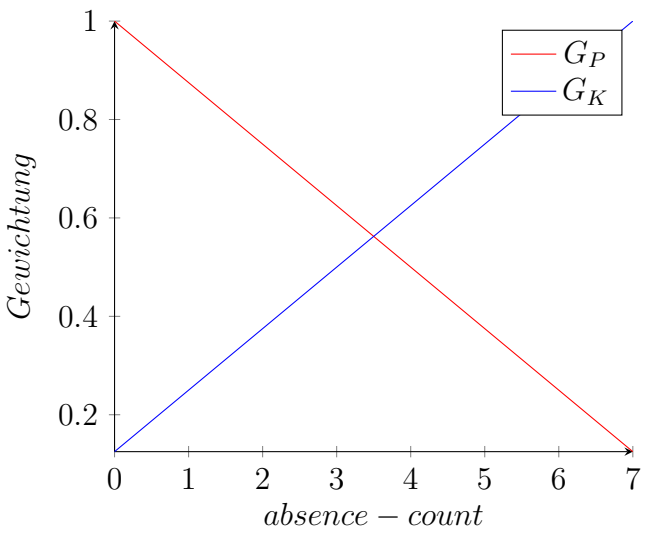



Abbildung 12: Teilgewichtung für $maxabsc = 7$

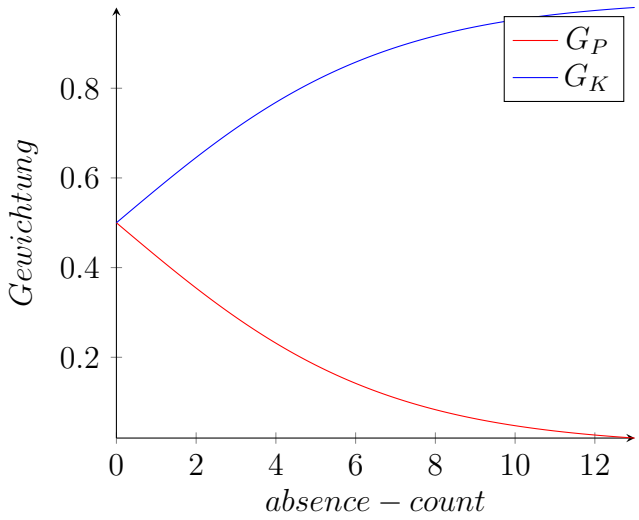



Abbildung 13: Sigmoid-Funktion für b=0.3

# 4 Experimentelle Untersuchung der entwickelten Methodiken

(a) (b)

Abbildung 14: Sequenzausschnitt aus dem Datensatz PETS09-S2L1

In (a) ist ein nicht-bearbeiteter Sequenzausschnitt gezeigt. In (b) ist der gleiche Ausschnitt mit berechneten Trajektorien und Bounding Boxen gezeigt.

Die in der vorliegenden Arbeit entwickelten Methodiken werden mit dem in Leal-Taixé et al. [2015] erarbeiteten Benchmarkdatensatz $PETS09 - S2L1$ experimentell überprüft und mit den in Bernardin and Stiefelhagen [2008] entwickelten Metriken $MOTA$ und $MOTP$ bewertet. In Abbildung 14 ist ein Ausschnitt aus der Sequenz dargestellt. Die Multiple Object Tracking Accuracy (MOTA) beschreibt die semantische Genauigkeit des Trackings:

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{7}$$

wobei $m_t$, $fp_t$ und $mme_t$ jeweils die Anzahl der verpassten Detektionen (Person im Bild wird nicht detektiert), der falsch-positiven Detektionen (Detektion einer Person im Bild, wo keine ist) und der Fehlzuordnungen (Identity Switch) zum Zeitpunkt $t$ darstellt; $g_t$ ist die Anzahl aller Personen im Bild zum Zeitpunkt $t$. Je größer der $MOTA$ ist, desto besser ist die semantische Genauigkeit des Trackings.

Die Multiple Object Tracking Precission (MOTP) beschreibt die geometrische Genauigkeit des Trackings:

Für die Auswertung der Experimente wird diese Metrik leicht verändert umgesetzt:

$$MOTP = 1 - \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{8}$$

Dabei ist $d_t^i$ die Distanz zwischen der prädizierten und der (durch den Benchmarkdatensatz bekannten) wahren Position zum Zeitpunkt $t$; $c_t$ ist zum Zeitpunkt $t$ die Anzahl aller Zuordnungen zwischen Detektionen und Trajektorien, also die Anzahl der Personen, für die $d_t^i$ berechnet werden kann.

Diese Umsetzung ermöglicht, die Metrik wie die Größe $MOTA$ zu interpretieren: Je größer der Wert, desto besser ist die geometrische Genauigkeit des Trackings.

## 4.1 Merkmalskonstellationen

Für die Bildsequenz sind die Positionen aller Personen im Bild bekannt, so dass für jede Klassifikation überprüft werden kann, ob diese korrekt ist. Die Bewertungsgröße für den Vergleich der Merkmalskonstellationen ist das prozentuale Verhältnis zwischen richtigen und allen Klassifikationen in der Bildsequenz. Zur Überprüfung der Annahme über den zu verwendenden Farbraum wurden die Experimente für die Querstreifen und die Ellipse sowohl im RGB- als auch im HSV-System durchgeführt. Die Ergebnisse sind in den jeweiligen Abschnitten mit aufgeführt.

### 4.1.1 Ellipse

Untersucht wurden die Ellipsen für Bounding Boxen der Größe 24x48px, 12x24px, 6x12px und 3x6px. Die Skalierung der Ellipsen wird vor dem Hintergrund untersucht, dass bei kleineren Ellipse der resultierende Merkmalsvektor deutlich kleinere Dimensionen annimmt: Für die in Klinger and Muhle [2012] verwendete Ellipsengröße von 24x48px hat der Merkmalsvektor 2847 Dimensionen, für die kleineren Skalierungen hat der Vektor 711, 177 respektive 42 Dimensionen. Die Ergebnisse der Untersuchung sind in Abbildung 15 dargestellt.

Die Abbildung 15 zeigt, dass die Verwendung des RGB-Farbraums zu deutlich schlechteren Ergebnissen führt. Ebenso ist zu erkennen, dass die Ellipsen der Größen 24x48, 12x24 und 6x12 ähnlich gute Ergebnisse liefern. Da spätere Experimente vor allem auf die ursprüngliche Ellipse der Größe 24x48 bezogen werden, und da diese nicht wesentlich schlechtere Ergebnisse liefert als die anderen Größen, wird diese Konstellation für die weiterführenden Experimente verwendet.

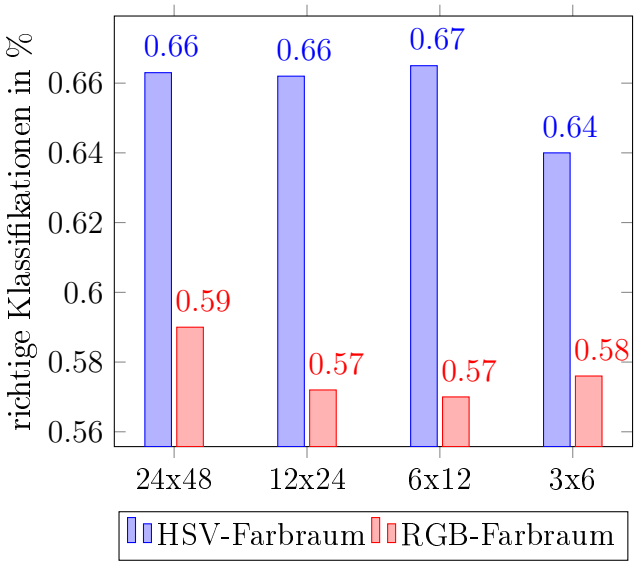



Abbildung 15: Untersuchung der Ellipsengröße und des Farbraums.

### 4.1.2 Querstreifen

In Bezug auf die horizontale Gewichtung der Pixel bei der Berechnung des Mittelwertes wird untersucht, welche Standardabweichung für die Gauß-Gewichtung bessere Ergebnisse liefert. Untersucht wurden diese Gewichtungen für eine Standardabweichung von 2 und 4, die Ergebnisse sind in Abbildung 16 dargestellt.
Untersucht wird auch, ob das vertikale Zusammenfassen der Farbinformationen die Genauigkeit des Klassifikators steigert. Die Zusammenfassung führt dazu, dass die Dimensionen des Merkmalsvektors weiter reduziert werden, und so weniger Trainingsdaten für den Klassifikator benötigt werden. Dazu wurden die Bounding Boxen in der Höhe auf 24, 12, 6 und 3 Pixel gestaucht, woraus sich entsprechend kleinere Merkmalsvektorgrößen von 144, 72, 36 bzw. 18 Dimensionen ergaben; bei einer Höhe der Bounding Box von 48 Pixel ist der Merkmalsvektor entsprechend 288-dimensional. Bei der Stauchung werden die Farbwerte vertikal zusammengefasst. Die Ergebnisse dieser Untersuchung sind in Abbildung 17 dargestellt.

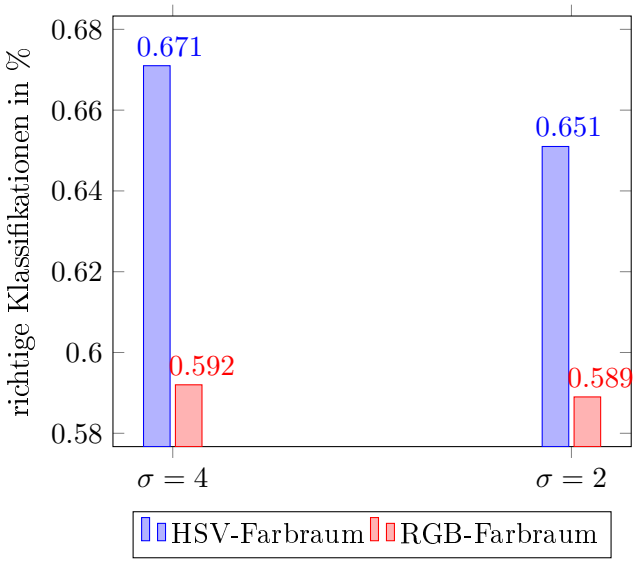



Abbildung 16: Untersuchung der Gaußverteilung und des Farbraums für Querstreifen. Diese Experimentreihe wurde mit einer Bounding Box-Höhe von 48px durchgeführt.

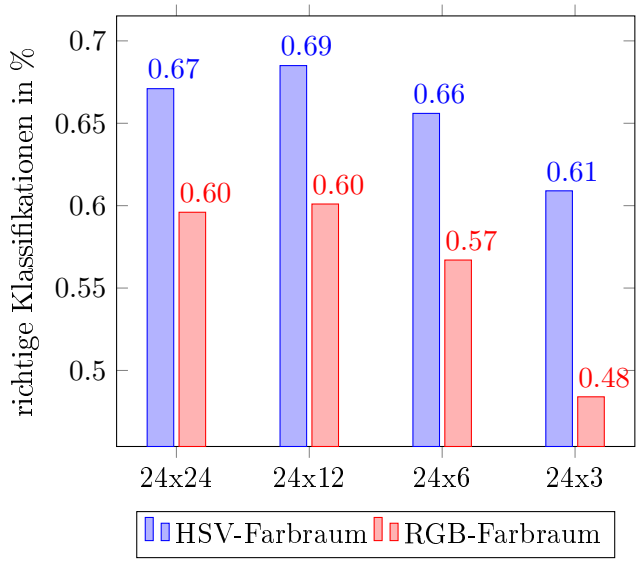



Abbildung 17: Untersuchung hinsichtlich der Größe und des Farbraums für Querstreifen.

In Abbildung 16 ist zu sehen, dass die Gewichtung mit $\sigma = 4$ zu besseren Ergebnissen führt, weshalb diese Gewichtung für weiterführende Experimente gewählt wird. Außerdem ist zu erkennen, dass die Verwendung des HSV-Farbraums bessere Ergebnisse liefert. Die Ergebnisse der Experimente in Abbildung 17 zeigen, dass sich vertikale Zusammenfassung von Informationen positiv auf die Klassifikatorgenauigkeit auswirkt. Für die weitergehenden Experimente wird daher die Konfiguration der vertikalen Stauchung auf 24x12 Pixel gewählt. Die Ergebnisse der Untersuchung bezüglich des Farbraums sprechen dafür, für die folgenden Experimente den HSV-Farbraum zu verwenden, da dessen Verwendung in

allen bisherigen Untersuchungen bessere Ergebnisse liefert, als die Nutzung des RGB-Farbraums.

### 4.1.3 Einteilung in Regionen

Da es für die Einteilung in Regionen keine veränderlichen Parameter gibt, die die Form der Einteilung oder die Berechnung der Kennzahlen des Merkmalsvektors beeinflussen, werden für diese Merkmalskonstellation keine vergleichenden Untersuchungen angestellt; die Einteilung in Regionen wird in der vorgestellten Form ab Kapitel 4.2 mit den anderen Konstellationen verglichen.

### 4.1.4 Einteilung in symmetrische Regionen

Für die Einteilung in symmetrische Regionen wurde überprüft, ob eine Gewichtung durch eine Gauß-Verteilung bessere Ergebnisse liefert; je näher Pixel an der Symmetrieachse liegen, desto mehr werden sie bei der Berechnung eines Mittelwertes gewichtet. Die Verteilung wurde hinsichtlich der Standardabweichung untersucht. Die Ergebnisse der Untersuchung sind in Abbildung 18 dargestellt.

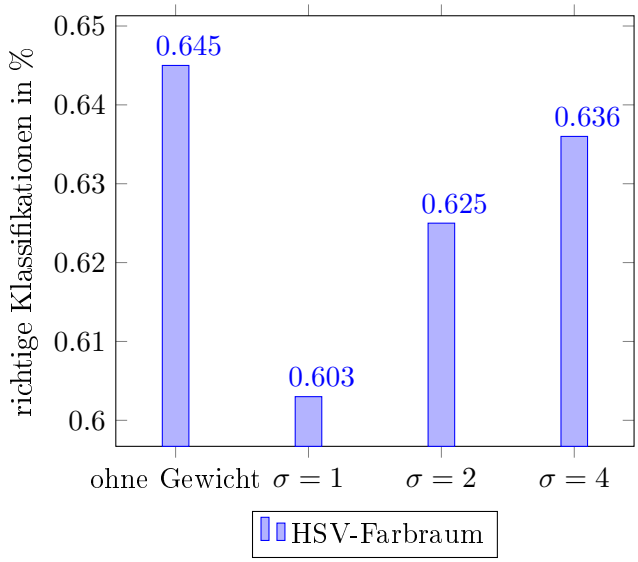



Abbildung 18: Untersuchung der Gaußverteilung für symmetrische Regionen.

Die Experimente zeigen, dass die Ergebnisse für eine höhere Standardabweichung besser werden. Wird die Standardabweichung bei der Gaußverteilung Unendlich, so sind die Gewichtungsanteile für die betrachteten Pixel gleich, was wiederum dem Ansatz ohne Gewichtung entspricht, weshalb für diese Merkmalskonstellation die Konfiguration ohne Gewichtung gewählt wird.

## 4.2 Zuordnungsproblem

Die im Folgenden vorgestellten Experimente wurden mit dem in der MOT-Challenge zur Verfügung gestellten Matlab-Skript ausgewertet. Die Metriken für den Vergleich der semantischen und geometrischen Genauigkeit des Trackings stellen die bereits vorgestellten Größen *MOTA* und *MOTP* dar. Beide Metriken können im Rahmen der Auswertung jeweils einen Wert zwischen 0% und 100% annehmen.
Da die Tests in den einzelnen Bäumen des ORF zufällig gewählt werden, wird in Hinblick auf die Vergleichbarkeit der Ergebnisse jedes Experiment viermal durchgeführt, die Ergebnisse werden gemittelt. Verglichen werden dementsprechend die Mittelwerte der Ergebnisse, welche in den Tabellen 1 für MOTA und 2 für MOTP dargestellt sind.
In Hinblick auf die Lösung der anfangs erklärten Probleme aktueller Tracking-Verfahren werden auch die Anzahlen der ID-Switches und Fragmentations für die durchgeführten Experimente verglichen. Die Ergebnisse sind in Tabelle 3 für die ID-Switches und Tabelle 4 für die Fragmentations dargestellt.
Für alle Ergebnisse wurde überprüft, ob sie mit einer Irrtumswahrscheinlichkeit von 5% besser als das Referenzergebnis (blau markiert) sind, wobei das Referenzergebnis aus den Experimenten mit den in Klinger and Muhle [2012] vorgestellten Ansätzen stammt. Grün markierte Ergebnisse sind signifikant besser, gelb markierte Ergebnisse sind es nicht. Für die Berechnung der Signifikanz wurden neben den Mittelwerten der Ergebnisse auch die Standardabweichungen berechnet, welche in den Tabellen 5 für MOTA, 6 für MOTP, 7 für ID-Switches und 8 für Fragmentations dargestellt sind.

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 85,56 | 85,29 | 86,09 | 86,80 |
| Additiver Ansatz | 81,87 | 82,24 | 85,33 | 83,98 |
| Teilgewichtung | 81,96 | 82,65 | 86,78 | 85,62 |
| Sigmoid-Gewichtung | 81,59 | 82,15 | 85,74 | 84,56 |

Tabelle 1: Vergleich der semantischen Genauigkeitsmerkmale (MOTA)

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 69,30 | 68,79 | 71,19 | 71,57 |
| Additiver Ansatz | 69,31 | 68,78 | 71,5 | 69,58 |
| Teilgewichtung | 69,94 | 68,76 | 71,92 | 70,86 |
| Sigmoid-Gewichtung | 69,03 | 68,83 | 71,6 | 71,38 |

Tabelle 2: Vergleich der geometrischen Genauigkeitsmerkmale (MOTP)

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 24 | 24,75 | 25,25 | 25 |
| Additiver Ansatz | 24,25 | 23,25 | 25,25 | 26,5 |
| Teilgewichtung | 29,5 | 24,25 | 25,75 | 26 |
| Sigmoid-Gewichtung | 25,25 | 25 | 25,25 | 27,25 |

Tabelle 3: Vergleich der Anzahl der aufgetretenen ID-Switches

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 30,5 | 32 | 29,75 | 26 |
| Additiver Ansatz | 36 | 36 | 33,5 | 32 |
| Teilgewichtung | 37,25 | 36,25 | 30,25 | 29,25 |
| Sigmoid-Gewichtung | 35,25 | 37 | 30 | 32,5 |

Tabelle 4: Vergleich der Anzahl der aufgetretenen Fragmentations

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 0,2 | 0,48 | 0,55 | 0,14 |
| Additiver Ansatz | 0,24 | 0,18 | 1,26 | 0,99 |
| Teilgewichtung | 0,16 | 0,17 | 0,44 | 1,35 |
| Sigmoid-Gewichtung | 1,35 | 0,25 | 1,07 | 1,24 |

Tabelle 5: Standardabweichungen der semantischen Genauigkeitsmerkmale (MOTA)

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 0,19 | 0,68 | 0,74 | 0,59 |
| Additiver Ansatz | 1,21 | 0,58 | 1,65 | 1,8 |
| Teilgewichtung | 0,85 | 0,99 | 1,46 | 1,86 |
| Sigmoid-Gewichtung | 0,53 | 0,17 | 2,47 | 1,38 |

Tabelle 6: Standardabweichungen der geometrischen Genauigkeitsmerkmale (MOTP)

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 0 | 1,71 | 1,71 | 0,82 |
| Additiver Ansatz | 2,5 | 0,96 | 0,5 | 3,32 |
| Teilgewichtung | 3,51 | 0,5 | 1,26 | 0 |
| Sigmoid-Gewichtung | 1,26 | 0,82 | 1,26 | 1,5 |

Tabelle 7: Standardabweichungen der Anzahl der aufgetretenen ID-Switches

| | Ellipse | Querstreifen | Symmetrische Region | Region |
|---|---|---|---|---|
| Multiplikativer Ansatz | 2,89 | 1,41 | 2,63 | 1,63 |
| Additiver Ansatz | 1,41 | 1,83 | 5,26 | 2,94 |
| Teilgewichtung | 2,87 | 2,22 | 2,36 | 3,1 |
| Sigmoid-Gewichtung | 0,5 | 2,45 | 2,83 | 2,08 |

Tabelle 8: Standardabweichungen der Anzahl der aufgetretenen Fragmentations

In den Tabellen 1 und 2 ist zu erkennen, dass die Merkmalskonstellation der symmetrischen Regionen, mit Ausnahme des multiplikativen Ansatzes, generell die besten Ergebnisse liefert. Daher wurde für Konstellation der symmetrischen Regionen mit der Teilgewichtung eine letzte Experimentreihe in Bezug auf die Größe *maxabsc* durchgeführt: Die Idee ist, dass eine Person nach einer Verdeckung nur dann wieder als die selbe Person klassifiziert werden kann, wenn die Wiedererkennung erwartet wird, die Trajektorie der Person also einige Bilder lang nicht verworfen wird. Je länger die Trajektorie erhalten bleibt, desto größer ist das Risiko, dass eine andere Person fälschlicherweise der Trajektorie der Person zugeordnet wird, die Wiedererkannt werden soll. Andererseits muss die Trajektorie lange genug erhalten bleiben, da sonst die Gefahr besteht, dass die Person noch nicht wieder ins Sichtfeld der Kamera getreten ist und somit noch nicht wieder detektiert werden kann. Untersucht wird, ob sich die Veränderung der Größe *maxabsc* auf die Tracking-Genauigkeiten oder die Anzahl der ID-Switches und Fragmentations auswirkt. Die Ergebnisse der Experimente sind in Tabelle 9 für die Größen $MOTA$ und $MOTP$ und in Tabelle 10 für die Anzahl der ID-Switches (#IDS) und Fragmentations (#Frag) dargestellt.

| max absc [Frames] | Mittelwert MOTA | StAbw. MOTA | Mittelwert MOTP | StAbw. MOTP |
|---|---|---|---|---|
| 5 | 86,52 | 0,18 | 71,8 | 1,36 |
| 6 | 86,76 | 0,41 | 72,09 | 1,51 |
| 7 | 86,78 | 0,44 | 71,92 | 1,46 |
| 8 | 86,92 | 0,22 | 73,42 | 0,42 |
| 9 | 87,13 | 0,2 | 73,64 | 0,48 |
| 10 | 86,67 | 1,02 | 72,9 | 1,52 |
| 14 | 86,17 | 0,52 | 71,03 | 2,18 |
| 30 | 83,93 | 0,51 | 70,53 | 0,74 |

Tabelle 9: Vergleich der Genauigkeitsmerkmale in Bezug auf den maximal absence count

| max absc[Frames] | Mittelwert #IDS | StAbw #IDS | Mittelwert #Frag | StAbw #Frag |
|---|---|---|---|---|
| 5 | 28,5 | 0,58 | 33,75 | 1,5 |
| 6 | 27,5 | 0,58 | 32,5 | 2,38 |
| 7 | 25,75 | 1,26 | 30,25 | 2,36 |
| 8 | 26,25 | 1,26 | 29,25 | 0,5 |
| 9 | 26,5 | 1 | 28 | 1,41 |
| 10 | 27 | 1,83 | 29,5 | 1 |
| 14 | 26,5 | 0,58 | 30,25 | 2,22 |
| 30 | 22 | 0 | 30,33 | 0,58 |

Tabelle 10: Vergleich der Anzahl der ID-Switches und Fragmentations in bezug auf den maximal absence count

# 5 Diskussion der Ergebnisse

Wie in den Tabellen 1 und 2 zu sehen ist, ist die semantische und geometrische Genauigkeit sowohl für die Merkmalskonstellation der symmetrischen Regionen zusammen mit dem Teilgewichtungsansatz als auch für die Konstellation der Regionen zusammen mit dem multiplikativen Ansatz signifikant besser als die Genauigkeit, die der in Klinger and Muhle [2012] vorgestellte Ansatz liefert. Trotzdem konnten mit den entwickelten Methoden der Merkmalskonstellation und der Überbrückung von Verdeckungen nicht die aufgeführten Probleme der ID-Switches und Fragmentations gelöst werden. Die meisten Ansätze liefern bezüglich dieser Probleme sogar schlechtere Ergebnisse als der Ansatz von Klinger and Muhle [2012], wie den Tabellen 3 und 4 zu entnehmen ist. Die Ausnahme stellt die Regionskonstellation in Verbindung mit dem multiplikativen Ansatz dar: Hier konnte die Anzahl der aufgetretenen Fragmentations signifikant verringert werden.
Bezüglich der Untersuchung auf die Veränderung von *maxabsc* ist der Tabelle 9 zu entnehmen, dass für $maxabsc = 9$ die besten Ergebnisse sowohl für die $MOTA$ als auch für die $MOTP$ erzielt werden. Ebenso zeigen die Ergebnisse, dass sich durch das Ändern von *maxabsc* die Tracking-Genauigkeiten signifikant ändern; dies wird an der Differenz zwischen den besten und schlechtesten Ergebnissen, also für $maxabsc = 9$ und $maxabsc = 30$ deutlich. Obwohl die Tracking-Genauigkeiten nur durch das Verändern von *maxabsc* nicht signifikant verbessert werden konnten, stellen die Ergebnisse für $maxabsc = 9$ mit den Symmetrieregionen als Merkmalskonstellation und der Teilgewichtung als Ansatz für die Überbrückung von Verdeckungen eine signifikante Verbesserung bezüglich dem in Klinger and Muhle [2012] vorgestellten Ansatz mit $maxabsc = 7$, der Ellipse als Merkmalskonstellation und dem multiplikativen Ansatz für die Überbrückung von Verdeckungen dar. Allerdings konnte auch für die *maxabsc*-Untersuchung die Anzahl der ID-Switches und Fragmentations nicht signifikant verringert werden (vgl. Tabelle 10). Die Ausnahme stellt die Anzahl der ID-Switches für $absc = 30$ dar: Hier konnte die Anzahl signifikant reduziert werden. Dennoch stellt dieser Ansatz keine Verbesserung bezüglich der Tracking-Genauigkeiten dar, wie Tabelle 9 zu entnehmen ist.

# 6 Fazit

In dieser Arbeit wurden verschiedene Merkmalskonstellationen entwickelt, auf ihre inneren Parameter hin untersucht und miteinander verglichen. Ebenso wurden verschiedene Gewichtungsansätze vorgestellt, die, abhängig von den Ergebnissen der Klassifikation und der Prädiktion, einer Detektion eine Trajektorie zuordnen. Das Trackingverfahren wurde mit den entwickelten Merkmalskonstellationen und Gewichtungsansätze hinsichtlich der semantischen und geometrischen Genauigkeiten untersucht und verglichen.
Die unter anderem aus Verdeckungen resultierenden Probleme der ID-Switches und Fragmentations konnte mit den entwickelten Ansätzen nicht gelöst werden, so dass hier weitere Untersuchungen nötig sind; es könnten zum Beispiel die aufgeführten Untersuchungen bezüglich des *maxabsc* auf andere Merkmalskonstellationen und Modelle für die Überbrückung von Verdeckungen bezogen werden. Im Hinblick auf das Lernen des Klassifikators könnte untersucht werden, ob das Reduzieren der Dimensionen der Merkmalsvektoren die Genauigkeit des Klassifikators verbessert; so könnte z.B. der Farbraum auf 2 Dimensionen für Farbe und Sättigung reduziert werden.
Die Experimente wurden anhand einer relativ unkomplexen Bildsequenz durchgeführt: Zu keinem Zeitpunkt sind mehr als 10 Menschen in der Sequenz zu sehen, ebenso bewegt sich die Kamera selbst nicht. Vor allem im Hinblick auf die Anwendungsgebiete des autonomen Fahrens und der automatischen Überwachung ist es sinnvoll, Experimente mit einer sich bewegenden Kamera und/oder deutlich größeren Menschenmengen durchzuführen.
Nichtsdestotrotz stehen am Ende der Untersuchung mit der Merkmalskonstellation der symmetrischen Regionen und dem Gewichtungsansatz der Teilungsgewichtung optimierte Ansätze bezüglich der Trackinggenauigkeiten für ein Tracking-by-Detection-Verfahren. Die erzielten Ergebnisse stellen eine signifikante Verbesserung in Bezug auf das in Klinger and Muhle [2012] vorgestellte Verfahren dar. Es konnten sowohl die semantischen wie auch die geometrischen Genauigkeiten verbessert werden.

# Literatur


G. Bammes. *Studien zur Gestalt des Menschen*. Verlag Otto Maier GmbH, 1990.

K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing, S. 1-10*, 2008.

G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools 25. Jg, Nr. 11, S. 120-126*, 2000.

N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), S. 886-893*, 2005.

M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, S. 2360-2367*, 2010.

T. Klinger and D. Muhle. Persistent object tracking with randomized forests. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, S. 403-407*, 2012.

L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, April 2015. URL `http://arxiv.org/abs/1504.01942`. arXiv: 1504.01942.

A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. *IEEE 12th International Conference on Computer Vision Workshops, S. 1393-1400*, 2009.

G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, S. 1815-1821*, 2012.