Tracking des Aufmerksamkeitsziels des Fahrers mittels eines Multi-Hypothesen Multi-Modell Filters

Julian Schwehr* und Volker Willert†

Zusammenfassung: Für eine sichere Übergabe der Fahraufgabe oder adaptive Warnstrategien sind Informationen über die Wahrnehmung einer Situation durch den Fahrer unerlässlich. In dieser Arbeit wurde ein Multi-Hypothesen Multi-Modell Tracking-Algorithmus entworfen, um das Aufmerksamkeitsziel des Fahrers zu schätzen und zeitlich zu verfolgen. Dabei werden sowohl Objektbewegungen als auch raumzeitliche Modellannahmen des menschlichen Blickverhaltens explizit in den Übergangsschritt eines probabilistischen Filters integriert, was zu einer robusten Schätzung des Blickziels führt. Durch die Verwendung dynamischer und statischer potenzieller Blickziele aus einer Objektliste bzw. einem Freiraum-Spline ist der Algorithmus prinzipiell unabhängig von der verwendeten Sensoranordnung. Die Ergebnisse des Filters werden anhand einer beispielhaften Szene aus realen Testdaten analysiert. Zusätzlich wird der Typ des Blickziels mit Hilfe semantischer Segmentierung im Kamerabild bestimmt und das Verfahren mit einem rein bildbasierten Ansatz ohne Tracking verglichen.

Schlüsselwörter: Bayesian Estimation, Fahrer-Umfeld-Fusion, Multi-Hypothesen Multi-Modell Tracking

1 Einführung

1.1 Motivation

Nach aktuellem Stand scheint der Weg zum hochautomatisierten Fahren vorgezeichnet. Mehrere Hersteller planen, ihre ersten Fahrzeuge nach SAE International Standard J3016 Level 4 mit Beginn des nächsten Jahrzehnts zu vermarkten [1]. Aufgrund der Verschiebung der Verantwortung vom Fahrer hin zu den Herstellern und den Herausforderungen durch abgelenkte Fahrer und der damit verbundenen Frage der sicheren Übergabe haben einige Hersteller angekündigt, Level 3 zu überspringen [2]. Dennoch ist davon auszugehen, dass auch zukünftig Fahrerassistenzsysteme (FAS) der Level 0 bis Level 3 weiter verbessert werden. Sowohl diese zukünftigen FAS, als auch das automatisierte Fahren können von einer Schätzung der Situationswahrnehmung des Fahrers profitieren, um ihre Handlungsstrategie besser an die jeweilige Situation anzupassen. Einer der zukünftigen Aspekte von FAS wird ein umfassendes Situationsverständnis der Assistenzfunktion sein, welches sowohl Fahrer als auch Umwelt miteinbezieht [3]. Da das menschliche Blickverhalten den wichtigsten Indikator für das Situationsbewusstsein des Fahrers während des Autofahrens darstellt, wird dieses bereits seit Jahrzehnten sowohl im Fahrsimulator als auch in Realfahrten untersucht.Seitdem wächst das Forschungsinteresse an der automatisierten Erkennung und Schätzung der Aufmerksamkeit und Situationswahrnehmung des Fahrers. Im Falle automatisierten Fahrens des Level 3 ist die Übergabe der Fahraufgabe vom Auto an den Fahrer entscheidend. Aktivitäts-, Aufmerksamkeits- und Wachsamkeitsanalysen sind notwendig, um die erforderliche Zeit für eine sichere Übergabe zu bestimmen. Aber auch Systeme der Level 0 bis 2 können durch

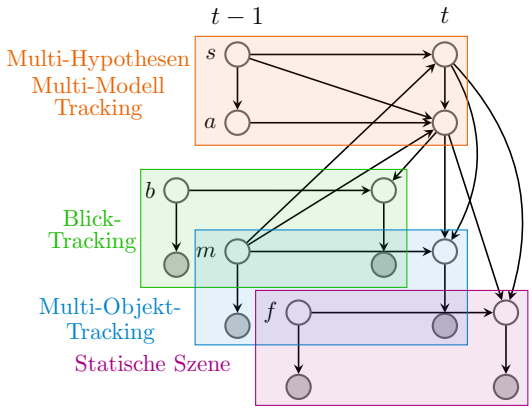
*Julian Schwehr ist Wissenschaftlicher Mitarbeiter am Fachgebiet Regelungsmethoden und Robotik, Technische Universität Darmstadt (e-mail: julian.schwehr@rmr.tu-darmstadt.de).

†Volker Willert ist Akademischer Oberrat am Fachgebiet Regelungsmethoden und Robotik, Technische Universität Darmstadt (e-mail: vwillert@rmr.tu-darmstadt.de).

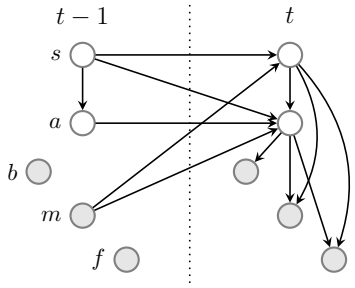
eine Schätzung der Fahreraufmerksamkeit verbessert werden. Diese Schätzung reicht bis zur Unterscheidung von vom Fahrer gesehene und übersehene Verkehrsteilnehmer. Aus technischer Sicht besteht die Aufgabe dann darin, das Blickziel, also das Objekt der gegenwärtigen visuellen Fixation, zu bestimmen und daraus die Wahrnehmung des Fahrers abzuleiten. Auf dieser Information aufbauend, könnte der Fahrer beispielsweise bei manueller Fahrt im Falle tatsächlicher Ablenkung schon bei geringerem Risiko frühzeitiger auf ein einscherendes Fahrzeug hingewiesen werden. Kreuzungs- und Notbremsassistenten könnten ihre Interventionsstrategie der Situation anpassen, z.B. wenn der Fahrer an einer Rechts-vor-Links Kreuzung nicht ausreichend visuell absichert. Bereits eine Adaption von wenigen 100 ms und ein besseres Verständnis des Fahrers kann die Falsch-Positiv-Rate eines Totwinkelassistenten verbessern und damit die Akzeptanz erhöhen. Systeme dieser Art könnten den Fahrer mit Hilfe des zusätzlichen Wissens demnach im Falle einer kritischen Situation entsprechend seiner Bedürfnisse, ähnlich eines aufmerksamen Beifahrers, warnen und eingreifen und damit zu einer intensiveren Kooperation zwischen Fahrer und Fahrzeug führen.

1.2 Stand der Forschung

Trotz der intensiven Bemühungen Fahrerverhalten und Fahrerblickbewegungen zu verstehen, ist es unmöglich, allein mit Hilfe von tragbaren oder berührungslosen kamerabasierten Eye-Tracking-Systemen, automatisiert valide Ground Truth Informationen darüber zu erhalten, *was* der Fahrer tatsächlich *gesehen* und *wahrgenommen* hat. Dieses Fehlen einer Ground Truth über das *Bewusstsein* des Fahrers über andere Verkehrsteilnehmer ist jedoch wahrscheinlich der Grund, warum die meisten Arbeiten in der Literatur, die sich mit der Blickzielschätzung befassen, auf einfachen Distanzkriterien beruhen [4–10]. Da das Blickziel nicht ohne Weiteres bestimmt werden kann, gilt das Ziel als gesehen, wenn der Blick in eine bestimmte Region fällt. Hierbei werden die Distanzen in unterschiedlichen Koordinatensystemen definiert [11], meist basierend auf der Wahl der Umfeldsensorik. Die meisten Arbeiten verwenden Kameras [4–6, 8, 10, 12], seltener werden Laser Scanner [7, 9] und Radar [11] genutzt. Allein in [7] wird die Modellierung um eine minimale Fixationszeit erweitert und in [8] wird eingeräumt, dass die Blickbewegung des Fahrers nicht allein durch das Auftreten von äußeren Objekten in bestimmten Regionen zuverlässig erklärt werden kann. Nichtsdestoweniger haben die vielen Simulator- und realen Fahrstudien einige allgemein beobachtbare Eigenschaften des Blickverhaltens aufgedeckt [13]. Darunter sind formulierbare Regeln, wie beispielsweise dass Fixationen auf aufgabenrelevante Objekte und Positionen fallen, einzelne Fixationen interpretierbare funktionale Rollen haben und Fixationen unregelmäßig unterbrochen sind. Beim Autofahren werden Fixationen auf relevante Ziele immer wieder durch Blicke nach vorne unterbrochen. Was in aktuellen Arbeiten bisher fehlt, ist die tatsächliche Verwendung dieser bekannten, wenn auch vagen, Verhaltensregeln als Modellwissen für eine gemeinsame Beschreibung des Fahrerverhaltens und der Umgebung zur Schätzung des Blickziels. Hierfür bieten sich Bayessche Filter an, die in zahlreichen Anwendungen gezeigt haben, die Zustandsschätzung, gegeben rauschbehafteter Messungen, mit Modellwissen verbessern zu können. In dieser Arbeit bauen wir auf unseren Ideen aus [11] auf mit dem übergeordneten Ziel, jeweils das Objekt zu extrahieren, dem der aktuelle Blick des Fahrers gilt. Zu diesem Zweck wird in dieser Arbeit ein Multi-Hypothesen Multi-Modell (MHMM) Tracking vorgestellt, welches für jedes Objekt explizit die Wahrscheinlichkeit schätzt, das aktuelle Blickziel zu sein. Zusammenhängende Phasen, die das identische Objekt als Blickziel ausgeben, können dann als Fixation detektiert werden und bei ausreichender Dauer als Indikator dienen, dass sich der Fahrer des entsprechenden Objekts bewusst ist. Vom Standpunkt der Motivation gesehen, ist [14] am ähnlichsten zu dieser Arbeit. Basierend auf ähnlicher Argumentation werden hier die relative Bewegungen der Objekte und des Fixationspunktes als Features genutzt, um das aktuelle Blickziel lernbasiert über ein Klassifikationsproblem zu bestimmen. Die



(a) Kombination des vorgeschlagenen Multi-Hypothesen Multi-Modell Filters zusammen mit probabilistischen Modellen für Blick, mehreren Objekten und statischer Szene.



(b) Vereinfachte Version des vorgeschlagenen Multi-Hypothesen Multi-Modell Filters mit Blick, Objekten und statischer Szene als Messungen.

Abbildung 1: Multi-Hypothesen Multi-Modell Filter in zwei verschieden Varianten mit Umschaltvariable s und Variable a für den Aufmerksamkeitsbereich. b, m, f beschreiben Blick, Objekte und statische Szene. Pfeile zu den Messungen zum Zeitpunkt $t-1$ wurden für bessere Übersicht nicht gezeichnet.

Blickrichtung wird hier jedoch über einen tragbaren Eye-Tracker erhalten.

2 Multi-Hypothesen Multi-Modell Tracking

In dieser Arbeit wird vorausgesetzt, dass die Wahrnehmungen der Verkehrsszene durch den Fahrer und das Fahrzeug konsistent sind. Es wird also einerseits angenommen, dass das Fahrzeug das Fixationsobjekt erfasst und die relative Position des Objektes ausreichend genau bestimmt hat. Andererseits wird angenommen, dass die gemessene Blickbewegung des Fahrers zur relativen Bewegung des jeweils aktuellen Fixationsobjektes passt. Mit dieser Prämisse ist das übergeordnete Ziel die Schätzung der a-posteriori Verbundwahrscheinlichkeit des probabilistischen Netzwerks in Abbildung 1a, welches die Wechselwirkung zwischen Fahrerblick und potenziellen Blickzielen beschreibt. Die Aufgabe des entworfenen Multi-Hypothesen Multi-Modell (MHMM) Filters auf der oberen Filterebene ist, den Aufmerksamkeitsbereich a und das Blickziel s mit geeigneten Bewegungsmodellen sowie den Blick- (b) und Objektschätzungen (m, f) zu schätzen und zu verfolgen. Im Falle exakter und vollständiger Inferenz werden die stochastischen Variablen auf der unteren Ebene des Filters (Blickrichtung, Objektposition und -geschwindigkeit) durch die Schätzung von s und a beeinflusst. Somit ist bei einer gemeinsamen Betrachtung ein Informationsgewinn in beiden Domänen, also Fahrerbeobachtung und Umgebungsschätzung, möglich: Korrekturen der Blickrichtungsmessung sind ebenso möglich wie eine Verbesserung der Positionsschätzung des fixierten Objekts. Statt der umfassenden probabilistischen Beschreibung in Abbildung 1a verwendet diese Arbeit eine vereinfachte Version des vollständigen Filters, welche in Abbildung 1b dargestellt ist und sich allein auf den MHMM-Filter konzentriert. Dieser Filter ist eine Kombination aus einem Multi-Hypothesen (MH) und einem Multi-Modell (MM) Tracking. Ein MH-Tracker stellt eine Verteilung über alle möglichen Objekt-Hypothesen dar und bestimmt die jeweiligen Gewichte jeder Hypothese basierend darauf, wie gut die Hypothesenvorhersagen zur Sensormessung passen. Im Unterschied dazu verfolgt ein MM-Filter verschiedene (Bewegungs-)Modelle für jeweils eine Hypothese und bestimmt das Übergangsmodell ebenfalls basierend darauf, wie gut die verschiedenen Modellvorhersagen zur Sensormessung passen [15].

Das Ziel dieser Arbeit ist die Schätzung der Verbundwahrscheinlichkeit $p(\{s,a\}^t|\{b,m,f\}^{1:t})$[1]. Die Schaltvariable s ist eine $(n+1)$-dimensionale binäre Zufallsvariable $s \in \mathbb{R}^{n+1}$ mit $s_i \in \{0,1\}$, $\sum_i s_i = 1$, wobei n die Anzahl der Objekte in der Umgebung beschreibt und die zusätzliche Dimension die statische Szene enthält. Jedes potenzielle Blickziel i besitzt eine gewisse Wahrscheinlichkeit $p(s_i = 1) = \pi_i$, dass es das aktuelle Blickziel des Fahrers ist und eine gewisse Ausdehnung $p(a|s_i = 1)$ modelliert durch eine Normalverteilung $\mathcal{N}(a|\mu_i, \Sigma_i), a, \mu_i \in \mathbb{R}^2, \Sigma_i \in \mathbb{R}^{2\times 2}$. Der Aufmerksamkeitsbereich des Fahrers wird durch die gewichtete Kombination der einzelnen Normalverteilungen als Gaußsches Mischmodell (engl. mixture of Gaussians)

$$p(\{s,a\}^t|\{b,m,f\}^{1:t}) = \prod_{i=1}^{n+1} \pi_i^{s_i} \mathcal{N}(a^t|\mu_i^t, \Sigma_i^t)^{s_i} \tag{1}$$

dargestellt. Die Messung $b \in \mathbb{R}$ repräsentiert den Gierwinkel der Blickrichtung gemessen in Fahrzeugkoordinaten durch das Eye-Tracking-System. $m \in \mathbb{R}^{n\times 5}$ ist eine geordnete Objektliste, welche Position (x,y), Ausrichtung (φ) und relative Geschwindigkeit $(v_{rel,x}, v_{rel,y})$ der Objekte enthält, basierend auf den Messungen der Front- und Seitenradare. Die Messung der statischen Umgebung $f = \{f_i\}$ ist die Menge der Kontrollpunkte der B-Spline-Kurve, die den Freiraum beschreibt [16].

2.1 Übergangswahrscheinlichkeiten (Prädiktion)

Bei der Modellierung der Übergangswahrscheinlichkeiten muss jede Möglichkeit der Blickbewegung berücksichtigt werden. Dabei können die folgenden Fälle auftreten: Der Blick kann auf einem spezifischen Objekt $i \in 1,...,n$ haften bleiben (Fixations- und Folgemodell) oder von einem Objekt i zu einem anderen Objekt $j \in 1,...,n$ oder auch der statischen Umgebung $(j = n+1)$ springen (Sakkadenmodell). Umgekehrt könnte auch die statische Umgebung das aktuelle Blickziel sein $(i = n+1)$ und der Fahrer könnte zum nächsten Zeitschritt mit seinem Blick an der gleichen Stelle bleiben, zu einer anderen Stelle in der statischen Umgebung springen (für beide Fälle gilt $j = n+1$) oder ein Objekt fixieren $(j \in 1,...,n)$. Aufgrund der entworfenen Filterstruktur in Abbildung 1b und der Formulierung der a-priori Verteilung (1) als Gaußsches Mischmodell kann der Übergang aufgeteilt werden in ein skalares Übergangsmodell der Schaltvariable und in ein räumliches Übergangsmodell

$$p(\{s,a\}^t|\{s,a,m\}^{t-1}) = p(s^t|\{s,m\}^{t-1}) p(a^t|s^t, \{s,a,m\}^{t-1}). \tag{2}$$

2.1.1 Übergang des Schaltzustands

Normalerweise sind die Übergangsgewichte konstante Entwurfsparameter. Hier wird der Übergang

$$p(s^t|\{s,m\}^{t-1}) = \prod_{j=1}^{n+1} \prod_{i=1}^{n+1} \left(\alpha(s^t, s^{t-1}) \beta(s^t, \{s,m\}^{t-1})\right)^{s_i s_j} \tag{3}$$

jedoch durch zwei Potenziale dargestellt, welche einen zeitlichen Aspekt $\alpha(s^t, s^{t-1})$ und einen räumlichen Aspekt $\beta(s^t, \{s,m\}^{t-1})$ modellieren. Dabei berücksichtigt das zeitliche Übergangsmodell α eine Veränderung der Übergangswahrscheinlichkeiten für Fixationen und Sakkaden über der Zeit. Je länger die detektierte Fixationsdauer ist, desto mehr steigt die Wahrscheinlichkeit für eine Sakkade, also einen Sprung hin zu einem anderen Ziel. Das räumliche Übergangsmodell β betrachtet dagegen die geometrischen Beziehungen der Objekte in der Szene und bezieht psychophysikalische Randbedingungen wie maximale Sakkadengeschwindigkeit und mittlere Blickrichtung mit ein.

[1] Wir nutzen die Notation $(\{\cdot,\cdot\}^{1:t})$ als verkürzte Schreibweise um mehrere Variablen mit dem gleichen Zeitindex zu beschreiben, normalerweise dargestellt durch $(\cdot^{1:t}, \cdot^{1:t})$.

2.1.2 Räumlicher Übergang

Wenn der Blick des Fahrers auf einem Objekt verharrt, die Augenbewegung also eine Fixation oder Folgebewegung (engl. smooth pursuit) beschreibt, sollte die entsprechende Wahrscheinlichkeitsverteilung die gleiche Bewegungsdynamik wie das fixierte Objekt aufweisen. Gleichzeitig sollte, für den Fall dass der Fahrerblick zu einem anderen Objekt j wechselt, die prädizierte Verteilung auch an der Stelle des Objekts j sein. Der räumliche Übergang wird daher als Normalverteilung

$$p(a^t|s^t, \{s,a,m\}^{t-1}) = \prod_{j=1}^{n+1} \prod_{i=1}^{n+1} \mathcal{N}(a^t|a^{t-1} + u_{ij}^{t-1}, \Gamma_{ij})^{s_i s_j} \tag{4}$$

modelliert, wobei u_{ij} einen Steuereingang darstellt und das Prozessrauschen durch die Kovarianzmatrix Γ_{ij} gegeben ist. Das Teilergebnis des späteren Inferenzschritts für den Übergang von Objekt i zu Objekt j ist dann die Normalverteilung $\mathcal{N}(a^t|\mu_i^{t-1} + u_{ij}^{t-1}, \Sigma_i^{t-1} + \Gamma_{ij})^{s_i s_j}$. Da für $i = j$ der Blick eine Bewegungsdynamik konsistent zur Messung der Objektdynamik aufweisen soll, wird der Steuereingang durch die relative Bewegung des Blickziels modelliert, also $u_{ij}^{t-1} = v_i^{t-1}\Delta t$, wobei Δt die Abtastzeit des Blicks beschreibt. Analog dazu wird der Steuereingang im Fall $i \neq j$ zu $u_{ij}^{t-1} = x_j^{t-1} - \mu_i^{t-1} + v_j^{t-1}\Delta t$ gewählt, sodass die prädizierte Position der Aufmerksamkeit auf der prädizierten Position des Objekts liegt. Für die statische Umgebung ($j = n+1$) besteht die relative Bewegung allein aus der Kompensation der Eigenbewegung und an Stelle der Objektposition x_j^{t-1} wird die letzte Schätzung μ_{n+1}^{t-1} genutzt. Zusätzlich wird das Prozessrauschen deutlich größer gewählt, da der Fahrerblick von der aktuellen Position auf dem Freiraum-Spline zu fast jeder anderen Position auf dem Spline springen kann.

2.2 Messmodelle

Es wird angenommen, dass die Messungen von Blick, dynamischen Objekten und der statischen Umgebung unabhängig voneinander sind, sodass die Likelihood des Filterschritts in die Terme für die einzelnen Messungen faktorisiert. Für die Blickmessung wird ein Modell verwendet, welches Positionen, die eine geringere Winkeldistanz zur Blickmessung aufweisen, eine höhere Likelihood zuordnet und eine geringe Likelihood an Positionen fern des Blickstrahls.

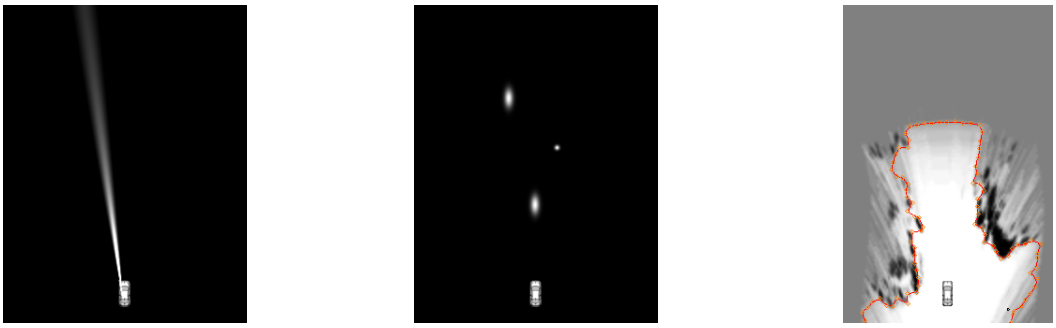
Bei den Objekten wird jedes einzelne Objekt als gewichtete Normalverteilung repräsentiert, deren Mittelwert der Position des jeweiligen Objekts entspricht. Die Kovarianzmatrix wird so gewählt, dass die Hauptachsen der 90% Kovarianzellipse der Breite und Länge des Objekts entsprechen. Die Matrix ist außerdem entsprechend der Ausrichtung des Objekts gedreht.

Die Messung der statischen Umgebung wird durch die Kontrollpunkte einer Freiraum-B-Spline-Kurve repräsentiert [16]. Da diese Spline-Kurve weder in geschlossener Form und schon gar nicht als Normalverteilung formuliert werden kann, dienen Abtastungen der Kurve als diskrete Realisierungen des Likelihood. Der Status der Kontrollpunkte gibt zudem an, ob es sich an der jeweiligen Position um ein statisches Hindernis oder einen Übergang zu unbekannter Umgebung handelt. Exemplarische Bilder der verschiedenen Messmodelle sind in Abbildung 2 dargestellt.

2.3 Inferenz

Im Inferenzschritt des entworfenen MHMM Tracking-Algorithmus treten zwei Schwierigkeiten auf:

1. Sowohl Blick- als auch Freiraum-Likelihood sind keine Normalverteilungen, weshalb die Inferenz nicht in geschlossener Form dargestellt werden kann.



(a) Blickmodell (hellere Bereiche beschreiben höhere Likelihood) (b) Likelihood für drei dynamische Objekte verschiedener Größe (c) Freiraum-Spline hinterlegt mit der Belegungskarte

Abbildung 2: Exemplarische Darstellung der Messmodelle.

2. Aufgrund der exponentiell steigenden Anzahl an Gaußschen Mischkomponenten (schon der erste Prädiktionsschritt führt auf $n+1$ Summen von $n+1$ Normalverteilungen) wird der Filter ohne geeignete Approximationen praktisch schnell unberechenbar.

Um dem ersten Problem zu begegnen wird jede Übergangswahrscheinlichkeit von Objekt i zu Objekt j durch Minimierung der Kullback-Leibler-Divergenz gegeben einer Menge an gewichteten Abtastungen x_k mit einer Normalverteilung approximiert. Um den Rechenaufwand des Filters beherrschbar zu halten, wird das zweite Problem angegangen, indem alle Mischkomponenten eines Blickziels mit einer einzelnen Normalverteilung approximiert werden. Dies entspricht dem Vorgehen des *generalisierten pseudo-Bayesschen Schätzers zweiter Ordnung* (engl. generalized pseudo-Bayesian estimator of second order (GPB2)) [15].

3 Bestimmung des Zieltyps mittels semantischer Segmentierung im Bild

Da bei der Modellierung in 2D-Fahrzeugkoordinaten nur der Blick-Gierwinkel verwendet wird, geht bei der zuvor vorgestellten Methode der Blick-Nickwinkel, und damit die Höhe des Fixationspunktes (engl. Point of Regard (PoR)) über dem Boden, verloren. Somit werden Blickmessungen potenziell mit Objekten verknüpft, ohne dass eine tatsächliche Übereinstimmung besteht. Um solche Falschzuordnungen zu detektieren, ist eine Möglichkeit, die Höhe des Blickstrahls am PoR zu überprüfen. Diese Vorgehensweise weist allerdings den Nachteil auf, dass die Höhe eines Objekts ebenso unbekannt sein kann. Stattdessen versuchen wir, den PoR in Fahrzeugkoordinaten ins Bild der Fahrzeugkamera rückzuprojizieren. Der PoR x^* in Fahrzeugkoordinaten ergibt sich als Mittelwert derjenigen Normalverteilung mit dem größten Gewicht

$$x^* = \mu_{i^*} \quad \text{mit} \quad i^* = \text{argmax}_i(\pi_i). \tag{5}$$

Die Höhe des Punkts wird über den gemessenen Blick-Nickwinkel und die Position von x^* im Fahrzeugsystem bestimmt. Anschließend wird dieser 3D-Punkt in das Bild rückprojiziert. Da die Pixelposition alleine jedoch keine Information liefert, wird der PoR im Bildraum mit einer semantischen Segmentierung des Bildes kombiniert. Hierzu verwenden wir das auf dem Cityscapes Datensatz [17] vortrainierte Faltungsnetz aus [18]. Die Art des Blickziels kann nun beispielsweise anhand des Pixellabels an der Stelle der Rückprojektion bestimmt (siehe Abbildung 3) und mit der Objektkategorie aus der Objektliste abgeglichen werden.

In [19] wird argumentiert, dass bei kleiner werdendem Aufmerksamkeitsbereich um den PoR herum, Kategorien mit steigendem proportionalem Anteil das tatsächliche Fixationsziel sind, während Kategorien mit sinkendem Anteil tendenziell nur in den Bereich um den PoR fallen,

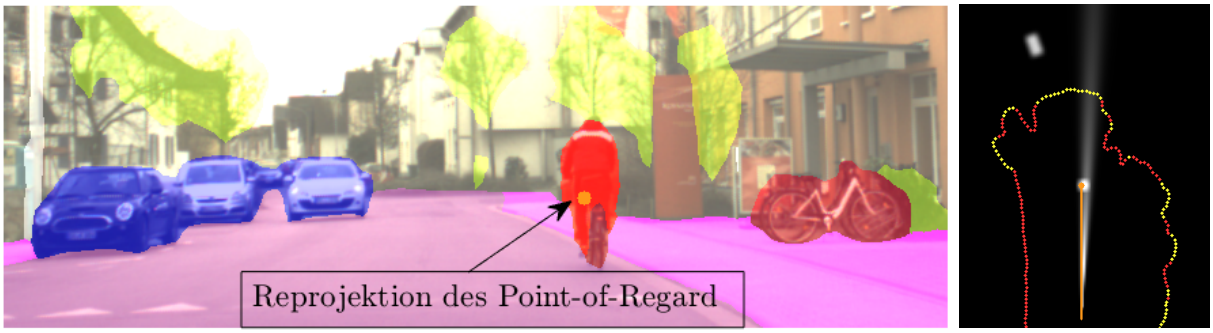


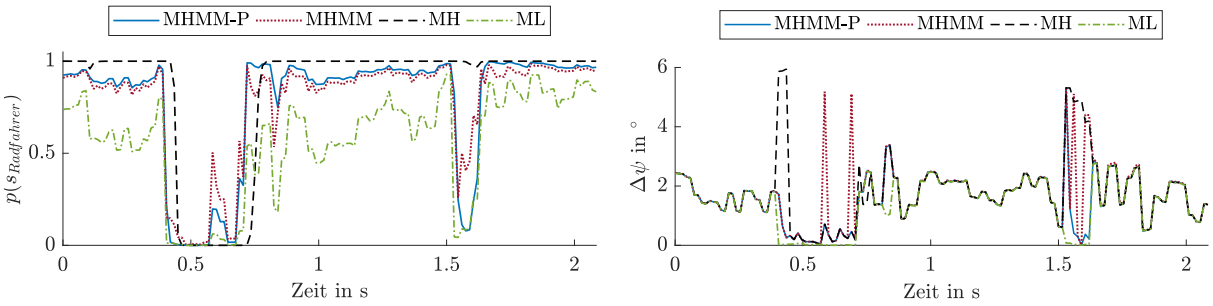
Abbildung 3: Bild der Fahrzeugkamera, überlagert mit dem Ergebnis der semantischen Segmentierung und der Rückprojektion des PoR aus dem MHMM Filter. Rechts die Visualisierung des PoR in Fahrzeugkoordinaten zusammen mit Freiraum-Spline, Objekten und Blickmessung.

da sie sich in der Nähe des eigentlichen Blickziels befinden. Diese Argumentation nutzen wir für eine weitere Möglichkeit zur Bestimmung des Zieltyps. Um den PoR werden Kreise mit den Radien 110 px und 20 px gezogen (entspricht bei der verwendeten Kamera in etwa 10 ° bis 2 ° Blick-Öffnungswinkel) und die proportionalen Anteile jeder Label-Kategorie $p_{k20}, p_{k110},\ k \in m$ bestimmt, wobei m die Anzahl an Kategorien ist. Der Zieltyp K ist dann diejenige Kategorie mit dem stärksten prozentualen Anstieg, d.h. $K^* = \text{label}(\text{argmax}_k(p_{k20}/p_{k110}))$ gegeben, wobei $\text{label}(k)$ das semantische Label der Kategorie k beschreibt.

4 Ergebnisse

4.1 Diskussion der Tracking-Ergebnisse anhand eines ausgewählten Szenarios

Wie eingangs erwähnt, ist es bisher nicht möglich, automatisiert valide Ground Truth Informationen darüber zu erhalten, welche Objekte der Fahrer tatsächlich visuell fixiert. In [14] werden daher händische Annotationen von Fixationen und Fixationszielen in den Daten durchgeführt. In dieser Arbeit soll anhand eines ausgewählten Abschnitts aus realen Messdaten, in dem der Fahrerblick klar zwischen zwei Blickzielen wechselt, das MHMM-Tracking argumentativ mit den degenerierten Versionen eines MH-Modells und eines Maximum-Likelihood-Modells (ML) verglichen werden. Im MH-Modell werden die Übergangsparameter von einem Objekt zu einem anderen zu null gesetzt, was in einem separaten Filter für jede Hypothese resultiert. Das ML-Modell besteht dagegen lediglich aus der Multiplikation der einzelnen Messmodelle. Außerdem wird der zusätzliche Nutzen einer komplexen räumlichen und zeitlichen Modellierung der Übergangswahrscheinlichkeiten hervorgehoben werden (das entwickelte Modelle wird hier abgekürzt mit MHMM-P). Das ausgewählte Szenario dauert 2 s und ist in Abbildung 3 dargestellt: in einem Wohngebiet fährt vor dem Fahrer ein Fahrradfahrer während ein anderes Fahrzeug entgegenkommt. In Abbildung 4a sind die Verläufe der Wahrscheinlichkeit $p(s_i)$ des Fahrradfahrers aus jedem der vier Modelle dargestellt und in Abbildung 4b die Abweichung des gemessenen Blick-Gierwinkels zur Richtung des jeweils wahrscheinlichsten Blickziels. Im direkten Vergleich sind folgende Eigenschaften der Modelle zu beobachten: Die ML-Schätzung ist im Vergleich zu den Filteransätzen anfällig für verrauschte Messungen, was eine inhärente Eigenschaft und die grundlegende Motivation für Bayessche Filter ist. Der Schätzungsverlauf des MH-Modells weist dagegen deutlichen Tiefpasscharakter auf, da Sprünge zwischen Blickzielen im direkten Vergleich verzögert auftreten oder sogar unterdrückt werden. Aufgrund der nicht vorhandenen Abwägung zwischen Fixation und Sakkade konvergiert die Wahrscheinlichkeit des „besten“ Modells zu Eins und unterdrückt die Wahrscheinlichkeit für mögliche Alternativen.



(a) Wahrscheinlichkeit $p(s_i)$, dass der Radfahrer vom Fahrer während der ausgewählten Sequenz angeschaut wird.

(b) Abweichung des gemessenen Blick- Gierwinkels ψ zur Richtung des jeweils wahrscheinlichsten Blickziels.

Abbildung 4: Tracking Ergebnis für verschiedene Abwandlungen des vorgestellten Modells.

Aus diesem Grund ist ein geeigneter Regularisierungsterm notwendig [15]. Der vorgeschlagene MHMM-Tracking-Algorithmus überwindet beide der genannten Probleme: Indem er die Möglichkeit verfolgt bei dem aktuellen Objekt zu bleiben, ist die Schätzung glatter als das Messmodell alleine und da er gleichzeitig die Möglichkeit für Sprünge zu anderen Zielen ermöglicht, ist das Modell in der Lage, der flexiblen und schnellen Dynamik des menschlichen Blicks zu folgen. Zu guter Letzt wird die Schätzqualität durch eine zeit- und szenenabhängige Modellierung der Übergänge mittels der Faktoren α und β verbessert (MHMM-P-Modell). Gerade direkt nach einem Sprung in der Wahrscheinlichkeit ist ersichtlich, dass das Modell den Kontrast zwischen den Wahrscheinlichkeiten für verschiedene Blickziele nochmal erhöhen kann.

Betrachtet man die Gierwinkelabweichungen, so ist erkennbar, dass alle Modelle für den Großteil der Sequenz in der Schätzung des Blickziels übereinstimmen (Fixation des Radfahrers), da die Abweichungen praktisch identisch sind. In diesen Abschnitten betragen die Abweichungen in etwa zwischen $0,5\,°$ und $2,5\,°$, was den Annahmen über Messgenauigkeiten und Öffnungswinkel des scharfen Sehens entspricht. Das bedeutet jedoch auch, dass das Messmodell (ML) bereits eine starke lokale Vorgabe über den Ort des Blickziels macht und die Tracking-Modelle vor allem die Gewichte π_i beeinflussen. Während der Phasen, in denen nicht der Radfahrer Ziel des Blicks ist, sondern der Freiraum-Spline, entspricht das ML-Modell praktisch der Messung selbst und beschreibt den Schnitt des Blickstrahls mit dem Freiraum-Spline. Auch bei den anderen Modellen sind die Abweichungen zwischen $t = 0,4\,\mathrm{s}$ und $t = 0,7\,\mathrm{s}$ tendenziell geringer, da im Falle einer Blickfolge auf den Freiraum-Spline kaum Annahmen über die Bewegung des Blicks gemacht werden. Beim MH-Modell ist erneut der Tiefpasscharakter erkennbar, da die Winkeldifferenz nach einer Sakkade, einem Wechsel von einem Blickziel zum anderen, erst verzögert korrigiert wird. Schließlich sind beim MHMM-Modell ebenfalls einzelne Fehlschätzungen erkennbar.

4.2 Verlauf des Blickziels im Bildbereich

Bei der Bestimmung des Zieltyps mittels semantischer Segmentierung im Bild vergleichen wir die Reprojektion des MHMM-P-Filters mit der rein bildbasierten Methode zur Bestimmung des PoR aus [12]: Im Tiefenbild (gemessen von einer Stereokamera) wird die Tiefe des rückprojizierten Blickstrahls mit der gemessenen Tiefe des entsprechenden Pixels im Bild verglichen. Die Stelle der geringsten Tiefendifferenz entspricht dem PoR. In Abbildung 5 ist das semantische Label des Blickziels über der Zeit dargestellt. Hierbei werden einerseits die beiden Methoden zur Bestimmung des PoR (MHMM-P und Stereo) und die beiden Methoden zur Bestimmung des Zieltyps (Pixel-Label an der Stelle des PoR und die Methode aus Abschnitt 3 in Anlehnung an [19]) verglichen. Zunächst ist ersichtlich, dass der rein bildbasiert bestimmte Zieltyp für beide

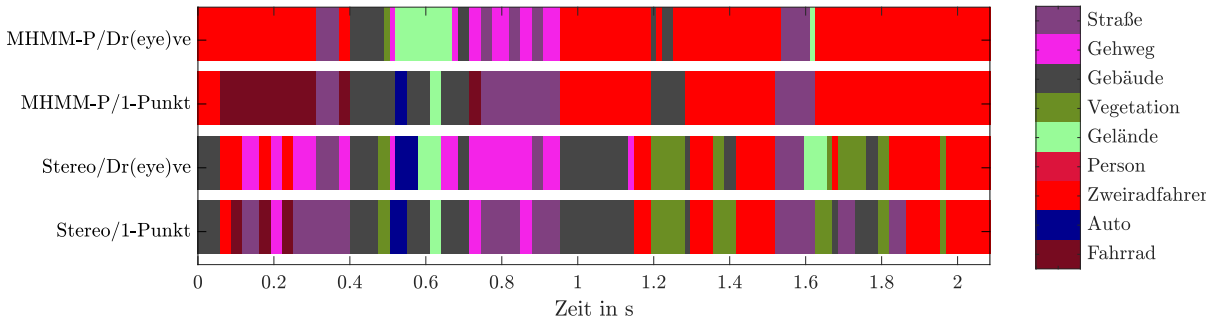


Abbildung 5: Label des Blickziels über der Zeit. MHMM-P: Der PoR stammt aus der Rückprojektion der Filterschätzung in Fahrzeugkoordinaten. Stereo: Der PoR stammt aus der Reprojektion des Blickstrahls in die Tiefenkarte; Dr(eye)ve: Zieltyp über den prozentualen Anstieg des Flächenanteils um den PoR in Anlehnung an [19]. 1-Punkt: Zieltyp entspricht dem Label des PoR-Pixels.

Abbildung 6: PoR im Vergleich, links bei $t = 0,72\,\mathrm{s}$ bis $0,76\,\mathrm{s}$, rechts bei $t = 1,2\,\mathrm{s}$. Orange: Rückprojektion des MHMM-P-Modells. Weiß: Schätzung über Tiefenbild. Zu diesen Zeiten schätzt das Tracking-Modell, dass der Fahrer den Radfahrer fixiert. Im Bildbereich wird unter Berücksichtigung des Blick-Nickwinkels ersichtlich, dass diese Aussage unter Umständen nicht korrekt ist und gegebenenfalls korrigiert werden muss.

Bestimmungsmethoden recht wechselhaft ist. Hieraus lässt sich keine Fixation auf ein Objekt ableiten, da die zusammenhängenden Abschnitte kaum mehr als 100 ms betragen. Dies ist nicht verwunderlich, da die Bestimmung des PoR auf den rauschbehafteten Blickmessungen basiert und keine Modellannahmen getroffen wurden. Im Gegensatz dazu kann aus der Rückprojektion des Filterergebnisses zuverlässig und zusammenhängend der Fahrradfahrer oder das dazugehörige Fahrrad als Blickziel extrahiert werden. Dies liegt zum einen daran, dass der Blick-Gierwinkel über das Tracking korrigiert wird und der PoR im Fahrzeugkoordinatensystem auf den Fahrradfahrer „gezogen“ wird, zum anderen passt jedoch auch der Blick-Nickwinkel zum Fahrradfahrer. Bei $t = 0,72\,\mathrm{s}$ bis $0,95\,\mathrm{s}$ widersprechen sich die Schätzungen des Trackings in Fahrzeug- und Bildkoordinaten jedoch. Wie Abbildung 6 zeigt, ist hier fragwürdig, ob der Blick dem Radfahrer zugeschrieben werden kann oder nicht. Zum Zeitpunkt $t = 1,2\,\mathrm{s}$ stimmen die Schätzungen zwar auch nicht überein, im Bildbereich wird jedoch ersichtlich, dass es sich womöglich nur um einen Fehler der Blick-Nickwinkelmessung handelt.

5 Fazit

In dieser Arbeit wurde ein MHMM-Tracking-Algorithmus zur Schätzung des Aufmerksamkeitsziels des Fahrers vorgestellt. Dieser beinhaltet im Vorhersageschritt sowohl Annahmen über Objektdynamik als auch über Blickverhalten. Die Fixationswahrscheinlichkeit jedes potentiellen Ziels resultiert direkt aus der probabilistischen Beschreibung, wodurch, im Gegensatz zu bestehenden Methoden, auch eine Abwägung zwischen mehreren Hypothesen erfolgt. Erste Evaluierungsergebnisse zeigen, dass die Methode grundsätzlich in der Lage ist, das Blickziel des Fahrers im zweidimensionalen Fahrzeugkoordinatensystem zu bestimmen. Über die Rückprojektion in die Fahrzeugkamera lässt sich das Ergebnis des Filters mit dem semantischen Label an der rückprojizierten Stelle verknüpfen und gegebenenfalls korrigieren.

In zukünftigen Arbeiten soll daher verstärkt auf die Kombination von Fahrzeug- und Bildkoordinaten eingegangen werden. Hierbei gibt es mehrere Ansatzmöglichkeiten: Die Schätzung des PoR in Bildkoordinaten kann über ein eigenes Trackingmodell unter Verwendung von robuster Objektdetektion und Objekttrackings sowie dichten Stereobildern verbessert werden. Robuste Objektdetektionen im Kamerabild könnten nicht nur pragmatischer sein als die detaillierte Darstellung in semantisch segmentierten Bildern, sondern versprechen zudem die Möglichkeit zur Verbesserung der radarbasierten Objektliste. Die Fusion des Aufmerksamkeitsbereichs basierend auf unterschiedlichen Modellen und die Bestimmung des Zieltyps im Bild ist über die Verwendung der Metrik *Intersection over Union (IoU)* denkbar.

Letztlich ist die größte Schwierigkeit jedoch eine ausführliche, sinnvolle und valide Auswertung und der Vergleich verschiedener Modelle. Da Methoden basierend auf berührungslosen kamerabasierten Eye-Tracking-Systemen zwar wie in dieser Arbeit argumentativ untereinander verglichen werden können, der Bezug zu einer vernünftigen Ground Truth jedoch fehlt, ist die Aussagekraft verbesserungswürdig. Die gleichzeitige Verwendung von berührungslosen und tragbaren Eye-Tracking-Systemen inklusive anschließendem manuellen Labeling ähnlich wie in [14,19] ist zur umfassenden Bewertung der vorgestellten Methoden vermutlich unvermeidbar.

Literatur

[1] J. Walker, "The Self-Driving Car Timeline – Predictions from the Top 11 Global Automakers," 2017. url: techemergence.com/self-driving-car-timeline-themselves-top-11-automakers; Stand 16. Juli 2018.

[2] B. Gain, "Why every car maker should skip level 3," 2017. url: driverless.wonderhowto.com/news/waymo-was-right-why-every-car-maker-should-skip-level-3-0178497; Stand 16. Juli 2018.

[3] R. Klette, "Vision-based Driver Assistance Systems," 2015. url: researchgate.net/publication/272199860_Vision-based_Driver_Assistance_Systems; Stand 16. Juli 2018.

[4] T. Langner, D. Seifert, B. Fischer, D. Goehring, T. Ganjineh, and R. Rojas, "Traffic awareness driver assistance based on stereovision, eye-tracking, and head-up display," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2016.

[5] S. J. Zabihi, S. M. Zabihi, S. S. Beauchemin, and M. A. Bauer, "Detection and recognition of traffic signs inside the attentional visual field of drivers," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017.

[6] L. Petersson, L. Fletcher, and A. Zelinsky, "A framework for driver-in-the-loop driver assistance systems," in *Proceedings. IEEE Intelligent Transportation Systems*, pp. 771–776, 2005.

[7] T. Bär, D. Linke, D. Nienhüser, and J. M. Zöllner, "Seen and missed traffic objects: A traffic object-specific awareness estimation," in *IEEE Intelligent Vehicles Symposium (IV)*, 2013.

[8] A. Doshi and M. Trivedi, "Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions," in *IEEE Intelligent Vehicles Symposium, (IV)*, 2009.

[9] M. M. Moniri, D. Merkel, M. Feld, and C. Müller, "Incorporating the Driver's Focus of Attention into Automotive Applications in Real Traffic and in Simulator Setups," in *12th Int. Conf. on Intelligent Environments (IE)*, 2016.

[10] S. Guasconi, M. Porta, C. Resta, and C. Rottenbacher, "A low-cost implementation of an eye tracking system for driver's gaze analysis," in *10th Int. Conf. on Human System Interactions (HSI)*, 2017.

[11] J. Schwehr and V. Willert, "Driver's Gaze Prediction in Dynamic Automotive Scenes," in *20th Int. IEEE Conf. on Intelligent Transportation Systems (ITSC)*, 2017.

[12] T. Kowsari, S. S. Beauchemin, M. A. Bauer, D. Laurendeau, and N. Teasdale, "Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene systems," in *IEEE Intelligent Vehicles Symposium Proceedings (IV)*, 2014.

[13] O. Lappi, P. Rinkkala, and J. Pekkanen, "Systematic Observation of an Expert Driver's Gaze Strategy–An On-Road Case Study," *Frontiers in psychology*, vol. 8, p. 620, 2017.

[14] S. Martin and A. Tawari, "Object of Fixation Estimation by Joint Analysis of Gaze and Object Dynamics," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[15] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: Theory, Algorithms and Software.* John Wiley & Sons, 2001.

[16] M. Schreier, V. Willert, and J. Adamy, "Compact Representation of Dynamic Driving Environments for ADAS by Parametric Free Space and Dynamic Object Maps," *IEEE Trans. Intell. Transport. Syst.*, vol. 17(2), pp. 367–384, 2016.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the Driver's Focus of Attention: the DR(eye)VE Project," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.