Gaze Target Tracking for Driver Assistance Systems

Dem Fachbereich
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades
eines Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von

Julian Jürgen Schwehr, M. Sc.

geboren am 27. August 1990 in Frankfurt am Main

Referent:	Prof. Dr.-Ing. J. Adamy
Korreferent:	Prof. Dr. rer. nat. H. Winner
Tag der Einreichung:	28. April 2020
Tag der mündlichen Prüfung:	15. Juli 2020

D17
Darmstadt 2020

Preface

This dissertation is the result of my work at the Control Methods and Robotics Lab of the Institute of Automatic Control and Mechatronics, TU Darmstadt. It was embedded in the PRORETA 4 project – a research cooperation with Continental. One of the first things I learned at the beginning of the research project was the origin of its name. The "proreta" was the officer at the bow on ancient Roman ships whose task was to keep an eye out for obstacles and shallows. The past five years have been an extraordinary journey on the sea called my life.

First of all, I want to thank my doctorate supervisor Prof. Jürgen Adamy for not only giving me the chance to go on this voyage but also for providing a pleasant working climate in his lab with the great possibility to sail wherever the wind of one's own ideas takes you. I am also grateful to Prof. Hermann Winner for his acceptance to act as second referee and even more for motivating and constructive feedback in the many PRORETA meetings providing fresh wind in one's sails.

I owe a special thanks to Volker Willert for all his time, support, ideas and confidence which helped me to sail around obstacles and shallows and not to capsize halfway on the trip.

Stefan Luthardt, Hien Dang, Maren Henzel and Nils Magiera were the best crew members I could wish for. The outstanding motivation to tackle any obstacle in the course of the project together made the daily work a real pleasure!

I am also deeply indebted to all Continental colleagues who contributed to the success of PRORETA and also this dissertation, namely Maximilian Höpfl, Benedikt Lattke, Christoph Wannemacher, Saman Khodaverdian, Ronald Bayer, Herbert Deckenbach, Johannes Eck, Alfred Eckert, Knut Ehm, Moritz Groh, Alexander Klotz, Ralph Lauxmann, Guido Mayer-Arendt, Karsten Michels, Rex Schilasky, Christian Thur, Manfred Wilck and all other colleagues who supported the project in the background.

To all my past colleagues at the Control Methods and Robotics Lab I want to say thanks for the great time and discussions giving fresh impetus to my work every day. Thanks to Hanno Winter and Moritz Bühler for proof reading this manuscript at an early stage and providing valuable

feedback.

I am very grateful to my parents who have been and still are supporting me in every step in the journey of my life no matter where I'm heading to. Thanks for being stable anchors in my life.

Last but not least I would like to give a special thanks to Katharina who started this project as my girlfriend, became my wife and ended the journey as mother of our son. Thanks for the everlasting support, for always cheering me up at times where nothing seemed to work and reminding me that there are more important things in life than this thesis.

Tettnang, in September 2020 Julian Schwehr

Car il suffit pour y voir clair de changer de perspective.

Antoine de Saint-Exupéry

Contents

Abbreviations and Symbols

Abbreviations

ABS	Anti-lock Braking System
Acc	Accuracy
ACC	Adaptive Cruise Control
AD	Automated Driving
ADAS	Advanced Driver Assistance System
ADF	Assumed Density Filter
DBN	Dynamic Bayesian Network
EEG	Electroencephalogram
EKF	Extended Kalman Filter
ESC	Electronic Stability Control
ETG	Eye-Tracking Glasses
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GPB	Generalized Pseudo-Bayesian
GUI	Graphical User Interface
HET	Head-Eye-Tracking
HMI	Human-Machine Interface
HMM	Hidden Markov Model
IMM	Interacting Multiple Model
IR	Infrared
IS	Intersection
KF	Kalman Filter
KL	Kullback-Leibler
LDS	Linear Dynamic System
LGS	Linear Gaussian System
LiLo	Looking in and Looking out
LKA	Lane Keeping Assist
MH	Multiple Hypothesis
MHMM	Multi-Hypothesis Multi-Model
MHMMP	Multi-Hypothesis Multi-Model tracking with gaze motivated parameter selection
MHMMPS	Multi-Hypothesis Multi-Model tracking with gaze motivated parameter selection and reduced sampling

ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MM	Multiple Model
MOT	Multiple Object Tracking
NCC	Normalized Cross Correlation
pdf	probability density function
PERCLOS	Percentage of eye closure
PF	Particle Filter
POI	Point of Interest
PoR	Point of Regard
Pr	Precision
Re	Recall
ROC	Receiver Operating Characteristic
TH	Threshold
TN	True Negative
TP	True Positive
TPR	True Postive Rate
TTC	Time To Collision
TTI	Time To Intervention
UKF	Unscented Kalman Filter
VO	Visual Odometry
VRU	Vulnerable Road User

Notation

x, X	Scalar
$\boldsymbol{x}$	Column vector
$\boldsymbol{x}^\top$	Row vector
$\boldsymbol{x}^{0:k}$	Sequence of vectors $\boldsymbol{x}^{0:k} = \{\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^k\}$
$\boldsymbol{x}^{0:t}$	Time sequence of vectors $\boldsymbol{x}^{0:t} = \{\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^t\}$
$\boldsymbol{x}^k$	kth vector of vector sequence $\boldsymbol{x}^{0:k}$
$\boldsymbol{x}^t$	Vector at time t of time sequence of vectors $\boldsymbol{x}^{0:t}$
$\{\boldsymbol{x}, \boldsymbol{y}\}^{0:t}$	short notation for $(\boldsymbol{x}^{0:t}, \boldsymbol{y}^{0:t})$
$\boldsymbol{X}$	Matrix
$\boldsymbol{X}^\top$	Transpose of matrix $\boldsymbol{X}$
$\boldsymbol{X}^{-1}$	Inverse of matrix $\boldsymbol{X}$
$\mathbf{0}$	Zero vector or matrix
$\mathbf{1}$	Matrix of ones

$\bar{\boldsymbol{x}}$	3D homogeneous coordinates of 2D point $\boldsymbol{x}$ (only in the context of projective geometry)
$\bar{\boldsymbol{X}}$	4D homogeneous coordinates of 3D point $\boldsymbol{X}$ (only in the context of projective geometry)
$f(\cdot)$	Scalar function
$\boldsymbol{f}(\cdot)$	Vector function
$\int_{\boldsymbol{x}}(\cdot)d\boldsymbol{x}$	Integration over the whole range of $\boldsymbol{x}$, e. g. if $\boldsymbol{x} \in \mathbb{R}^n$, then $\int_{\boldsymbol{x}}(\cdot)d\boldsymbol{x} = \int_{x_1} \dots \int_{x_n}(\cdot)dx_1 \dots dx_n$
$\int_{\boldsymbol{x}^{0:k}}(\cdot)d\boldsymbol{x}^{0:k}$	Integration over the sequence of vectors $\boldsymbol{x}^{0:k}$, i. e. $\int_{\boldsymbol{x}^{0:k}}(\cdot)d\boldsymbol{x}^{0:k} = \int_{\boldsymbol{x}^0} \dots \int_{\boldsymbol{x}^k}(\cdot)d\boldsymbol{x}^0 \dots d\boldsymbol{x}^k$
$p(\boldsymbol{x})$	Probability density function if $\boldsymbol{x}$ is a continuous random vector or probability mass function if $\boldsymbol{x}$ is a discrete random vector
$p(\boldsymbol{x}, \boldsymbol{y})$	Joint pdf of $\boldsymbol{x}$ and $\boldsymbol{y}$
$p(\boldsymbol{x}\|\boldsymbol{y})$	Conditional pdf of $\boldsymbol{x}$ given $\boldsymbol{y}$
$p(x_j)$	probability that a discrete binary random vector $\boldsymbol{x}$ is in state j, i. e. $p(x_j) = p(x_j = 1)$
$q(\boldsymbol{x}\|\boldsymbol{\theta})$	Approximating distribution with parameters $\boldsymbol{\theta}$
$\mathcal{N}$	Normal distribution

Important Functions and Transformations

$p(\boldsymbol{x}^k\|\boldsymbol{z}^{1:k})$	Filtering distribution or current belief of $\boldsymbol{x}^k$ given $\boldsymbol{z}^{1:k}$
$p(\boldsymbol{x}^k\|\boldsymbol{x}^{k-1})$	Transition density
$p(\boldsymbol{z}^k\|\boldsymbol{x}^k), \ell(\boldsymbol{z}^k\|\boldsymbol{x}^k)$	Measurement likelihood
$\mathcal{N}(x\|\mu, \sigma)$	Normal pdf of a random variable x with mean μ and variance σ
$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal pdf of a random vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathrm{KL}(p(\boldsymbol{x})\|\|q(\boldsymbol{x}))$	Kullback-Leibler divergence between distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$
$\mathrm{E}[\boldsymbol{x}]$	Expectation of random vector $\boldsymbol{x}$

Symbols

Latin Capital Letters

$\boldsymbol{A}$	System matrix of discrete-time linear dynamic system
$\boldsymbol{B}$	Input matrix of discrete-time linear dynamic system
$\boldsymbol{C}$	Measurement matrix of discrete-time linear dynamic system
K	number of samples
K'	number of free space samples
$\boldsymbol{K}$	Kalman gain matrix; Camera calibration matrix
$\boldsymbol{P}$	State error covariance matrix
$\boldsymbol{Q}$	Process noise covariance matrix
$\boldsymbol{R}$	Measurement noise covariance matrix
$\boldsymbol{R}_{cv}$	Rotation matrix between vehicle coordinate system and camera coordinate system
$\boldsymbol{X}_{g,c}$	3D point in camera coordinate system
$\boldsymbol{X}_{g,v}$	3D point in vehicle coordinate system
$\boldsymbol{X}_{\mathrm{PoR}}$	3D Point of Regard
Z_d	Depth of the scene at image point $\boldsymbol{x}_g$ obtained from stereo disparity map
Z_g	Depth of $\boldsymbol{X}_{g,c}$ in camera coordinates

Latin Lowercase Letters

b	Posterior belief
c	Normalization constant
d	Distance between point $\boldsymbol{x}$ and gaze origin $\boldsymbol{x}_o$ in 2D vehicle coordinates
d_{fs}	Interpolation distance between free space sample points
$\boldsymbol{f}$	Static environment measurement variable
g	Gaze yaw angle measurement variable
i	Running index, used at time $t-1$
j	Running index, used at time t
k	Running index, used for time sequence and sample points
l	Running index, used for particle filter sample points
$\boldsymbol{m}$	Dynamic objects measurement variable

n	State vector dimension; number of dynamic objects in object list
p	Input vector dimension
p_{ij}	Transition probability from model i to model j
q	Measurement vector dimension
r	Number of prediction models
$\boldsymbol{r}$	B-Spline curve with curve parameter s
s	Curve parameter of B-spline
$s^\star$	Curve parameter of B-spline of relevant spline points
$\boldsymbol{s}$	Discrete binary switching variable
$\boldsymbol{t}_{cv}$	Translation vector between vehicle coordinate system and camera coordinate system
$\boldsymbol{u}$	Input vector
$\boldsymbol{v}$	Measurement noise vector
$\boldsymbol{v}_{\mathrm{rel}}$	Relative velocity to ego-vehicle
$\boldsymbol{w}$	Process noise vector
$\boldsymbol{x}$	State vector; spatial coordinate
$\boldsymbol{x}_{\mathrm{a}}$	State vector of area of attention
$\boldsymbol{x}_{\mathrm{fs}}$	Free space sample point
$\boldsymbol{x}_g$	2D image coordinates (reprojection) of 3D point $\boldsymbol{X}_{g,c}$
$\boldsymbol{x}_o$	gaze origin, i. e. location of the driver's eyes
$\boldsymbol{x}_{\mathrm{P}}$	B-spline control point
$\boldsymbol{x}_{\mathrm{PoR}}$	Point of Regard in 2D image coordinates
z_c	Weighting factor of normal distribution obtained from multiplication of two normal distributions
$\boldsymbol{z}$	Measurement vector

Greek Letters

α	Gaze yaw angle (heading)
α_{ij}	Temporal switching weight when switching from target i to target j
β	Gaze pitch angle (heading)
β_{ij}	Spatial switching weight when switching from target i to target j
γ_{ij}	Process noise vector when switching from target i to target j
$\boldsymbol{\Gamma}_{ij}$	Process noise covariance matrix of γ_{ij}

δ_0	Constant model parameter in sigmoid function of spatial switching weights
ϵ	Weighting of points which do not belong to free space sample points
η	Normalization constant
θ	Angular difference between measured gaze yaw direction and 2D point in vehicle coordinates
θ_{s_i}	Angular difference between measured gaze yaw direction and closest point of object i
$\boldsymbol{\theta}$	Parameters of approximating distribution $q(\boldsymbol{x}\|\boldsymbol{\theta})$
$\boldsymbol{\theta}^\star$	Parameters of best approximation $q(\boldsymbol{x}\|\boldsymbol{\theta})$ of $p(\boldsymbol{x})$ in terms of the KL-Divergence
κ	Constant model parameter in sigmoid functions of switching weights
λ	Sample weight; mean sample weight
Λ	Mode likelihood function
$\boldsymbol{\mu}$	Mean vector
$\boldsymbol{\mu}_{oj}$	Measured position of object j
ν	Measurement noise variable of gaze measurement
π_i	Mode probability of model i
π_{ij}	Merging probability
$\boldsymbol{\Pi}$	Model transition probability matrix
$\boldsymbol{\Pi}_0$	Canonical projection matrix
ρ	Weighting of free space sample points
σ	Variance of gaze likelihood
$\boldsymbol{\Sigma}$	Covariance matrix
$\boldsymbol{\Sigma}_{oj}$	Covariance matrix with Eigenvalues and Eigenvectors such that major axes of the 90 % covariance ellipse correspond to width, length, and heading of object j
τ_0	Constant model parameter in sigmoid function of temporal switching weights
φ	Measured heading angle of objects in object list

Abstract

Despite many supporting systems, so-called advanced driver assistance systems (ADAS), human error is still by far the main cause of traffic accidents. In the development of new driver assistance concepts, systems and functions monitoring the driver while driving and classifying their behavior in the driving context are therefore increasingly coming to the fore. In this context, this dissertation deals with the question what the driver perceived in their environment. For this purpose, the information of the environment model has to be merged with measured gaze data. Given a precise calibration of the individual sensors, visual fixations of the driver on road users are modeled.

Based on the realization that simple geometric approaches cannot answer this question of visual fixations precisely enough, characteristics of human gaze behavior are identified and integrated as model knowledge into a probabilistic tracking approach. This tracking model considers every object which is classified as a dynamic object and thus as a potential road user by the vehicle's environment perception module as a possible hypothesis for the driver's current visual attention target. In addition, two different motion models of eye movements for fixations and saccades are integrated, so that the estimation of the gaze target can follow the special dynamics of human gaze and recognize specific connected time spans. The advantage of this novel resulting Multi-Hypothesis Multi-Model (MHMM) filter is the confidence which is characteristic to probabilistic approaches, indicating the probability of each object being fixated by the driver.

A challenge is the evaluation of such new algorithms. For the statement which object the driver actually visually fixates, ground truth information is necessary. However, this cannot be covered by questionnaires. For this reason, a reference data set is created in which the recordings of the remote eye-tracking system installed in the vehicle are extended with the data of wearable eye-tracking glasses. With the help of these recordings, different model approaches are now compared on a quantitative and not only qualitative basis.

The prototypical City Assistant System, which was co-developed as part of this work, shows how the newly gained information about the driver's

gaze behavior can be incorporated into new assistance concepts. It adapts its warning and recommendation cascade in urban intersection scenarios to the driver's driving style and gaze behavior. Through this orientation towards the driver's need for support, the City Assistant System contributes to higher acceptance of warning and recommending systems and ultimately to increased road safety.

Kurzfassung

Trotz vieler unterstützender Systeme, sogenannter Fahrerassistenzsysteme (FAS), sind menschliche Fehler immer noch mit großem Abstand die Hauptursache für Verkehrsunfälle. Bei der Entwicklung von neuen Fahrerassistenzkonzepten rücken daher verstärkt Systeme und Funktionen in den Vordergrund, die den Fahrer während der Fahrt beobachten und sein Verhalten im Fahrkontext einordenen und bewerten. In diesem Rahmen behandelt die vorliegende Dissertation die Frage, was der Fahrer in seinem Umfeld wahrgenommen hat. Hierzu sind die Informationen des Umfeldmodells mit den Messdaten der Blickrichtung zu fusionieren. Eine präzise Kalibrierung der einzelnen Sensoren vorausgesetzt werden visuelle Fixationen des Fahrers auf Verkehrsteilnehmern modelliert.

Basierend auf der Erkenntnis dass einfache geometrische Ansätze diese Frage nach visuellen Fixationen nicht klar genug beantworten können, werden zunächst Eigenschaften des menschlichen Blickverhaltens identifiziert und als Modellwissen in einen probabilistischen Trackingansatz integriert. Dieses Trackingmodell berücksichtigt jedes Objekt, welches von der Umfelderfassungssensorik des Fahrzeugs als dynamisches Objekt und damit als potentieller Verkehrsteilnehmer engestuft wird, als mögliche Hypothese für das aktuelle visuelle Aufmerksamkeitsziel des Fahrers. Zusätzlich sind für Fixationen und Sakkaden der Augenbewegungen zwei verschiedene Bewegungsmodelle integ-riert, sodass die Schätzung des Aufmerksamkeitsziels der speziellen Dynamik des menschlichen Blicks folgen und gezielt zusammenhängende Zeitspannen erkennen kann. Der Vorteil dieses neuen resultierenden Multi-Hypothesen Multi-Modell (MHMM) Filters besteht in der für probabilistische Ansätze charakteristischen Konfidenz, die für jedes Objekt angibt, gerade vom Fahrer angesehen zu werden.

Eine Herausforderung besteht in der Bewertung solch neuer Algorithmen. Für die Aussage, welches Objekt der Fahrer tatsächlich visuell fixiert, sind Referenzwerte notwendig, die nicht über Fragebögen abgedeckt werden können. Aus diesem Grund wird ein Referenzdatensatz erstellt, bei dem die Aufnahmen des entfernten, im Fahrzeug verbauten Eye-Tracking-Systems mit den Daten einer tragbaren Eye-Tracking-Brille erweitert werden. Mit Hilfe dieser Aufnahmen werden verschiedene Modellansätze nun quantitativ

und nicht mehr nur qualitativ miteinander verglichen.

Wie die neu gewonnene Information über das Blickverhalten des Fahrers in neue Assistenzkonzepte einfließen kann, zeigt der prototypische City Assistant, welcher im Rahmen dieser Arbeit mitentwickelt wurde. Dieser passt seine Warn- und Empfehlungskaskade in innerstädtischen Kreuzungsszenarien an den Fahrstil und das Blickverhalten des Fahrers an. Durch diese Orientierung am Unterstützungsbedarf des Fahrers leistet der City Assistant einen Beitrag zu höherer Akzeptanz von warnenden und empfehlenden Systemen und letztlich zu höherer Verkehrssicherheit.

1 Introduction

From a control point of view, driving a car is a relatively simple task, consisting eventually only of accelerating and braking as well as steering. The difficulties arise by the complexity of traffic which a driver has to perceive and understand in order to act accordingly. But already in the seventies, it was no secret that human factors were the most frequent cause of traffic accidents. In a study for the U.S. National Highway Traffic Association (NHTSA), Treat et al. reported in 1979 that human errors definitely caused at least 64 % of accidents and were probably causes in over 90 % of the investigated accidents [188]. In order to increase road safety, besides passive safety systems such as the safety belt or the airbag, the last three decades experienced a large development effort for active safety systems. These active safety systems are intended to "improve road safety in terms of crash avoidance, crash severity mitigation and protection, and automatic post-crash notification of collision. [...] More generally, some driver support systems are intended to improve safety whereas others are convenience functions" [44]. This definition of Advanced Driver Assistance Systems (ADAS) is eventually not restricted to modern systems like Adaptive Cruise Control (ACC) or Lane Keeping Assist (LKA), it also applies to the groundbreaking systems of anti-lock braking (ABS) and electronic stability control (ESC)[1]. One goal of ADAS is to relieve the driver from parts of the driving task or to additionally support and protect the driver leading to higher passenger safety and with increased market penetration also to overall higher traffic safety. Current ADAS have reached a high degree of maturity and diversity some of which have computational perception and function models which are expected to enable conditional automated driving. Nowadays, conditional automated systems such as so-called highways pilots and fully automated systems for restricted areas such as automated valet parking are in the starting blocks. These systems form the basis for the transition to higher automated levels of driving in the area of individual mobility (cf. Fig. 1.1). By taking the driver out of the loop, full automation is linked to the expectation of higher road

[1] While early active safety systems have originally been termed 'Driver Assistance Systems' only, the above mentioned definition does not specifically differentiate between

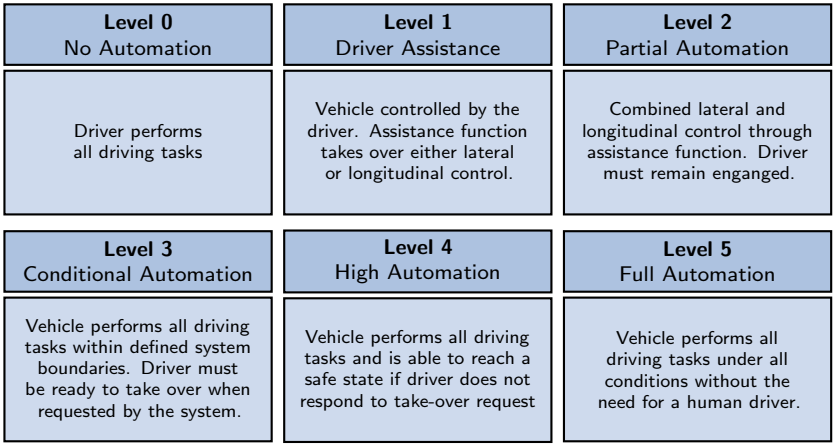


Figure 1.1: Simplified description of the levels of automation defined in the SAE J3016 standard [148].

efficiency and to the promise that drivers can use their driving time freely. Nevertheless, open regulatory and societal questions, technical hurdles and eventually also the open question on affordability let to expect that human drivers will in parts stay in control and in responsibility and that ADAS and lower levels of automation will coexist next to emerging systems of higher automation. This leads to the proposition that current ADAS will still be further improved and their functionality extended. Based on the supposed maturity of current systems, the legitimate question is what future assistance systems are to achieve.

Current accident statistics from 2018 in Germany show that the percentage of human errors as accident cause are still on a comparable level as in the seventies. According to [171], human error was responsible for 88.4 % of all reported accidents. In recent years, the total number of registered accidents increased while accidents with personal injury in the same period slightly decreased. This relationship suggests that the increasing market penetration of ADAS indeed contributes to an overall crash severity mitigation. The question is, though, whether ADAS can be designed such that even more accidents could be prevented. One hurdle lies in the warning dilemma especially of visually or acoustically warning systems and active intervention systems. A warning is more effective the earlier it is raised.

‘advanced’ and ‘normal’ systems.

Yet, at the same time, the probability of a false warning is increasing due to the potential of a changing environment meaning that the expected situation of which is warned is not happening. The dilemma is given by the fact that with increasing false system reactions the acceptance of a system decreases [45]. For this reason, on the one hand, current systems are designed to warn and intervene as late as possible making interventions heavier and warnings more noticeable than necessary. On the other hand, current ADAS still exhibit unused potential in the field of user acceptance. One contribution to the mitigation of the classical warning dilemma is the incorporation of driver knowledge, e. g. an estimate of the driver's mental state, in the design and development of ADAS. An extended comprehension of the driving situation including the driver can lead to a better appraisal of the driver's actual need for assistance. It can furthermore be used to adapt the warning strategy or simply increase the time to intervention. The potential benefit of such a system is thus that it can act like an always attentive co-pilot that specifically aids the driver without overloading them[2] with distracting unnecessary advice [106], leading to a closer cooperation between the vehicle and the driver. This goal involves the deployment of new driver monitoring sensors in the vehicle's interior in order to consider the driver's behavior and reactions in the modeling of a new ADAS. Driver monitoring will be one of the topics of the future. This is also shown by the intention of the Euro NCAP 2025 roadmap to include driver monitoring in the passenger safety testing scheme by 2024 [43]. This proposition will probably lead to a large coverage of vehicles equipped with driver monitoring systems. Within this broader area, research activities on camera based driver gaze estimation and its matching to the surrounding environment have been growing in recent years. The reason why a large share of the research effort is dedicated to the driver's gaze is the fact that most traffic relevant information (about 80-90 %) is perceived via the visual channel [1]. Where and at what the driver is looking is an indicator for their situational awareness. At the same time, what the driver has not seen in combination of what is actually relevant determines their true need for assistance. As a consequence, future ADAS need to have the ability to reliably perceive and identify such situations.

A fictitious example of such a novel ADAS is given in 2003 in [60]. A distracted driver is approaching an intersection with red traffic lights. Since the car is not slowing down, pedestrians are hesitating whether it is safe to

[2]Since this work mentions the driver and his or her attention, gaze, etc. many times, the author decided to use singular 'they' throughout this work in order to be as gender-neutral as possible [16].

cross the road. The (at this time) imaginary ADAS warns the driver since it has registered the red traffic light on the one hand *and* has detected that the driver has overlooked it on the other hand. A similar example is quoted almost ten years later. "Imagine a driving assistance system that records the driver's eye movements and analyzes them to warn the driver about entities (e. g. traffic participants) [they] might have overlooked. An essential requirement for such a system is the online analysis of the driver's scanpath[3] with respect to the entities that appear on the visual scene" [176].

It is the author's hope that after reading until the end of this thesis, the reader finishes with the impression that these fictitious scenarios are nowadays within reach.

1.1 Contributions

The first challenge when measurements of the driver's gaze are available as information source in the vehicle is to relate it to the available situation information. A suitable system calibration and time synchronization provided, the first question addressed in this thesis is:

1. How can the driver's gaze be combined with knowledge about the environment based on the given sensors?

As will be argued later in this thesis, the driver's gaze motion is highly task dependent and it can be expected that they will look at other road users that are relevant to the driving task. Based on the obtained insights from the first question, the second question is thus formulated as:

2. How can knowledge on human gaze behavior be used for the determination at which traffic participant the driver is looking?

Since human perception can hardly be measured, it would be at least desirable to know whether supposedly seen road users have actually been looked at by the driver. The estimation of the driver's gaze target[4] leads thus automatically to the third question of this thesis:

[3] The visual scanpath is the sequential order in which 'things' are looked at.
[4] The object of current visual attention is denoted as gaze target in this work.

3. How can the determination of the current gaze target be evaluated? How could a quality assessment look like?

Finally, now that all this effort has been undertaken to find out at what the driver has probably been looking, the natural following question is "what for?":

4. How can the driver's gaze and viewing behavior be incorporated into the design of future ADAS?

This thesis tackles the above introduced research questions leading to the contributions of this work which can be summarized as follows.

- In-depth analysis and discussion of fusion spaces and basic geometric models for gaze target estimation
- Novel gaze target tracking algorithm [158–160]
  - application of probabilistic tracking methods
  - incorporation of knowledge of human gaze behavior
  - obtain gaze target as part of the existing environment representation, i. e. a probability is assigned to each entity in the scene
- Novel evaluation approach for remote eye tracking systems in automotive scenarios [156]
  - creation and publication of reference dataset
  - implicit gaze calibration procedure
- Realization of gaze based prototypical ADAS [157]
  - adaptation of warning strategy according to gaze and acceleration behavior of the driver
  - implicit gaze calibration procedure

Just like the research questions, the contributions build harmoniously on each other, starting with a reflected view on how the driver's gaze can be incorporated as additional source of information. The next step involves the actual data fusion which has been implemented on the project's test vehicle followed by the attempt to assess the developed model's performance in comparison to existing models from the literature. In the final step, the

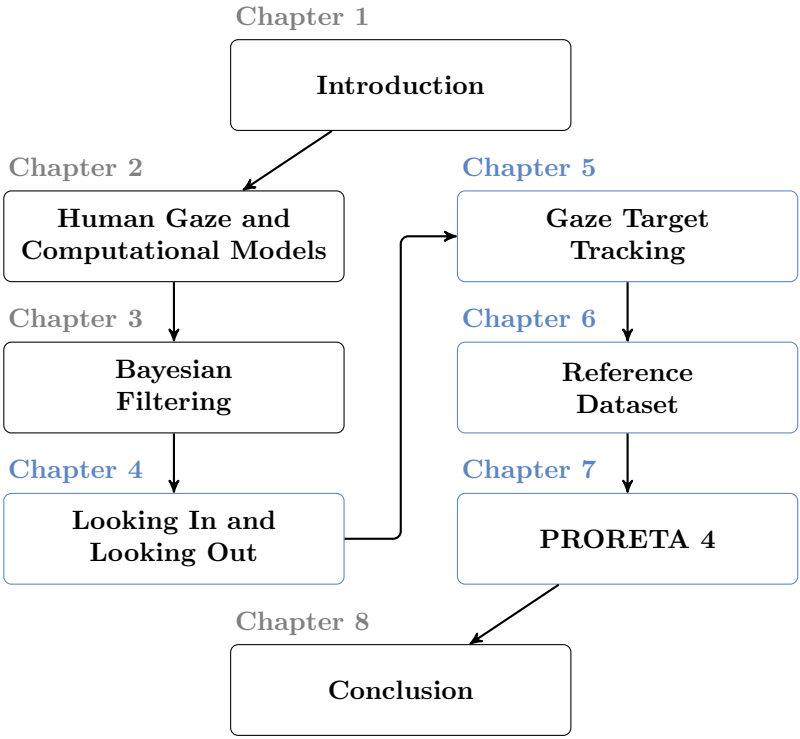


Figure 1.2: Outline of the dissertation. Chapters containing contributions are highlighted in blue.

gaze target estimation is embedded into a prototypical ADAS in a test vehicle operating in real world traffic.

Individual contributions have been published to the Intelligent Transportation Systems community in the author's publications [156, 158–160] and the project's joint publication [157].

1.2 Outline of the Dissertation

The outline of this dissertation is depicted in Fig. 1.2. Even though all chapters are self-contained and can be read individually, the author's chain of thought can be best comprehended if the Chapters 4 – 6 are

read one after another with the foundation Chapters 2 and 3 before. Following this introduction, the thesis is started out with foundations on human gaze behavior in Chapter 2. Along with the psychophysical and physiological foundations which are picked up in the course of the thesis at various instances, also related work on computational models is presented. These include the tracking of gaze with camera systems, individual gaze calibration, fixation detection algorithms and computational models for human visual attention. Completely switching the topic, Chapter 3 introduces probabilistic Bayesian filters with a focus on the employed algorithms. In Chapter 4, the automotive frame for the main contributions of this work in the subsequent chapters is presented. Here, related work on inside-outside sensor fusion approaches and estimation of the driver's awareness of road users is presented. Along with the discussion on related work, different fusion approaches and their drawbacks are discussed in more detail. Chapter 5 presents the author's approach to the driver's gaze target estimation in the form of a specially tailored tracking approach. The probabilistic formulation of what the driver looks at allows to incorporate knowledge on human gaze behavior in form of model knowledge and takes measurement noise into account. One drawback of the specific research area is the lack of proper reference data. This problem is tackled in Chapter 6, where the creation of a reference dataset is outlined. Using this information, different gaze target estimation approaches are quantitatively compared. Up to that point in the thesis, the focus is on the explicit determination where and at what the driver looks. In Chapter 7, this insight into the driver's visual behavior is set in a more general context. First, the potential benefit of knowledge on the driver's situation awareness for ADAS is discussed and with the PRORETA 4 City Assistant System, which the author co-developed in the course of the thesis, an exemplary implementation in a prototypical assistance system is presented. The final chapter summarizes the main results and draws a conclusion of the presented work providing also an outlook to possible future work.

2 Foundations of Human Gaze and its Computational Models

In this chapter, a short overview on the main aspects of human gaze are introduced that form the foundations of this work. Additionally, a brief overview on the state-of-the-art computational models related to each topic is given as these models partially build the general technical basis and related work of this thesis. When the proposed gaze target tracking algorithm and evaluation method are presented later in the Chapters 5 and 6, it is helpful to have an idea of the anatomy of the human eye, its motion characteristics as well as how people direct their perception and attention. Throughout the thesis, different aspects of this chapter are taken up.

2.1 The Human Eye and Gaze Estimation

2.1.1 Anatomy

The geometric and physical structure of the human eye is quite complex. A schematic view of an eye with only a selection of ocular axes and their angles taken from [128] is depicted in Fig. 2.1. When light is falling into the eye from the *fixation point* (marked with FP in the image), in the gaze tracking context also called *Point of Regard* (PoR), it is refracted on various surfaces including cornea and lens. This *line of sight* falls on the retina at the *fovea centralis*, a small area with a high density of cones responsible for sharp view. The fovea is the region of highest acuity on the retina of the human eye, enclosed by the parafovea and the peripheral region. The area of sharp view corresponds to approximately the central 2° of vision. The acuity in the parafovea, which extends out to approximately 5° on either side of the fovea, is not nearly as good [144] as the density of cones is highly decreased due to a predominant density of rods [88]. As can be seen in Fig. 2.1, the *visual axis* connects the fovea and the fixation point by way

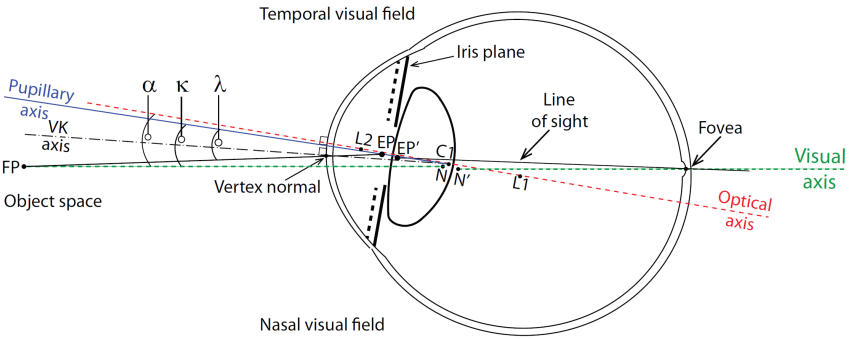


Figure 2.1: Schematic sketch of a selection of ocular axes. Image taken from [128][6].

of the nodal points[5]. It is composed of the line segments that connect the fovea with the first nodal point of the eye and the segment that connects the second nodal point with the fixation point. These two segments, which are strictly speaking no straight line, have the same inclination angle [10]. The visual axis is usually close to the line of sight and can thus be regarded as approximation of the same. This becomes relevant in Section 2.1.2 dealing with computational models of gaze estimation. The *optical axis* of the eye can be seen as the "best fit" of a line through the centers of curvature of each refracting surface. The optical axis, also called *line of gaze* [57], has no particular importance by itself for the optics of the eye but it can be seen as useful reference [10]. In the literature, deviations of the optical and visual axis of about 5° are assumed while measurements report a range of +17° (nasal object space, i. e. towards the nose) to −2° (temporal object space, i. e. towards the temple) horizontally and 2° to 3° vertically [128].

2.1.2 Gaze Estimation

The main part of this work, namely the fusion of the driver's gaze with the surrounding scene, relies on the gaze estimation output of a remote eye tracking module. Therefore, the technique behind gaze estimation based on

[5] Nodal points have the special property that a ray passing from an off-axis point towards the first nodal point appears to pass through the other nodal point on the side of the lens [10]. The concept comes from the field of optics when dealing with thick lenses.

[6] The items α, κ, λ in the sketch do not refer to quantities in this work.

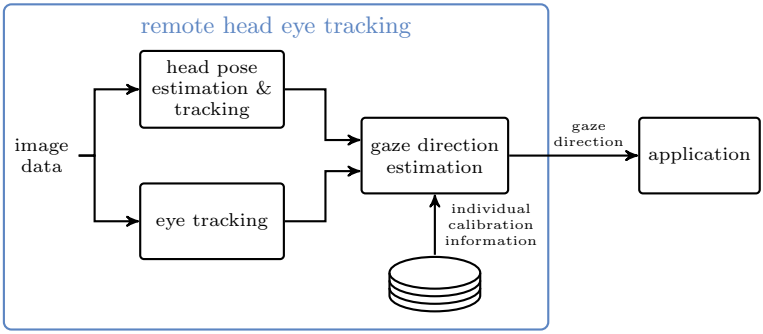


Figure 2.2: Components of the gaze estimation process. The output of the gaze tracking module, i. e. gaze origin and direction coordinates at each image frame, can be utilized in possible applications. The contributions of this work are within the application. There, the gaze estimation is fused with the surround environment (see Chapters 4 to 7).

corneal reflections in digital videos[7] is outlined. Fig. 2.2 shows a simplified depiction of the necessary components of the gaze tracking process each of which being a research field on its own. Prior to the actual interpretation of the eye position in the image, the detection of existence and location of the eye is necessary. When dealing with video data, the detection goes hand in hand with tracking the eyes' positions and shapes over multiple image frames. Quite obviously, the knowledge of the eye positions in the input images is a necessary prerequisite for computing the gaze direction. For remote eye tracking, this also involves the detection and tracking of the head since a person's gaze is determined by the head pose and the eyeball orientation [57, 126]. In case of 3D model based gaze estimation, knowledge about the 3D eyeball positions may obviate head tracking since the head pose can be inferred from the eye positions [57]. However, in practical applications, head tracking often remains one preceding step due to robustness. For the two topics of head and eye tracking, the interested reader is referred to further literature since the vast amount of methods is out of scope of this dissertation. The detection and tracking of head motion is surveyed in [126] while [57] gives an extensive overview on the detection and tracking of eyes as well as gaze. An overseeable summary on both is given by [3]. Large parts from eye and head detection all the

[7]This is state of the art in many non-intrusive gaze estimation systems. Other techniques and a short overview on the history of gaze tracking are presented in [39].

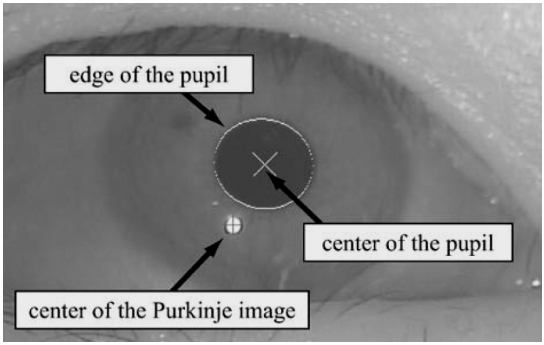


Figure 2.3: Image of an eye under IR illumination with corneal reflection also called glint or Purkinje image. Image taken from [56].

way to possible applications are covered in [56] and a practical guide to the methods and measures of eye tracking is given by [66].

Nevertheless, the gaze estimation itself is shortly outlined. The goal is to compute the line of sight as accurately as possible as three-dimensional vector in space starting at the eye position. For this, it has proven helpful to use a calibrated multi-camera system with multiple infrared (IR) light sources [55, 57]. A prerequisite step is the intrinsic and extrinsic calibration of the camera system in order to set up a common three dimensional Cartesian coordinate system. The IR light sources are used since each of them yields a reflection in every eye. These corneal reflections are called *first Purkinje images*[8] or *glints*. An example is shown in Fig. 2.3. Beside that the glints are relatively easy to locate even in challenging lighting conditions, their position can be related to the optical axis through the physiology of the eye as well as the hardware system's configuration. This made feature and 3D model based estimation the most popular approach for gaze estimation [57]. In automotive applications, methods based on active IR illumination are furthermore the only way to handle the broad variety of conditions that the system faces such as night or sunlight. Typical 3D approaches assume spherical eyeball and cornea surfaces. Given these geometric model assumptions, together with the positions of the glint(s) and an estimate of the pupil center via methods of image processing, the optical axis can be reconstructed [55]. The main error sources in this procedure

[8] Purkinje images are named after the Czech anatomist Jan Evangelista Purkyně (1787–1869).

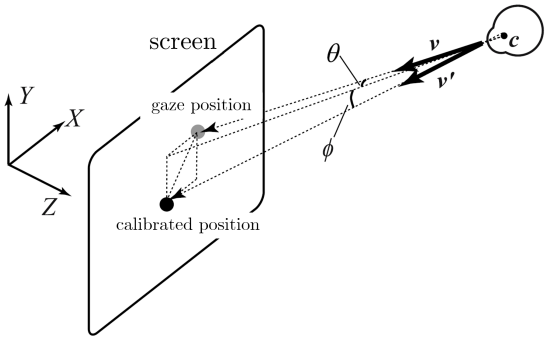


Figure 2.4: Personal calibration of the user's gaze. The deviation between optical and visual axis is compensated by the calibration. Image taken from [56].

according to [55] are the discrepancy between the shape of real and assumed corneas, as well as the noise in the computation of the pupil center and the glint position. Since the true line of sight is assumed to coincide with the visual axis in the vast majority of computational models (cf. Fig. 2.1) [57], the second step thus deals with the reconstruction of the visual axis from the optical axis. For this, their deviation due to anatomic reasons needs to be defined through a personal calibration process[9], in which the measured optical axis is compared to a "known" visual axis. The visual axis is known because the subject is asked to fixate a point with known 3D coordinates, e. g. a calibrated point on a screen. The idea is shown in Fig. 2.4. In the case of a multi-camera system with multiple light sources, it is possible to reduce the process to a simple one-point calibration procedure, since no knowledge of the subject-specific eye parameters[10] are necessary [55]. Even though the subject-specific deviation of optical and visual axis can

[9]If no multi-camera system with multiple light sources is used, even more reasons for personal calibration exist [56]: 1.) Personal difference in eyeball size and shape. 2.) Due to the refraction at the corneal surface, the observed pupil position differs from its real position. Using multiple calibrated cameras and multiple light sources, terms containing eye parameters affecting 1) and 2) can be discarded in the set of equations [55]. 3.) Differences in screen positions and resolutions which, however, only become relevant if a mapping of gaze on a screen is performed. 4.) Eye positioning error due to free head movement. By applying multiple cameras and 3D models, the 3D position of the eyes can be recovered.

[10]The parameters are: radius of cornea, distance between pupil center and center of corneal curvature and the effective combined index of refraction of the aqueous humor and the cornea.

be estimated from only one point, many systems rely on several calibration points. From the determined pan and tilt angles between the two axes together with the head pose, the line of sight can be recovered. Due to the usage of multiple cameras, the 3D position of the eyeball, also called *gaze origin*, can as well be computed. As all information is represented in three dimensions, gaze can be expressed as vector in 3D camera coordinates with known origin.

In this work, gaze direction data from a commercially available *head-eye-tracking* (HET) system are used. The system used throughout this work is a multi-camera SmartEye system of version 6.2 installed in the prototype vehicle. It consists of four cameras and two IR illumination flashes. The system performs head and gaze tracking as well as individual gaze calibration if points with known 3D positions are provided to the system. Together with an extrinsic calibration of the head-eye-tracking system in vehicle coordinates[11], gaze and head information can be transformed from camera coordinates to vehicle coordinates in order to enable the fusion of the driver's gaze with the current situation (see Chapters 4 and 5). Except for Chapter 7, it is assumed that subjects are individually calibrated so that the measurement output of the tracking system corresponds to a measurement of the line of sight. Fig. 2.5 shows the gaze estimation in the prototype vehicle.

2.2 Gaze Motion Characteristics

2.2.1 Basic Gaze Motion

As presented in the previous Section 2.1.1, the highest acuity on the retina of the human eye is on the fovea and it decreases drastically within a few degrees off the fovea. Thus, in order to perceive a scene, the eyes must be moved so that the point of interest (POI) falls on the fovea making eye movements essential for the perception of a scene. Even though cognitive perception involves not only the eye motion but also memory, cognitive attention, cognition and decision making [84], guiding the eyes to the region of interest is part of the first steps in perception. Since all neural and motor processes require time [68], gaze motion can be observed to exhibit periods of relative stability which are interspersed with rapid jerky movements [97]. These two basic movements are called fixations and saccades. Furthermore, the fixational movement can also be a smooth

[11] cf. Appendix A for further information on the system calibration.

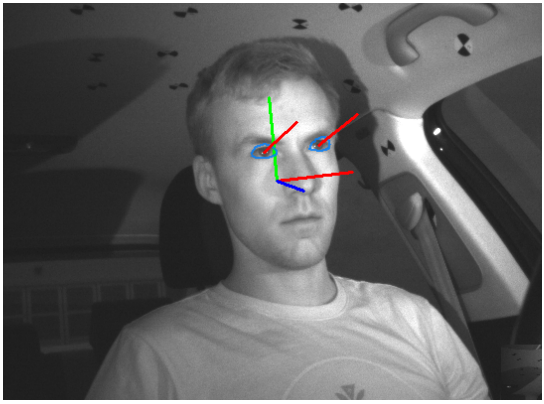


Figure 2.5: Gaze estimation in the prototype vehicle. Gaze direction as well as eye and head position are directly transformed from camera coordinates and output in vehicle coordinates, as the vehicle and the head-eye-tracking system are fully calibrated.

pursuit if the point of interest is moving relatively to the observer. In this section, these three most relevant movement characteristics are pointed out in more detail, as they are most relevant for the gaze tracking model presented in Chapter 5. As human eye motion and cognitive perception are important and well researched topics in the fields of physiology and psychology, the interested reader is referred to the book of Land and Tatler [95]. A thorough collection on eye movement research is given by Kowler [90] or Lappi [97]. Explanations with connections to eye tracking are given in the book of Duchowski [39]. The following information on eye movements is gathered from these mentioned sources.

Saccades Saccades[12] describe the rapid eye movements which change the image on the fovea by redirecting the direction of gaze to a new location in the visual environment. The maximum frequency of saccades is only about $4\,\mathrm{s}^{-1}$ with fixations or smooth pursuits in between [39]. Even though different durations of saccades between around 10-300 ms have been reported [97], the typical duration is below 100 ms [39]. In reading tasks for example, the typical saccade duration is around 30 ms whereas in a scene perception task, saccades last about 50 ms [144]. Due to the high speed of

[12]The French term *saccade* means jerk.

up to 500-800°/s [95, 97], the human brain is incapable of processing the incoming information during a saccade. This effective blindness[13] is called saccadic suppression [97]. Saccadic movements can be either voluntarily executed or occur reflexively if a stimulus in the environment catches the viewer's attention. Furthermore, they are performed conjugately, meaning that they are simultaneous, of the same size, and in the same direction for both eyes [95].

Fixations During fixations, the eye is kept relatively stable with almost no motion. In a simplifying understanding of human perception, the image on the fovea during a fixation is the current focus of attention of the viewer and if the fixation duration is sufficiently long, the scene information is also perceived. In [97], eye fixation is therefore compared to a "window" through which the environment is perceived interrupted by saccades. This can be interpreted such that humans are constantly sampling information from the world. The minimum fixation duration depends on tasks and subjects. Fixations have minimum durations of approximately 100 ms [98], however, common durations are reported to be about 200-300 ms [84, 144] but can take much longer [84].

In [97] it is pointed out that the term "fixation" is used interchangeably when referring to oculomotor fixation in physiological literature and to fixating the point of regard in the naturalistic eye movement literature. A fixation in the physiological sense describes stabilization of the eye in relation to the head. When stabilizing the gaze target on the fovea, however, the actual oculomotor control is highly dependent on the task and a relative motion between viewer and target. In many naturalistic studies, the eye movement pattern is characterized by fixate-saccade-fixate sequences without strict references to the physiological eye motion types. As [97] admits, there is no clarifying notation so that in the remainder of this thesis, the term fixation is used as in the naturalistic eye movement literature. Often, smooth pursuits as described in the following paragraph are contained in this notion.

Smooth Pursuits This third kind of eye movement occurs when the eye follows a small moving target[14]. By moving the eye slowly and consistently

[13] According to [95], this means that roughly one tenth of the waking day, about 1.5h, humans are effectively blind.

[14] The target can also be a certain point on a larger target, e. g. the licence plate or emblem on a car.

with the target's motion, the image of the target is kept fixed on the fovea and thus, information during a pursuit is also processed. In contrast to saccades, which can be voluntarily directed towards any location in space, pursuits cannot be performed voluntarily without a target to track. An attempt to do so normally results in a sequence of saccades [97]. Interestingly, humans are able to maintain the pursuit even if the target is momentarily occluded [97]. The tracking of an object via a smooth pursuit alone is only possible to speeds up to about 15°/s. At higher speeds of the target, the gaze tends to lag behind. Then, gaze motion is supplemented by catch-up saccades to compensate this lag [95, 97]. These limits refer to the angular velocity of the eye relative to the head. As gaze motion is usually accompanied with head motion, the velocity ranges of targets that can be tracked with smooth pursuits is actually extended.

As the driver of a car as well as other road users are mostly moving, gaze targets in the surrounding environment are in relative motion to the driver. Thus, it can be assumed that most fixational eye movements[15] in the domain of driving are smooth pursuits. In the gaze target tracking model of Chapter 5, fixations are included in the model as a special case of targets moving with zero relative velocity.

Further Types of Eye Movements When dealing with the topic of eye movements, the interested reader may stumble across further types of eye movements reported in the literature including drift, glissades, vergence, micro-saccades, vestibulo-ocular response or optokinetic response. However, in remote eye tracking, these types do not play a major role. This is mainly due to the basic model of information intake which assumes that only during fixations information can be processed. These additional motion patterns are either too small to be accurately measured by a remote eye tracking system or they are side effects of world and head motion in order to keep the gaze target stable on the retina.

In the list of the above, it is worth to shortly discuss vergence since theoretically, the point of regard could be determined from the intersection of the two lines of sight of both eyes. Vergence is responsible for adjusting the angle between the eyes so that the fixation point is on the line of sight of each eye similar to a point that is visible in the two images of a stereo camera. Consequently, vergence normally occurs in the opposite direction for each eye relative to the other [95]. As the pupil distance

[15] Remember that the meaning of fixation is used in the understanding of the naturalistic eye movement literature referring to stabilize the visual target on the retina.

is often very small in comparison to the distance of the fixation point, computing its 3D location based on the angular difference between left and right eye would be very inaccurate. It is therefore more common that eye tracking systems provide a joint estimate of the gaze direction based on the directions measured for both eyes. Thus, vergence is of no further importance for this thesis, however, it is pointed out that for very small workspaces at distances like a standard desktop, the computation of the 3D position of the point of regard can be computed using vergence [61].

2.2.2 Computational Models of Fixation and Saccade Detection

Due to the characteristic described in the section above that visual information intake only occurs during fixations or smooth pursuits, the detection of fixations and saccades is of interest in many fields of research and development dealing with human visual attention. A short overview on fixation detection algorithms is presented in [84], taken mostly from [66]. According to these, fixation detection approaches can roughly be classified in three main categories: dispersion based, velocity and acceleration based as well as probabilistic models based on statistical Markov[16] models.

Dispersion based methods are based on the distance between subsequent data points in a time window. Fixations are identified by finding sequences of data points that are close enough to one another. All other events are collected in one "no fixation" category. The second group bases the classification on the velocity and/or the acceleration from the previous data point to the following. Velocity and acceleration based models are quite simple and exhibit very low latency since they only have to consider two data points. However, these methods have their difficulties with detecting smooth pursuits [84] since the angular velocity of pursuits with underlying head motion can be close to the angular velocity of slower saccades. Therefore, some methods rather focus on the detection of clear saccades and assign everything else except for blinks to fixation clusters [66]. The third group, probabilistic models, are popular for online detection due to the ability to handle successively arriving data points. They estimate the probability of a data sample belonging to the class of fixations or saccades based on the previous state and the transition probability. In [84], a probabilistic model was enhanced with postprocessing to detect smooth

[16] The Markovian assumption is introduced in the next chapter in the context of probabilistic models.

pursuits based on eigenvector and eigenvalue analysis of the distribution of gaze points. The tracking module presented in Chapter 5 can as well be interpreted as rough fixation/saccade detection system even though this is not its aspiration. From this point of view, the model would be closest to the class of probabilistic models as it is also based on a Markov chain.

2.3 Visual Attention

2.3.1 Characteristics of Visual Attention and Gaze Behavior

An informal definition of attention is given for example in [29]: "Attention defines the mental ability to select stimuli, responses, memories, or thoughts that are behaviorally relevant among the many others that are behaviorally irrelevant." This definition has many aspects but stresses the fact that attention above all is directed towards regions of interest. At each moment, only small parts of a scene are analyzed and only by changing the focus of attention, a rich representation of the visual world is obtained [50]. Most of the time, the visual focus of attention is the region that is currently fixated by the eyes, however, the cognitive focus of attention can lie elsewhere [50]. As a person's mental state is very difficult to measure, their gaze has often been used substitutionally to estimate attention [22]. To make this more evident, the term "visual attention" is used. While the former section dealt with the mechanical motion of the eyes, the concept of visual attention is mainly concerned with the computational mechanisms that guide the gaze over a scene [22]. Visual scenes contain too much information to process them as a whole [58]. Together with the characteristics of the eye, this leads to an active sampling of the environment through a sequence of fixations. Even though it is possible to mentally attend to locations in the peripheral view (*covert attention*), visual attention is mostly associated with eye movements, i. e. fixations (*overt attention*) [50]. A relevant example in the context of driving is given by [22]: a driver can keep their eyes on the road overtly monitoring the car's position in the lane while at the same time covertly monitoring the status of lights or the motion of other road users. The fact that visual intake from the peripheral field of view can be perceived is an important insight and must be acknowledged. E. g., another road user does not necessarily have to be fixated by the driver in order that they are aware of them. However, covert attention and subsequent saccades towards interesting locations enabling perception at a higher resolution often work

together [22, 50]. The possibility of covert attention and peripheral vision should be kept in mind throughout this thesis. However, as outlined here and argued later, it is very unlikely that traffic hazards are perceived without visual fixations. Thus, the focus remains on the mechanisms that guide gaze direction and fixation sequences over a scene.

Why do we look where we do? This question for the processes behind visual attention can also be reformulated as how do humans decide where to look. It has been raised in early works such as in [26] by Buswell which were technologically limited to simple ocular observations or photographs [39]. Following the scanpath theory by Noton and Stark [127], the interest in the particular fixation sequence has vastly increased [6] and nowadays is an extensive interdisciplinary research field including psychophysics, cognitive neuroscience, and computer science [39]. Yet, up to now, it has not been possible to predict the sequence of fixations, neither temporally nor spatially, of an observer looking at an arbitrary scene [21, 50, 155]. However, two major categories of factors that drive attention have been made out and are widely accepted. These are *bottom-up factors* and *top-down factors*. Bottom-up cues are derived from the visual scene itself and are often referred to as stimulus driven. So-called *salient* regions attract the attention of the observer through a sufficiently discriminative feature with respect to surrounding features [50]. Such features can consist of characteristics like motion, color, size, shape, texture, specific patterns or a combination of those. Bottom-up attention is fast and involuntary [22]. Top-down factors on the other hand consist of cognitive factors like knowledge, expectations, reward, and current goals [22]. This implicates that the intentional information gathering for the task at hand heavily influences the gaze behavior and the actual sequence of fixations. Examples for this behavior are intuitive and numerous. When turning left, a driver will probably exhibit a different fixation sequence and will scan different road users compared to when turning right. One of the most famous works showing that eye movements depend on the current task is of Yarbus [197]. The participants' eye movements differed considerably depending on the given task when looking at the same scene. The scanpath theory by Noton and Stark [127] explains the process of vision by stating that an internal cognitive model of the scene controls the visual perception as well as the fixation and saccade sequence [140]. Even without any task at hand, subjects tend to fixate "informative details" [39]. Top-down visual attention is comparatively slow, task-driven and voluntary [22].

Task-driven gaze behavior can be and is in many cases learned [58]. For simple laboratory tasks, human gaze behavior strategies are, under certain contraints like a minimal fixation duration, close to optimal to solve these tasks [68]. But also in complex tasks, observers must have learned which regions are important. E. g. in walking, people learn how often to look at other pedestrians in order to avoid collisions [58]. Other regions than the attended one are often largely ignored, shown e. g. in experiments for change blindness [165] or intentional blindness [164]. In the latter, participants of the study did not notice the person in a gorilla costume walking through a scene when counting ball passes. In many studies, "free viewing" is applied as an attempt to isolate task-free visual processes. However, as pointed out in [179], subjects are probably choosing their own internal priorities, which are unknown to the interviewer. According to [155] and [58], the question for future research is to understand how exactly the combination of different factors and tasks influence the gaze target selection process. The insights obtained in this section automatically lead to the question of how these concepts transfer to the special case of driving a car.

Why do we look where we do while driving? Seven "qualitative laws" of natural gaze behavior in the wild have been summarized by Lappi [97] from the vast amount of research on gaze behavior indicated in the section above. The term "natural gaze behavior in the wild" refers to everyday tasks outside of "sterile" laboratory settings which became possible to investigate by the advent of wearable and remote video based eye-tracking. Each "law" is formulated quite openly to cover the many observations. The subsequent work from Lappi et al. [98] summarizes these rules for the specific task of driving and exemplarily presents the occurrence of those in one driving sequence of an expert driver. An overview of their work is presented in Table. 2.1.

The first two rules are also the most important ones in the context of this work. The first rule states that, given task constraints, gaze behavior shares characteristics within and between individuals. This underlines the influence of the task itself on the gaze behavior. Not only in driving, it has been shown that single fixations are mainly focused on those targets that are most relevant to the task [98]. This does not only include specific objects but also regions of interest defined by other factors such as maneuver planning or knowledge on traffic rules [180]. E. g., a large, red, and fast approaching truck, which should strongly attract visual attention according

Rule	Comment	Driving example
1. Gaze patterns are repeatable and stereotypical	Stereotypy is found within and between individuals	visually securing intersections; glances at tangent point in curve driving
2. Gaze is focused on task relevant objects and locations	Corresponds to top-down control rather than bottom-up behavior	Scanning intersections, traffic, potential hazards, etc.
3. Fixations have interpretable functional roles	These roles are not always intuitive	see [98] for more information.
4. Targets are fixated "just in time"	Targets are fixated when they become relevant for the next (sub)task	The time delay from gaze to steering action is about 1-2 s
5. Visual sampling is intermittent; just in time fixations are interleaved with look-ahead fixations	"fixations to [targets] that will be relevant in a later subtask implying short-term memory" [97]	Glances at the instruments, traffic signs or wayposts
6. Memory is used to orient gaze in 3D space	Few fixations on irrelevant targets imply a mental model of where to expect task-relevant targets	Traffic signs are better noticed when placed at expected locations
7. Gaze control is embedded in head and body motion and control	Gaze shifts are supported by head and body motion; conversely, body motion is compensated by gaze stabilizing effects.	Turning the head to look at another road user

Table 2.1: Seven "qualitative laws" of gaze strategies in naturalistic tasks according to and mainly taken from [98].

to the concept of bottom-up cues, is irrelevant to the task of one's own lane-keeping as long as this truck stays on its own lane. In this case, it is unnecessary to observe the truck. Only if the truck departs from its trajectory creating a collision risk, it is highly probable that the driver guides their gaze towards the approaching hazard. Task dependency is furthermore emphasized by the observation that during driving, the largest part of the gaze falls into a small region towards the driving direction called the "eyes-on-road" region. It was reported in [180] that a simple and narrow Gaussian baseline is already a great predictor for the gaze direction in driving tasks.

Even though gaze behavior in driving is mainly task-driven, two concurring aspects are mentioned. First, as pointed out in [180], bottom-up features compete with the top-down cues such as objects that pop out behind occluding objects, e. g. a child jumping out behind a car, flashing of indicators or brake lights, which are all salient features. Second, driving often is a task which does not require full attention from the driver. This leads to what is called "visual spare capacity" by [86] meaning that drivers have time to attend to targets that are not relevant for driving. These phases can occur at any time and it can be assumed that during those, the

driver's gaze can land anywhere. This free-viewing behavior[17] is interfering and alternating with purely task-dependent factors.

2.3.2 Computational Models of Visual Attention

In contrast to psychological studies concerned with the cognitive processes that guide gaze motion patterns, computational models of visual attention try to use the observed concepts to predict where and in what order humans visually attend to in a static or dynamic scene. Again, the extent of research in this field is vast and beyond the scope of this thesis. However, if one thinks about possible applications of a suitable visual attention module in ADAS, such a model could predict where the driver should be looking. This information could then be mapped to the actually measured attended regions and objects, i. e. the topic of this thesis. For this reason, some literature and survey articles on the field of visual attention are presented.

General Models As already formulated, it has not been possible to predict the precise sequence of fixations, neither temporally nor spatially, of an observer looking at an arbitrary scene [21, 155]. Rather, whole scanpaths and motion patterns are analyzed via statistical methods [21] or visualization techniques [20] and computational attention models are commonly validated against those measured eye movements of human observers [22]. Extensive survey papers of attentional models giving a good overview on the topic are [22, 50].

Prediction of Driver's Gaze Newly emerging models for visual attention in specific domains such as driving extend the feature based idea by Itti et al. [72] of detecting salient regions. Instead of using handcrafted features that are believed to attract gaze, these features are learned in the form of deep neural networks from large datasets [31, 134, 135, 180, 181, 196]. The characteristics of saliency are then encoded in the model parameters of the neural network and can thus not only capture bottom-up cues but also top-down cues from the training dataset. The motivation can be twofold. First, the goal is to predict where the driver should look in order to be able to compare this result with where the driver actually looked. And second, the knowledge of where a driver normally pays attention to opens

[17] As stated before, the "task" is decided by the person itself, e. g. looking at an advertising board.

up new possibilities for the vehicle's vision system on where to pay more detailed attention on.

In contrast to this prediction of gaze locations, models using reinforcement learning have been developed to predict temporal aspects of human gaze in natural situations such as driving [80, 174] or walking [168, 169]. It is argued that humans actively gather relevant information in their environment depending on the current task or sub-task and guide their gaze accordingly in order to achieve goals. This top-down strategy can be formalized in reward functions extended by uncertainty terms. Comparisons to recorded human gaze behavior support these approaches [173, 187].

2.4 Summary

This chapter presented basic aspects of human eye physiology and vision as well as perception and attention. In combination, computational models to each subtopic were outlined. The take home message from this chapter has three aspects, one from each section. Firstly, gaze trackers need a personal calibration procedure due to physiological differences between human eyes. Secondly, the simplest categorization of gaze motion is the division into fixations, together with smooth pursuits, and saccades. During fixations and smooth pursuits, the image on the retina within the fovea is kept stable and the information from the visual scene in this region can be cognitively processed. Thirdly, visual attention is governed by bottom-up (scene-driven) and top-down (task-driven) processes. Strategies of the latter can be and are often learned. Gaze behavior in driving is mainly governed by top-down processes and fixations focus on targets relevant to the task. However, predicting the precise sequence of fixations is impossible. The main algorithmic contribution in Chapter 5 makes use of some of the presented concepts and incorporates them into the general approach and the tracking model assumptions.

3 Bayesian Filtering

This chapter firstly provides the theoretical foundations of Bayesian filters for dynamic systems as a special case of Bayesian inference, a well-founded theory for reasoning under uncertainty. Secondly, it gives an overview over practical algorithms starting with the well-known and widely-used Kalman filter followed by extensions to approximate the Bayesian filter equations in case of nonlinear, non-Gaussian or multiple model assumptions. Within the latter, explanations are mostly tailored to the tools necessary for the design of the gaze target tracking filter in Chapter 5. All major derivations and explanations in this chapter can be found in at least one of the extensive text books from Bar-Shalom [14], Bishop [18], Murphy [124] or Barber [15]. The book from Thrun et al. [186] furthermore provides an extensive overview on applications of Bayesian inference in the domain of robotics. A well written, thorough and comprehensive introduction to Bayesian inference for linear and nonlinear filtering tasks with numerous background information can also be found in [153]. The derivations and explanations in this chapter are closely leaned on these mentioned sources.

3.1 Optimal Bayesian Filter

The key concept of probabilistic filtering is that of *belief*, i. e. the internal knowledge about the system's state is given by a *probability density function* (pdf). It is the goal of the Bayes filter to constantly update the belief on the system state when new measurement data is arriving. In order to process such sequentially arriving data, the Bayes filter performs two essential steps. These are called *prediction* and *measurement update*. Derivations of the probabilistic formulation of the optimal Bayesian filter equation can be found in [18, 151, 153, 186].

Starting from the joint pdf of the sequence of all observed measurements $\boldsymbol{z}^{1:k} = \{\boldsymbol{z}^1, ..., \boldsymbol{z}^k\}$ and hidden, i. e. unobserved, states $\boldsymbol{x}^{0:k} = \{\boldsymbol{x}^0, ..., \boldsymbol{x}^k\}$, the belief of the state sequence can be expressed as probability conditioned

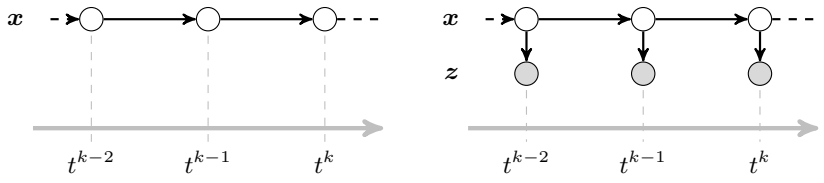
on the observed measurements sequence

$$p(\boldsymbol{x}^{0:k}|\boldsymbol{z}^{1:k}) = \frac{p(\boldsymbol{x}^{0:k}, \boldsymbol{z}^{1:k})}{p(\boldsymbol{z}^{1:k})}. \tag{3.1}$$

In Bayesian optimal filtering, however, not the full joint posterior is of interest but the marginal posterior state density $p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k})$ which describes the belief of the current system state given all observed measurements up to the current sequence step k. In order to design computationally feasible filters, two common assumptions on stochastic independence are made. First, measurements only depend on the system state at the same time and the system state itself only depends on the previous system state. This is known as first order *Hidden Markov Model* (HMM) [15].

Hidden Markov Model Just like any other joint pdf, the pdf of a dynamic system can be factorized in conditional probabilities according to a graphical model representation [18]. Directed edges represent the dependencies of random variables which are denoted by the nodes. Observed, or measured, variables are often drawn as shaded or gray nodes, while unobserved, or hidden, variables are presented by white nodes. The graphical representation of a joint pdf corresponds to one specific factorization into conditional probabilities. In the case of dynamic systems, the conditional independence assumptions are visualized in so-called *Dynamic Bayesian Networks* (DBN).

A hidden Markov model describes a Markov chain on hidden variables [15]. The sequence of hidden system states $\boldsymbol{x}^{0:k}$ forms the Markov chain that obeys the Markov assumption. This assumption in its simplest form states that past and future data are independent if the current state $\boldsymbol{x}^k$ is known [186]. In terms of conditional dependence, this means that the current state $\boldsymbol{x}^k$ is independent of all older states $\boldsymbol{x}^{0:k-2}$ if the previous state $\boldsymbol{x}^{k-1}$ is given. The corresponding graphical model of a first order Markov chain is shown in Fig. 3.1a. The intuition behind this modeling is that the influence of the immediate past is more relevant than of the remote past [15], or, that the current state contains all past effects. Extending the Markov chain, the hidden Markov model is formed by the *emission* $p(\boldsymbol{z}^k|\boldsymbol{x}^k)$. The current observation $\boldsymbol{z}^k$ of the system state $\boldsymbol{x}^k$ is independent from other measurements and previous system states depending only on the current system state. In filtering tasks of dynamic systems, the observation is formed by the noisy measurement of the system state or of a measurable derivation of the state. Fig. 3.1b shows such a hidden Markov model.



(a) Markov Chain (b) HMM

Figure 3.1: First order Markov chain and Hidden Markov Model. The Markov chain fulfills the Markov assumption and the hidden model emerges from the Markov chain through the observations of the state variable.

Recursive Filter Equation Now given these properties of an hidden Markov model of the dynamic system, the numerator on the right side of equation (3.1) can be factorized according to the stated independencies and the product rule of probabilities

$$p(\boldsymbol{x}^{0:k}|\boldsymbol{z}^{1:k}) = \frac{p(\boldsymbol{x}^{0:k-1}, \boldsymbol{z}^{1:k-1})p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1})p(\boldsymbol{z}^k|\boldsymbol{x}^k)}{p(\boldsymbol{z}^{1:k})}. \tag{3.2}$$

By marginalizing over all previous states $\boldsymbol{x}^{0:k-1}$ one obtains using the sum rule of probabilities

$$p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k}) = \frac{p(\boldsymbol{z}^k|\boldsymbol{x}^k)}{p(\boldsymbol{z}^{1:k})} \int_{\boldsymbol{x}^{k-1}} \cdots \int_{\boldsymbol{x}^0} p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1})p(\boldsymbol{x}^{0:k-1}, \boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{0:k-1}. \tag{3.3}$$

This term describes already the belief of the current system state $\boldsymbol{x}^k$ conditioned on all observed measurements. The multiple integral over the old system states can be further written as

$$\begin{aligned}
&\int_{\boldsymbol{x}^{k-1}} \cdots \int_{\boldsymbol{x}^0} p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1})p(\boldsymbol{x}^{0:k-1}, \boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{0:k-1} \\
&\quad = \int_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1}) \int_{\boldsymbol{x}^{k-2}} \cdots \int_{\boldsymbol{x}^0} p(\boldsymbol{x}^{0:k-1}, \boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{0:k-2}d\boldsymbol{x}^{k-1} \\
&\quad = \int_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1})p(\boldsymbol{x}^{k-1}, \boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{k-1}.
\end{aligned}$$

Since the measurements are assumed to occur independently from each other over time, $p(\boldsymbol{z}^{1:k})$ decomposes into $p(\boldsymbol{z}^{1:k}) = \prod_{i=1}^{k} p(\boldsymbol{z}^{i})$. At the same time, the joint probability of the old state and all measurements up to time $k-1$ under the integral can be factorized by applying again the product rule of probabilities to

$$p(\boldsymbol{x}^{k-1}, \boldsymbol{z}^{1:k-1}) = p(\boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})p(\boldsymbol{z}^{1:k-1}). \tag{3.4}$$

By making use of these reformulations, (3.3) now yields

$$p(\boldsymbol{x}^{k}|\boldsymbol{z}^{1:k}) = \frac{p(\boldsymbol{z}^{k}|\boldsymbol{x}^{k})}{p(\boldsymbol{z}^{k})p(\boldsymbol{z}^{1:k-1})} \int\limits_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^{k}|\boldsymbol{x}^{k-1})p(\boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})p(\boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{k-1}. \tag{3.5}$$

After canceling out the term $p(\boldsymbol{z}^{1:k-1})$, the powerful optimal Bayesian filter equation

$$\overbrace{\underbrace{p(\boldsymbol{x}^{k}|\boldsymbol{z}^{1:k})}_{\text{Bel}(\boldsymbol{x}^{k})}}^{\text{Posterior}} = \frac{\overbrace{p(\boldsymbol{z}^{k}|\boldsymbol{x}^{k})}^{\text{Emission}}}{\underbrace{p(\boldsymbol{z}^{k})}_{\text{Normalization}}} \overbrace{\int\limits_{\boldsymbol{x}^{k-1}} \underbrace{p(\boldsymbol{x}^{k}|\boldsymbol{x}^{k-1})}_{\text{Transition}} \underbrace{p(\boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})}_{\text{Bel}(\boldsymbol{x}^{k-1})} d\boldsymbol{x}^{k-1}}^{\text{Prior}} \underbrace{\qquad\qquad}_{\text{Prediction } p(\boldsymbol{x}^{k}|\boldsymbol{z}^{1:k-1})} \tag{3.6}$$

is obtained. It is this equation that all algorithms concerned with probabilistic filtering try to solve or approximate. As can be seen in (3.6), the current state estimate, or belief, only depends on the previous state estimate, i. e. the old belief, the transition and the measurement, also called emission or likelihood. Given those, only an initial state pdf is needed to start the recursion. The prediction is also called *prior* or *a-priori probability*, because it describes the state probability *before* the observation. In the same way, the belief is also called *posterior* or *a-posteriori probability* because it is the pdf after the observation has been made [18]. Since $p(\boldsymbol{z}^{k})$ is simply a scalar factor normalizing the term on the right side to a true pdf that integrates to 1, equation (3.6) can be written more practically as

$$p(\boldsymbol{x}^{k}|\boldsymbol{z}^{1:k}) = \eta\ p(\boldsymbol{z}^{k}|\boldsymbol{x}^{k}) \int\limits_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^{k}|\boldsymbol{x}^{k-1})p(\boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{k-1}, \tag{3.7a}$$

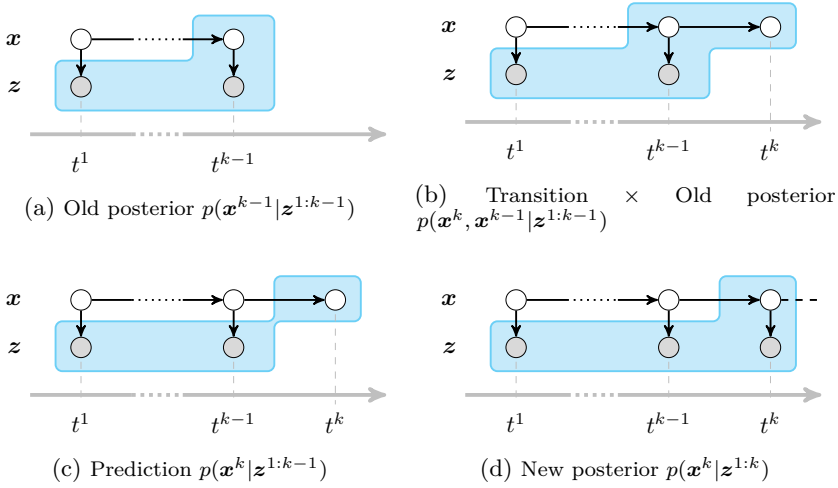
$$p(\boldsymbol{x}^{k}|\boldsymbol{z}^{1:k}) \sim p(\boldsymbol{z}^{k}|\boldsymbol{x}^{k}) \int\limits_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^{k}|\boldsymbol{x}^{k-1})p(\boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{k-1}, \tag{3.7b}$$

where η is a normalization constant. As soon as the right hand side of (3.7b) is evaluated, the normalization factor can often be calculated and it is therefore omitted whenever possible. From (3.6), the two essential steps, the prediction and measurement update, also called correction, become visible. By multiplying the old belief with the transition probability and marginalizing out the old system states, the prediction $p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k-1})$ is obtained. The measurement update then incorporates the newly observed knowledge by multiplying the predicted belief with the likelihood[18] $p(\boldsymbol{z}^k|\boldsymbol{x}^k)$ that the observation $\boldsymbol{z}^k$ may have been observed. This recursive procedure is visualized in Fig. 3.2. In order to start the recursion, only formulations of the transition $p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1})$ and the measurement likelihood $p(\boldsymbol{z}^k|\boldsymbol{x}^k)$ are necessary besides an assumption for the initial state pdf $p(\boldsymbol{x}^0)$. Based on this concept of Bayesian filtering, practical filter algorithms are introduced and their characteristics discussed in the subsequent sections, starting with the widely used Kalman Filter.

3.2 Kalman Filter

Gaussian filters are the earliest and most widely used tractable realizations of the Bayesian filter who share the basic idea that beliefs are modeled as multivariate Gaussian distributions [186]. The *Kalman Filter* (KF), originally published in [82], is one of the most commonly used techniques for filtering and prediction in *Linear Dynamic Systems* (LDS) disturbed by additive Gaussian noise, which are also called *Linear Gaussian Systems* (LGS). The wide spread of the Kalman filter in various technical domains for several decades has led to a multitude of tutorials and derivations of the filter equations. Here, the probabilistic derivation based on the previously presented optimal Bayesian filter shall be presented as it lines up well with the explanations on further extensions of the approach. It can be found in chapter 13.3 of the book of Bishop [18] with the necessary equations on marginal and conditional probabilities of Gaussians in Chapter 2.3 of [18] that are also listed here in the following. The derivation presented here is closely leaned on the internal documents of [37] and [190], which themselves refer to [18].

[18] The term *likelihood* for $p(\boldsymbol{z}|\boldsymbol{x})$ is only used if viewed as a function of $\boldsymbol{x}$ for observed, and thus fixed, $\boldsymbol{z}$. Its integral does not (necessarily) integrate to 1 which makes it not a valid pdf over $\boldsymbol{x}$. In Chapter 5, the likelihood is explicitly denoted as $\ell(\cdot)$ to emphasize this difference.



(a) Old posterior $p(\boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})$

(b) Transition × Old posterior $p(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})$

(c) Prediction $p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k-1})$

(d) New posterior $p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k})$

Figure 3.2: HMM with the single steps of the Bayes filter step. The nodes enclosed by the blue areas denote the variables of the respective distribution with the respective "most recent" system state and measurements. Starting from the old posterior (a), the transition to the new state (b) is computed. Marginalization yields the prediction (c) and incorporating the new measurement the new posterior (d).

Marginal and Conditional Gaussians – Toolkit The only equations on marginal and conditional probabilities needed for the derivation of the Kalman filter are the following. Let a marginal Gaussian distribution for $\boldsymbol{x}$ and a conditional Gaussian distribution for $\boldsymbol{y}$ given $\boldsymbol{x}$ be given by

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x), \tag{3.8a}$$
$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{C}\boldsymbol{x} + \boldsymbol{c}, \boldsymbol{\Sigma}_{y|x}), \tag{3.8b}$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal, or Gaussian, distribution of $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is assumed to have a mean that is a linear function of $\boldsymbol{x}$ given by the matrix $\boldsymbol{C}$ and the vector $\boldsymbol{c}$ and a covariance $\boldsymbol{\Sigma}_{y|x}$ which is independent of $\boldsymbol{x}$ [18]. Then, the marginal distribution of $\boldsymbol{y}$ and the conditional distribution of $\boldsymbol{x}$ given

$\boldsymbol{y}$ are obtained by

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \mathcal{N}(\boldsymbol{y}|\boldsymbol{C}\boldsymbol{\mu} + \boldsymbol{c}, \boldsymbol{\Sigma}_{y|x} + \boldsymbol{C}\boldsymbol{\Sigma}_x\boldsymbol{C}^\top), \tag{3.9a}$$

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \tag{3.9b}$$

with

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}(\boldsymbol{C}^\top\boldsymbol{\Sigma}_{y|x}^{-1}(\boldsymbol{y} - \boldsymbol{c}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}), \tag{3.9c}$$

$$\boldsymbol{\Sigma}_{x|y} = (\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{C}^\top\boldsymbol{\Sigma}_{y|x}^{-1}\boldsymbol{C})^{-1}. \tag{3.9d}$$

Stochastic State Space System The Kalman filter provides an exact filtering solution for linear Gaussian systems, i. e. linear, but possibly time-variant, discrete time stochastic state space systems of the form

$$\boldsymbol{x}^{k+1} = \boldsymbol{A}^k\boldsymbol{x}^k + \boldsymbol{B}^k\boldsymbol{u}^k + \boldsymbol{w}^k, \qquad \boldsymbol{w}^k \sim \mathcal{N}(0, \boldsymbol{Q}^k) \tag{3.10a}$$

$$\boldsymbol{z}^k = \boldsymbol{C}^k\boldsymbol{x}^k + \boldsymbol{v}^k, \qquad \boldsymbol{v}^k \sim \mathcal{N}(0, \boldsymbol{R}^k) \tag{3.10b}$$

with state vector $\boldsymbol{x}^k \in \mathbb{R}^n$, input vector $\boldsymbol{u}^k \in \mathbb{R}^p$, measurement vector $\boldsymbol{z}^k \in \mathbb{R}^q$, process noise vector $\boldsymbol{w}^k \in \mathbb{R}^n$ and measurement noise vector $\boldsymbol{v}^k \in \mathbb{R}^q$. $\boldsymbol{A}^k \in \mathbb{R}^{n\times n}$ is the system matrix, $\boldsymbol{B}^k \in \mathbb{R}^{n\times p}$ is the input matrix and $\boldsymbol{C}^k \in \mathbb{R}^{q\times n}$ is the measurement matrix. The noise sequences are assumed to be independent, zero-mean, white, Gaussian with process noise covariance $\boldsymbol{Q}^k \in \mathbb{R}^{n\times n}$ and measurement noise covariance $\boldsymbol{R}^k \in \mathbb{R}^{q\times q}$. For better overview, the time indices of the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{R}$ are left aside during the derivation of the filter equations. Nevertheless, the matrices can be either time variant or time invariant. If they are independent of time, the system is *stationary*.

Given these prerequisites and that the initial state distribution is also Gaussian [153], the transition and emission probabilities (cf. equation (3.6)) follow normal distributions in the form of

$$p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1}, \boldsymbol{u}^{k-1}) = \mathcal{N}(\boldsymbol{x}^k|\boldsymbol{A}\boldsymbol{x}^{k-1} + \boldsymbol{B}\boldsymbol{u}^{k-1}, \boldsymbol{Q}), \tag{3.11}$$

$$p(\boldsymbol{z}^k|\boldsymbol{x}^k) = \mathcal{N}(\boldsymbol{z}^k|\boldsymbol{C}\boldsymbol{x}^k, \boldsymbol{R}). \tag{3.12}$$

The initial state distribution is often based on an initial measurement and can be stated in the form of

$$p(\boldsymbol{x}^1|\boldsymbol{z}^1) = \mathcal{N}(\boldsymbol{x}^1|\boldsymbol{\mu}^1, \boldsymbol{P}^1) \tag{3.13}$$

with expectation $\boldsymbol{\mu}^1$ and covariance matrix $\boldsymbol{P}^1$ in order to start the recursion[19]. As it is a variant of the optimal Bayesian filter, the Kalman filter can be comprehensibly formulated by a prediction and correction step which can be derived from the equations (3.11)-(3.13) using the equations (3.8a)-(3.9b).

Prediction Step The prediction step computes the probability distribution of the current state $\boldsymbol{x}^k$ given all measurements $\boldsymbol{z}^{1:k-1}$ up to the previous time step $k-1$. The predicted belief $p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k-1})$ can be obtained from equation (3.9a) by interpreting (3.13) as the marginal probability (3.8a) and (3.11) as the conditional probability with known dependency (3.8b). This yields

$$p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k-1}) = \int_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^k|\boldsymbol{x}^{k-1})p(\boldsymbol{x}^{k-1}|\boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{k-1} \tag{3.14a}$$

$$= \int_{\boldsymbol{x}^{k-1}} \mathcal{N}(\boldsymbol{x}^k|\boldsymbol{A}\boldsymbol{x}^{k-1} + \boldsymbol{B}\boldsymbol{u}^{k-1}, \boldsymbol{Q})\mathcal{N}(\boldsymbol{x}^{k-1}|\boldsymbol{\mu}^{k-1}, \boldsymbol{P}^{k-1})d\boldsymbol{x}^{k-1} \tag{3.14b}$$

$$= \mathcal{N}(\boldsymbol{x}^k|\underbrace{\boldsymbol{A}\boldsymbol{\mu}^{k-1} + \boldsymbol{B}\boldsymbol{u}^{k-1}}_{\check{\boldsymbol{\mu}}^k}, \underbrace{\boldsymbol{A}\boldsymbol{P}^{k-1}\boldsymbol{A}^\top + \boldsymbol{Q}}_{\check{\boldsymbol{P}}^k}). \tag{3.14c}$$

Thus, the predicted belief, or prior, is itself of Gaussian shape with the predicted mean $\check{\boldsymbol{\mu}}^k$ and covariance $\check{\boldsymbol{P}}^k$.

Measurement Update Step By inserting the result of the prediction step (3.14c) and the emission (3.12) into the Bayes filter equation (3.6), the correction step can be performed in a similar way. By interpreting (3.14c) as (3.8a) (again, the marginal probability) and (3.12) as (3.8b) (the conditional probability), the posterior follows from (3.9b)

$$p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k}) \sim p(\boldsymbol{z}^k|\boldsymbol{x}^k)p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k-1}) \tag{3.15a}$$

$$\sim \mathcal{N}(\boldsymbol{z}^k|\boldsymbol{C}\boldsymbol{x}^k, \boldsymbol{R})\mathcal{N}(\boldsymbol{x}^k|\check{\boldsymbol{\mu}}^k, \check{\boldsymbol{P}}^k) \tag{3.15b}$$

$$\Rightarrow p(\boldsymbol{x}^k|\boldsymbol{z}^{1:k}) = \mathcal{N}(\boldsymbol{x}^k|\boldsymbol{\mu}^k, \boldsymbol{P}^k) \tag{3.15c}$$

[19] Sometimes, the starting point is defined as $p(\boldsymbol{x}^0) = \mathcal{N}(\boldsymbol{x}^0|\boldsymbol{\mu}^0, \boldsymbol{P}^0)$, which is, however, only a question of mathematical definition without further relevance for the filtering task.

with

$$\boldsymbol{\mu}^k = \boldsymbol{P}^k(\boldsymbol{C}^\top \boldsymbol{R}^{-1}\boldsymbol{z}^k + (\check{\boldsymbol{P}}^k)^{-1}\check{\boldsymbol{\mu}}^k), \tag{3.16a}$$

$$\boldsymbol{P}^k = ((\check{\boldsymbol{P}}^k)^{-1} + \boldsymbol{C}^\top \boldsymbol{R}^{-1}\boldsymbol{C})^{-1}. \tag{3.16b}$$

Mean $\boldsymbol{\mu}^k$ and covariance $\boldsymbol{P}^k$ of the filter estimate are at the same time starting point for the next recursion. For completeness, also the normalization factor is given here. It follows from (3.9a) as

$$p(\boldsymbol{z}^k) = \mathcal{N}(\boldsymbol{z}^k | \boldsymbol{C}\check{\boldsymbol{\mu}}^k, \boldsymbol{C}\check{\boldsymbol{P}}^k\boldsymbol{C}^\top + \boldsymbol{R}). \tag{3.17}$$

Since the gaze target tracking module in Chapter 5 cannot be formulated in closed form, the reformulation of the above result into the common Kalman filter equations is negligible but provided here for completeness. Using the Woodbury matrix identity [136], the state covariance $\boldsymbol{P}^k$ can be reformulated as

$$\boldsymbol{P}^k = ((\check{\boldsymbol{P}}^k)^{-1} + \boldsymbol{C}^\top \boldsymbol{R}^{-1}\boldsymbol{C})^{-1} \tag{3.18a}$$

$$= \check{\boldsymbol{P}}^k - \underbrace{\check{\boldsymbol{P}}^k\boldsymbol{C}^\top(\boldsymbol{C}\check{\boldsymbol{P}}^k\boldsymbol{C}^\top + \boldsymbol{R})^{-1}}_{\boldsymbol{K}^k}\boldsymbol{C}\check{\boldsymbol{P}}^k \tag{3.18b}$$

$$= (\boldsymbol{I} - \boldsymbol{K}^k\boldsymbol{C})\check{\boldsymbol{P}}^k, \tag{3.18c}$$

using the commonly known Kalman gain matrix $\boldsymbol{K}^k$. The state expectation $\boldsymbol{\mu}^k$ can be similarly re-arranged yielding

$$\boldsymbol{\mu}^k = \check{\boldsymbol{\mu}}^k + \boldsymbol{K}^k(\boldsymbol{z}^k - \boldsymbol{C}\check{\boldsymbol{\mu}}^k). \tag{3.19}$$

This last equation for the state estimate shows how the Kalman filter trades off between model prediction and observed measurement based on the assumed noise covariance matrices which are normally subject of model parameter choice. Increased measurement noise, reflected by "large" $\boldsymbol{R}$ leads to a "small" Kalman gain $\boldsymbol{K}^k$ and thus increased confidence in the model prediction. Conversely, better measurements and higher process noise or variance of the previous state result in a "larger" Kalman gain and thus more weight on the measurement. The clear advantage of the Kalman filter is that because all probabilities are purely Gaussian, the filter is analytically solvable and all parameters can be computed in closed form for any time step. On the other side, its main disadvantage is its restriction to linear Gaussian systems. In many of today's applications, this prerequisite is not fulfilled so that plain Kalman filter are only applicable to the most trivial problems which is why extensions and variations of the Kalman filter have emerged over the last decades. In the following sections, only a few are discussed.

3.3 Approximations of the Optimal Bayesian Filter

3.3.1 Nonlinear Transition and Emission

In many applications, the transition and measurement equations can be modeled in form of a nonlinear discrete time stochastic state space system of the form

$$\boldsymbol{x}^{k+1} = \boldsymbol{f}(\boldsymbol{x}^k, \boldsymbol{u}^k) + \boldsymbol{w}^k, \qquad \boldsymbol{w}^k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^k) \tag{3.20a}$$

$$\boldsymbol{z}^k = \boldsymbol{h}(\boldsymbol{x}^k) + \boldsymbol{v}^k, \qquad \boldsymbol{v}^k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}^k) \tag{3.20b}$$

with state vector $\boldsymbol{x}^k \in \mathbb{R}^n$, input vector $\boldsymbol{u}^k \in \mathbb{R}^p$, measurement vector $\boldsymbol{z}^k \in \mathbb{R}^q$, process noise vector $\boldsymbol{w}^k \in \mathbb{R}^n$ and measurement noise vector $\boldsymbol{v}^k \in \mathbb{R}^q$. $\boldsymbol{f}(\boldsymbol{x}^k, \boldsymbol{u}^k)$ denotes the system function and $\boldsymbol{h}(\boldsymbol{x}^k)$ is the measurement function. Just like the linear Gaussian system, the nonlinear system can be time variant. The noise sequences are again assumed to be independent, zero-mean, white, Gaussian with process noise covariance $\boldsymbol{Q}^k \in \mathbb{R}^{n \times n}$ and measurement noise covariance $\boldsymbol{R}^k \in \mathbb{R}^{q \times q}$.

Due to the nonlinearities in the system and measurement function, the posterior distribution becomes distorted. Usually, this makes it impossible to perform the recursion step exactly [186] and the filter update can no longer be evaluated in closed form. For the state estimate of such nonlinear stochastic systems, mainly two procedures have been established, namely the *Extended Kalman Filter* (EKF) [9, 14] and the *Unscented Kalman Filter* (UKF) orginally presented in [81] and extended by [191]. Since both algorithms belong to the state of the art of nonlinear filtering tasks for many years now, discussions of the advantages and disadvantages can be found in many textbooks such as [124, 163, 186]. A good introduction to both can also be found in [192].

Extended Kalman Filter

The extended Kalman filter uses the same representation of the posterior belief as the Kalman filter, namely a Gaussian, but the difference is that this posterior is only approximate in contrast to the Kalman filter where it is exact as presented in the previous section. The key idea of the extended Kalman filter is to approximate the system and measurement functions through their first order Taylor series expansion resulting in linear approximations that are tangent in the operating point. For the

system function, the operating point is the mean of the last posterior $\boldsymbol{\mu}^{k-1}$ and the control input $\boldsymbol{u}^{k-1}$ while for the measurement function the mean of the current prediction $\check{\boldsymbol{\mu}}^{k}$ is used. Based on the Jacobian matrices of the two functions, the prediction and correction step of the extended Kalman filter are analogous to the procedure of the standard Kalman filter. Even though the extended Kalman filter provides a rather simple solution to the extension to nonlinear systems, it suffers from several drawbacks. First, with large nonlinearities or uncertainties, the filter solution becomes less accurate [124]. Second, the extended Kalman filter can diverge if the linearization is done at false operating points [14]. These can not only result from unfavorably chosen initial conditions, but can also arise from the lower accuracy. And third, the computation of the Jacobians can be error prone or even impossible if $\boldsymbol{f}$ or $\boldsymbol{h}$ is non-differentiable at the operating point [153]. The unscented Kalman filter described in the following has proven to often perform better.

Unscented Kalman Filter

The reason why the unscented Kalman filter performs better in many situations is that it suffers less in case of large nonlinearities or uncertainties. This is because the nonlinearities of system and measurement function are captured better. Instead of linearizing the functions, a set of so-called Sigma-points is chosen from the initial Gaussian distribution, which are then propagated through the nonlinear functions. A Gaussian approximation of the posterior is then retrieved from the nonlinearly propagated Sigma-points to start the next filter iteration. The idea behind this is that "it is easier to approximate a Gaussian distribution than it is to approximate an arbitrary nonlinear function" [81]. Hereby, the unscented Kalman filter captures mean and covariance of the posterior to the third order Taylor series approximation for any nonlinearity for Gaussian inputs and at least to the second order for non-Gaussian inputs [192]. The standard extended Kalman filter on the other hand provides only first order approximation. Beside its superior performance, the unscented Kalman filter has further advantages. Its computation does not require explicit calculations of derivatives such as Jacobians and its overall computational complexity is of the same order as of the extended Kalman filter. Furthermore, the additive white Gaussian noise case is just a special case of the basic filter equations so that the unscented Kalman filter generally is also able to handle non-additive non-Gaussian noise without further approximations [192].

Extended and unscented Kalman filter are mentioned at this point since

they are relatively simple approaches to realize the Bayesian filter task for nonlinear stochastic state space systems which are encountered in many practical problems, e. g. in the domain of robotics. As stated, the main problem is that the predict-update-cycle can not be computed exactly in closed form due to the nonlinearities of the system. In contrast, the next section introduces another approximate inference approach for the case when the recursion can be performed exactly but the resulting prior or posterior is of inconvenient form to start the next recursion step.

3.3.2 Assumed Density Filter

According to [121], the *Assumed Density Filter* (ADF) has been independently proposed in statistics, artificial intelligence and control. Here, it is referred to the latter, originally introduced in [118] from where also the name "assumed density filtering" originates. The idea is as follows. The posterior of the previous step is assumed to be of a certain density, i. e. a suitably restricted family of probability distributions, e. g. Gaussian. Furthermore, the one-step update is assumed to be tractable and analytically determinable [124]. However, the resulting posterior is no longer representable with the assumed density form, i. e. it is no longer part of the family of tractable distributions [124]. To overcome this problem, after updating, the best tractable approximation is sought by projecting the result back onto the space of the assumed density family. Assuming that the resulting "exact" posterior is $p(\boldsymbol{x})$ and the goal is to find the best approximation $q(\boldsymbol{x}|\boldsymbol{\theta}^\star)$ from the family of assumed densities parameterized by $\boldsymbol{\theta}$, this can be achieved by minimizing the Kullback-Leibler (KL) Divergence [93]

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \mathrm{KL}(p(\boldsymbol{x})||q(\boldsymbol{x}|\boldsymbol{\theta})) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x}|\boldsymbol{\theta})}\right) d\boldsymbol{x} \quad (3.21)$$

with respect to the parameters $\boldsymbol{\theta}$ of the approximating distribution $q(\boldsymbol{x}|\boldsymbol{\theta})$. Once $q(\boldsymbol{x}|\boldsymbol{\theta}^\star)$ is determined, it marks the starting point for the new recursion step of the filter. The common *predict* $\rightarrow$ *update* cycle is thus extended by a *project*-step. It can be shown that this minimization of the KL-Divergence is equivalent to the maximum likelihood estimation (MLE) of the parameters of $q(\boldsymbol{x}|\boldsymbol{\theta})$ [124] and furthermore that in case that q is in the exponential family, the parameters are given by *moment matching* [18, 121, 124, 186]. Simply speaking, in the case of q being a Gaussian, mean and covariance of q are just set equal to mean and covariance of p.

Formally, this can be expressed as

$$\mathrm{E}_q[\boldsymbol{x}] = \mathrm{E}_p[\boldsymbol{x}], \tag{3.22a}$$

$$\mathrm{E}_q[\boldsymbol{x}^T\boldsymbol{x}] = \mathrm{E}_p[\boldsymbol{x}^T\boldsymbol{x}], \tag{3.22b}$$

where $\mathrm{E}[\cdot]$ describes the expected value. If, for example, the posterior distribution is assumed to have Gaussian shape and the transition is given by r different linear models, the posterior distribution results in a mixture of r Gaussians

$$p(\boldsymbol{x}) = \sum_{i=1}^{r} \pi_i \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3.23}$$

with mixture weights π_i. Since the mixture of Gaussians does not belong to the family of assumed densities, i. e. is no longer Gaussian, the mixture components need to be merged to one single component. For this special case, mean and variance of the approximating distribution q are given by [15]

$$\boldsymbol{\mu}_q = \sum_{i=1}^{r} \pi_i \boldsymbol{\mu}_i, \tag{3.24a}$$

$$\boldsymbol{\Sigma}_q = \sum_{i=1}^{r} \pi_i (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top) - \boldsymbol{\mu}_q \boldsymbol{\mu}_q^\top. \tag{3.24b}$$

This example of a generalized pseudo-Bayesian estimator of first order belongs to the larger class of multiple-model filters which will be treated in more detail later.

The gaze target tracking module in Chapter 5 applies only linear transition models in combination with multi-model assumptions and a non-Gaussian likelihood. Therefore, the assumed density filter in combination with sampling is a suitable tool for the means of this work. However, arbitrary posterior densities can also be tackled with sampling methods. Thus, for completeness, an example for a possible approach that is able to handle non-parametric distributions and also nonlinear transitions is shortly outlined.

3.3.3 Particle Filter – Example of Non-Parametric Filter

For dynamical systems which are not linear, but especially not Gaussian, sampling based methods provide accurate (but not exact) alternatives

to the previously presented approaches. More precisely, *Particle Filters* (PF) belong to the larger class of non-parametric filters[20] and can thus be applied to arbitrary multi-modal posterior distributions and any kind of system dynamics. Introductions can again be found in any major textbook [15, 18, 124, 186] or in the well written tutorial article of Arulampalam [9].

The particle filter, a sequential Monte Carlo algorithm, approximates the posterior distribution by a finite number of weighted samples, called particles. The particle set is propagated through time according to the transition distribution. The new set of particles represents then an approximation of the state prediction. In the update step, the so-called *importance factors* $p(\boldsymbol{z}^k|\boldsymbol{x}_l^k)$, i. e. the likelihood of the measurement given the sample $\boldsymbol{x}_l^k$, incorporate the measurements into the particle set. After each recursion, a resampling procedure with replacement is performed which eliminates samples with lower weights adding to the robustness of the filter.

The great advantage of sequential Monte Carlo methods is that they are able to handle arbitrary distributions more accurately than parametric filters such as the unscented Kalman filter. The major drawback on the other hand is the need for large particle sets making the computation comparatively slow [124]. It is thus reasonable to work with parametric approaches whenever possible.

3.4 Multiple Model Filtering

Even though the introduced assumed density filter and particle filter are in principle able to handle nonlinear and non-Gaussian systems, so far, it has been assumed that the considered system behaves exactly according to one dynamic model. In many applications, however, the system behavior is changing over time. By that, not a potential time variance of a linear system is meant since this case can be easily handled as a special case of the Kalman filter and its extensions. Rather, the system dynamics function changes structurally due to a *switch* of the system behavior and obeys to one of r different models. An example for such a change in dynamics can be found in tracking maneuvering targets, i. e. objects or agents that change their motion behavior over time from moving straight with constant velocity to performing a coordinated turn [153]. Obviously, the state of the agent can switch from one motion model to the other and back. Often,

[20] Non-parametric filters approximate the posterior over a continuous space with a finite number of values and do not rely on a fixed functional form such as a Gaussian [186].

like in this maneuvering example, this change is abrupt compared to the time that one specific model is in place. Two problems occur in a practical application if this change in behavior is not taken into consideration in the filter design. First, the filter result is sub-optimal because the system function does not match the true system behavior. And second, data association becomes more difficult and error-prone due to the sub-optimal state estimate since the measurement potentially lies further away from the predicted state. However, if it is possible to formulate structurally different models each of which describes one of the potentially occurring system behaviors, the model in place can be estimated together with the system state. This is the objective of multiple model filters.

3.4.1 Multiple Model Optimal Bayesian Filter

The goal of the outlined problem is to estimate the joint pdf $p(\boldsymbol{x}^k, \boldsymbol{s}^k|\boldsymbol{z}^{1:k})$ of a continuous random variable $\boldsymbol{x}^k$ describing the system state, and a discrete random variable $\boldsymbol{s}^k \in \mathbb{R}^r$ being the switching variable which determines the current dynamic model in place given all observations $\boldsymbol{z}^{1:k}$ up to the current time step. The resulting hybrid system is called *jump Markov system* or *switching state space model* and the respective filter approaches are termed *multiple model* filters. In the following, the already introduced optimal Bayesian filter is extended to multiple models followed by a brief introduction of different practical implementations.

Extension of the Optimal Bayesian Filter to Switching Models

In the jump Markov system in its general form, the evolution of the switching variable is assumed to obey a first order Markov chain. Furthermore, the estimated system state at each time step depends on the model in place, leading to the graphical model shown in Fig. 3.3. The joint probability $p(\boldsymbol{x}^k, \boldsymbol{s}^k|\boldsymbol{z}^{1:k})$ of system state and switching variable is given from the Bayesian filter equation (3.6)[21]

$$p(\boldsymbol{x}^k, \boldsymbol{s}^k|\boldsymbol{z}^{1:k}) = \frac{p(\boldsymbol{z}^k|\boldsymbol{x}^k, \boldsymbol{s}^k)}{p(\boldsymbol{z}^k)} \sum_{\boldsymbol{s}^{k-1}} \int_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^k, \boldsymbol{s}^k|\boldsymbol{x}^{k-1}, \boldsymbol{s}^{k-1}) \, p(\boldsymbol{x}^{k-1}, \boldsymbol{s}^{k-1}|\boldsymbol{z}^{1:k-1}) d\boldsymbol{x}^{k-1}. \quad (3.25)$$

[21] By considering the set of the variables $\boldsymbol{x}$ and $\boldsymbol{s}$ as one joint stochastic variable, (3.25) follows straightforwardly from (3.6) by plugging in the new joint variable.

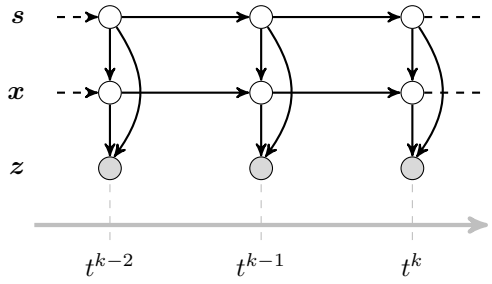


Figure 3.3: Standard version of a jump Markov system: the higher level is the switching variable, the middle level is the state estimate and the shaded variables are observations of the state.

Here, only the joint probability $p(\boldsymbol{x}^k, \boldsymbol{s}^k|\boldsymbol{z}^{1:k})$ is considered since the expectation of the state $\boldsymbol{x}$ can be obtained by marginalizing over all models, i. e. over $\boldsymbol{s}^k$. Vice versa, the probability of the switching variable, the so-called mode probability $p(\boldsymbol{s}^k|\boldsymbol{z}^{1:k})$, is obtained after marginalizing out the system state. Due to the factorization of the random variables as shown by the graphical model in Fig. 3.3, the joint pdf from (3.25) can be factorized to

$$p(\boldsymbol{x}^k, \boldsymbol{s}^k|\boldsymbol{z}^{1:k}) = \frac{p(\boldsymbol{z}^k|\boldsymbol{x}^k, \boldsymbol{s}^k)}{p(\boldsymbol{z}^k)} \sum_{\boldsymbol{s}^{k-1}} \int_{\boldsymbol{x}^{k-1}} p(\boldsymbol{s}^k|\boldsymbol{s}^{k-1}) p(\boldsymbol{x}^k|\boldsymbol{s}^k, \boldsymbol{x}^{k-1}) p(\boldsymbol{x}^{k-1}, \boldsymbol{s}^{k-1}|\boldsymbol{z}^{1:k-1}) d\boldsymbol{x}^{k-1}. \quad (3.26)$$

The switching variable $\boldsymbol{s}$ is generally formulated as r-dimensional binary random variable [18] having a 1-of-r representation in which only one s_i is equal to one and the rest of the elements are zero, i. e. $s_i \in \{0, 1\}, \sum_i^r s_i = 1$. The interpretation is that at a certain time step, only one of the system behavior models is active. The probability of model i at time step k is given in terms of the mixing coefficient $p(s_i^k = 1) = \pi_i$ and the notation $s_i^k = 1$ is abbreviated as s_i^k in the following.

The term $p(\boldsymbol{s}^k|\boldsymbol{s}^{k-1})$ describes the model transition probability and is generally a time invariant Markovian model transition probability matrix

$$\boldsymbol{\Pi} = \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & \ddots & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix} \in \mathbb{R}^{r \times r} \quad (3.27)$$

whose entries p_{ij} are design parameters. Each entry p_{ij} describes the probability $p(s_j^k|s_i^{k-1})$ that the model in place switches from model i to model j ($i \to j$). In case of $i = j$, the model stays the same. Note that index j is used here at time-step k and index i at time-step $k-1$ which will be continued in the following.

From the dynamic Bayesian network in Fig. 3.3 and the representation of $\boldsymbol{s}$ as binary random variable, the posterior at each time step can be factorized as

$$p(\boldsymbol{x}^k, \boldsymbol{s}^k|\boldsymbol{z}^{1:k}) = p(\boldsymbol{s}^k|\boldsymbol{z}^{1:k})p(\boldsymbol{x}^k|\boldsymbol{s}^k, \boldsymbol{z}^{1:k}) \tag{3.28a}$$

$$= \prod_{j=1}^{r} p(s_j^k|\boldsymbol{z}^{1:k})^{s_j} p(\boldsymbol{x}^k|s_j^k, \boldsymbol{z}^{1:k})^{s_j} \tag{3.28b}$$

$$= \prod_{j=1}^{r} p(\boldsymbol{x}^k, s_j^k|\boldsymbol{z}^{1:k})^{s_j}. \tag{3.28c}$$

Therefore, it is sufficient to consider each model j separately in the filter step. Additionally, since the switching variable $\boldsymbol{s}$ has a 1-of-r representation in which only one s_i is equal to one, the sum over the previous vector states $\sum_{\boldsymbol{s}^{k-1}}$ can be written as sum over the single elements $\sum_{i=1}^{r}$. Starting from (3.26) and using the same factorization from (3.28b) for the old posterior, this yields the one step filtering equation

$$\begin{aligned} p(\boldsymbol{x}^k, s_j^k|\boldsymbol{z}^{1:k}) = {} & \frac{p(\boldsymbol{z}^k|\boldsymbol{x}^k, s_j^k)}{p(\boldsymbol{z}^k)} \\ & \sum_{i=1}^{r} \int_{\boldsymbol{x}^{k-1}} p(s_j^k|s_i^{k-1}) p(\boldsymbol{x}^k|s_j^k, \boldsymbol{x}^{k-1}) \\ & \prod_{i=1}^{r} p(s_i^{k-1}|\boldsymbol{z}^{1:k-1})^{s_i} p(\boldsymbol{x}^{k-1}|s_i^{k-1}, \boldsymbol{z}^{1:k-1})^{s_i} d\boldsymbol{x}^{k-1}. \end{aligned} \tag{3.29}$$

Since the product and the sum both run over i, the above equation can be further simplified to

$$\begin{aligned} p(\boldsymbol{x}^k, s_j^k|\boldsymbol{z}^{1:k}) = {} & \frac{p(\boldsymbol{z}^k|\boldsymbol{x}^k, s_j^k)}{p(\boldsymbol{z}^k)} \\ & \sum_{i=1}^{r} \int_{\boldsymbol{x}^{k-1}} p(s_j^k|s_i^{k-1}) p(\boldsymbol{x}^k|s_j^k, \boldsymbol{x}^{k-1}) p(s_i^{k-1}|\boldsymbol{z}^{1:k-1}) \\ & p(\boldsymbol{x}^{k-1}|s_i^{k-1}, \boldsymbol{z}^{1:k-1}) d\boldsymbol{x}^{k-1}. \end{aligned} \tag{3.30}$$

Rearranging terms in (3.30) yields

$$p(\boldsymbol{x}^k, s_j^k|\boldsymbol{z}^{1:k}) = \sum_{i=1}^{r} \left(\frac{p(\boldsymbol{z}^k|\boldsymbol{x}^k, s_j^k) \int_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^k|s_j^k, \boldsymbol{x}^{k-1}) p(\boldsymbol{x}^{k-1}|s_i^{k-1}, \boldsymbol{z}^{1:k-1}) d\boldsymbol{x}^{k-1}}{1} \times \frac{p(s_j^k|s_i^{k-1}) p(s_i^{k-1}|\boldsymbol{z}^{1:k-1})}{p(\boldsymbol{z}^k)} \right) \tag{3.31}$$

and expanding with the model likelihood $\Lambda_{ij}^k = p(\boldsymbol{z}^k|s_j^k, s_i^{k-1})$ leads to

$$p(\boldsymbol{x}^k, s_j^k|\boldsymbol{z}^{1:k}) = \sum_{i=1}^{r} \underbrace{\overbrace{\frac{p(\boldsymbol{z}^k|\boldsymbol{x}^k, s_j^k) p(\boldsymbol{x}^k|s_j^k, s_i^{k-1}, \boldsymbol{z}^{1:k-1})}{p(\boldsymbol{z}^k|s_j^k, s_i^{k-1})}}^{\text{mode matched filtering / state inference stage}}}_{p(\boldsymbol{x}^k|s_j^k, s_i^{k-1}, \boldsymbol{z}^{1:k})} \underbrace{\overbrace{\frac{p(\boldsymbol{z}^k|s_j^k, s_i^{k-1}) p(s_j^k, s_i^{k-1}|\boldsymbol{z}^{1:k-1})}{p(\boldsymbol{z}^k)}}^{\text{mode learning stage}}}_{p(s_j^k, s_i^{k-1}|\boldsymbol{z}^{1:k})}. \tag{3.32}$$

As can be seen in comparison to (3.6), the first fraction corresponds to one specific filter given the previous model i and the current model j. This is known as *mode matched filtering*. In case of a static multiple model filter, switches between models are inhibited, i. e. jumps from one model s_i to another mode s_j are not considered and $\mathbf{\Pi}$ is equal to the identity matrix. This case corresponds to one separate filter for each model and the equation (3.32) stays tractable[22]. This is since the sum over previous models vanishes as $p(s_j^k|s_i^{k-1})$ is equal to zero if $i \neq j$. Otherwise, as can be seen by the sum over all previous models in (3.32), each model transition $i \to j$ at time step k induces one probability mixture component resulting in r terms for one single starting component i or, when regarded the other way round, a sum of r terms for a given model j. In the subsequent step, each of these r terms corresponds again to one starting component resulting in an exponentially increasing number of mixture components for the complete mode sequence [14, 162]. This is exemplarily shown in Fig. 3.4 for the case of two models.

[22] However, this static estimator is restricted since the correct model must be among the formulated hypotheses. Otherwise the mode probability of the filter converges to the "best" model regardless of how well the hypothesis fits the system behavior.

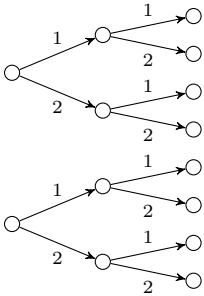


Figure 3.4: Intractability of a mutliple model filter. At every step, each model induces one component letting the number of components grow exponentially. Here, only two models are illustrated. Figure according to [124].

Murphy [124] summarizes three different general approaches for the jump Markov system, all involving approximate inference, i. e. they approximate the optimal filter recursion equation (3.32). The first approach consists of cutting off branches with low probability of the discrete tree in Fig. 3.4. This procedure is also called *pruning* [186]. The second consists of applying Monte-Carlo estimation, previously introduced in the frame of particle filters in the last section. The third approach is using an assumed density filter which is also used in the later presented gaze target tracking algorithm. When using assumed density filters, the exponential growth of mixture components is inhibited by projecting the posterior pdf back to a smaller mixture density after each time step.

3.4.2 Approximating the Multiple Model with ADF

For convenience in the explanation, let's assume that the posterior belief is given as weighted combination of individual normal distributions, i. e. is modeled by a mixture of Gaussians[23] (one component per discrete state)

$$p(\boldsymbol{x}^k, \boldsymbol{s}^k | \boldsymbol{z}^{1:k}) = p(\boldsymbol{s}^k | \boldsymbol{z}^{1:k}) p(\boldsymbol{x}^k | \boldsymbol{s}^k, \boldsymbol{z}^{1:k}) \tag{3.33a}$$

$$= \prod_{i=j}^{r} \underbrace{(\pi_j^k)}_{p(s_j^k, \boldsymbol{z}^{1:k})}^{s_j} \underbrace{\mathcal{N}(\boldsymbol{x}^k | \boldsymbol{\mu}_j^k, \boldsymbol{P}_j^k)}_{p(\boldsymbol{x}^k | s_j^k, \boldsymbol{z}^{1:k})}^{s_j}. \tag{3.33b}$$

[23] Since the marginalized state probability $p(\boldsymbol{x}^k | \boldsymbol{z}^{1:k}) = \sum_{i=1}^{r} p(\boldsymbol{x}^k, \boldsymbol{s}^k | \boldsymbol{z}^{1:k}) = \sum_{i=1}^{r} \pi_i^k \mathcal{N}(\boldsymbol{x}^k | \boldsymbol{\mu}_i^k, \boldsymbol{\Sigma}_i^k)$ is a sum of Gaussians, this version of multiple model filters is also known as Gaussian sum filter [167].

Let's further assume that each model j obeys the linear Gaussian system[24] assumption, so that all necessary filter components can be directly obtained from the Kalman filter equations, thus

$$p(\boldsymbol{x}^k|s_j^k, \boldsymbol{x}^{k-1}) = \mathcal{N}(\boldsymbol{x}^k|\boldsymbol{A}_j\boldsymbol{x}^{k-1}, \boldsymbol{Q}_j) \tag{3.34}$$

$$p(\boldsymbol{z}^k|\boldsymbol{x}^k, s_j^k) = \mathcal{N}(\boldsymbol{z}^k|\boldsymbol{C}_j\boldsymbol{x}^k, \boldsymbol{R}_j). \tag{3.35}$$

The mode matched prediction step is then performed as

$$\int_{\boldsymbol{x}^{k-1}} p(\boldsymbol{x}^k|s_j^k, \boldsymbol{x}^{k-1})p(\boldsymbol{x}^{k-1}|s_i^{k-1}, \boldsymbol{z}^{1:k-1})d\boldsymbol{x}^{k-1}$$

$$= p(\boldsymbol{x}^k|s_j^k, s_i^{k-1}, \boldsymbol{z}^{1:k-1}) \tag{3.36a}$$

$$= \mathcal{N}(\boldsymbol{x}^k|\check{\boldsymbol{\mu}}_{ij}^k, \check{\boldsymbol{P}}_{ij}^k) \tag{3.36b}$$

and the normalizer $p(\boldsymbol{z}^k|s_j^k, s_i^{k-1})$, which is at the same time the mode likelihood function, is given by

$$p(\boldsymbol{z}^k|s_j^k, s_i^{k-1}) = \mathcal{N}(\boldsymbol{z}^k|\boldsymbol{C}_j\check{\boldsymbol{\mu}}_{ij}^k, \boldsymbol{C}_j\check{\boldsymbol{P}}_{ij}^k\boldsymbol{C}_j^\top + \boldsymbol{R}_j) = \Lambda_{ij}^k \tag{3.37}$$

yielding the *mode-conditioned* state probability according to (3.15c) - (3.16b)

$$p(\boldsymbol{x}^k|s_j^k, s_i^{k-1}, \boldsymbol{z}^{1:k}) = \mathcal{N}(\boldsymbol{x}^k|\boldsymbol{\mu}_{ij}^k, \boldsymbol{P}_{ij}^k). \tag{3.38}$$

The second fraction of (3.32) is directly obtained as

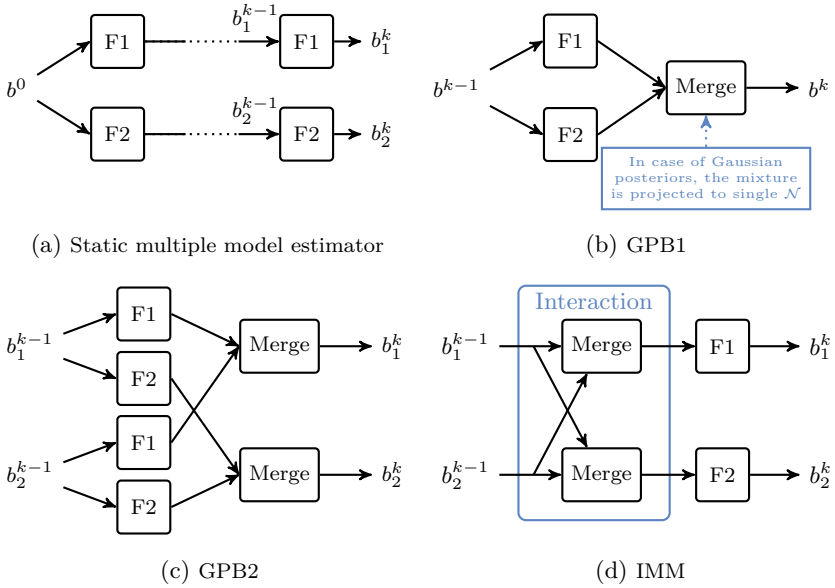
$$p(s_j^k, s_i^{k-1}|\boldsymbol{z}^{1:k}) = \pi_{ij}^k = \frac{\Lambda_{ij}^k p_{ij}\pi_i^{k-1}}{p(\boldsymbol{z}^k)} \tag{3.39}$$

with

$$p(\boldsymbol{z}^k) = \sum_{j=1}^{r}\sum_{i=1}^{r}\Lambda_{ij}^k p_{ij}\pi_i^{k-1}, \qquad j = 1, ..., r. \tag{3.40}$$

Taking this as starting point, depending on the simplifying assumptions, different practical multiple model algorithms are obtained including the *generalized pseudo-Bayesian estimators of first and second order*. A schematic comparison over some common multiple model algorithms is presented in Fig. 3.5.

[24] This assumption is only for simplicity to show the basic functionality. The multiple model approaches can of course be combined with other system behaviors. In case of non linear system behavior, the Kalman filter matched to each mode can be replaced by extended Kalman filter/unscented Kalman filter.



(a) Static multiple model estimator

(b) GPB1

(c) GPB2

(d) IMM

Figure 3.5: Block diagrams of different multiple model filter algorithms. Each F1/F2 denotes one mode matched filter. Posterior beliefs are abbreviated with b. The static filter inhibits transitions between models and can only handle them with suitable regularization. The GPB1 needs r filters whereas the GPB2 applies r^2 filters. The IMM is an intermediate approach applying only r filters with comparable results to the GPB2. Figures 3.5c)-3.5d) according to [124].

GPB1

The *first order generalized pseudo-Bayesian estimator* (GPB1) approximates the old posterior as single Gaussian distribution which is propagated through j filter models. After each cycle, the resulting r Gaussians are collapsed to one single hypothesis to start the next recursion as visualized in Fig. 3.5b. The first order generalized pseudo-Bayesian estimator thus requires r filters running in parallel using the previously combined estimate. For the mode conditioned state estimate, this implicates that all dependencies on s_i^{k-1} vanish and thus

$$p(\boldsymbol{x}^k|s_j^k, \boldsymbol{z}^{1:k}) = \mathcal{N}(\boldsymbol{x}^k|\boldsymbol{\mu}_j^k, \boldsymbol{P}_j^k). \tag{3.41}$$

However, the previous modes are used to compute the new mode probabilities according to

$$\pi_j^k = \frac{1}{c}\Lambda_j^k \sum_{i=1}^{r} p_{ij}\pi_i^{k-1}, \qquad j = 1, ..., r \tag{3.42}$$

where

$$c = \sum_{j=1}^{r} \Lambda_j^k \sum_{i=1}^{r} p_{ij}\pi_i^{k-1} \tag{3.43}$$

is the normalization constant. The combined state and covariance for the next iteration are then given by inserting $\boldsymbol{\mu}_j^k, \boldsymbol{P}_j^k$ and π_j^k into (3.24a) and (3.24b).

GPB II

In contrast to the first order generalized pseudo-Bayesian estimator, the *second order generalized pseudo-Bayesian estimator*(GPB2) only approximates $p(\boldsymbol{x}^{k-1}|s_i^{k-1}, \boldsymbol{z}^{1:k-1})$ by a Gaussian and thus the posterior of each time step is represented by a mixture of Gaussians. As stated above in (3.32), each model transition $i \to j$ induces one mixture component resulting in a sum of r terms for a given model j. The idea of the second order generalized pseudo-Bayesian estimator is to project all components resulting from different previous model behaviors i but the same current model behavior j onto one mixing component, i. e. onto one Gaussian. This results again in a total of r terms in the subsequent filtering step. Thus, a total of r^2 filters are needed to compute all histories given the previous and current model as visualized in Fig. 3.5c.

For the case of a Gaussian mixture posterior combined with linear Gaussian system dynamics, all equations for the one-step filter update are already given by the Kalman filter equations. The merging probabilities π_{ij}^k are first normalized for each j

$$\pi_{ij}^k = \frac{\Lambda_{ij}^k p_{ij}\pi_i^{k-1}}{\sum_{i=1}^{r} \Lambda_{ij}^k p_{ij}\pi_i^{k-1}} \tag{3.44}$$

before they are used to collapse the r different Gaussians to one component. The state estimate and covariance belonging to s_j^k are again given by inserting $\boldsymbol{\mu}_{ij}^k, \boldsymbol{P}_{ij}^k$ and π_{ij}^k into (3.24a) and (3.24b) yielding $\boldsymbol{\mu}_j^k$ and $\boldsymbol{P}_j^k$.

The updated mode probabilities π_j^k are given by first merging π_{ij}^k with subsequent normalization so that $\sum_i^r \pi_j^k$ sums up again to one

$$\pi_j^k = \frac{\sum_{i=1}^r \pi_{ij}}{\sum_{j=1}^r \sum_{i=1}^r \pi_{ij}} = \frac{1}{c} \sum_{i=1}^r \Lambda_{ij}^k p_{ij} \pi_i^{k-1}, \qquad j = 1, ..., r \tag{3.45}$$

where c is the normalization constant

$$c = \sum_{j=1}^r \sum_{i=1}^r \Lambda_{ij}^k p_{ij} \pi_i^{k-1}. \tag{3.46}$$

Finally, but only for output purposes, a combined state estimate and covariance can be computed by repeated use of (3.24a) and (3.24b). By inserting $\boldsymbol{\mu}_j^k, \boldsymbol{P}_j^k$ and π_j^k, one obtains the combined estimates $\boldsymbol{\mu}^k$ and $\boldsymbol{P}^k$.

Due to the r^2 filters needed in each time step, the second order generalized pseudo-Bayesian estimator is comparably more demanding than the first order generalized pseudo-Bayesian estimator but it generally leads to better results. An often preferred intermediate method for multi-model tracking, e. g. of maneuvering targets, is the *Interacting Multiple Model Filter (IMM)* shortly mentioned in the following which is similarly effective as the second order generalized pseudo-Bayesian estimator at the computational cost of the first order generalized pseudo-Bayesian estimator. Later in this work, however, the multi-model approach is used to track more than one gaze target hypothesis modeled by multiple modes. In that case, each mixture component of the posterior is directly linked to one object. Starting from each of these hypotheses separately, multiple gaze motion models are considered making the second order generalized pseudo-Bayesian estimator the preferable method for the gaze target tracking. Nevertheless, the interacting multiple model filter is shortly mentioned as it is often explained together with the two previous approaches and for example for single object tracking with multiple models it would be the preferred choice.

IMM

Just like the second order generalized pseudo-Bayesian estimator filter, the interacting multiple model filter uses a Gaussian mixture to represent the posterior belief at each time step. However, it uses only r filters making the interacting multiple model filter comparable in computational cost to the first order generalized pseudo-Bayesian estimator. The advantage, however, lies in the different inputs compared to the first order generalized

pseudo-Bayesian estimator. Instead of merging different models after the filtering step, the merging is done before in the so-called *interaction step*. First, mixing coefficients are computed for each model j and all previous models i that are combined into one Gaussian via moment matching [14]. Then each of the resulting mixing components is filtered with one filter as visualized in Fig. 3.5d. Strictly speaking, the interacting multiple model filter does not belong to the class of assumed density filter but presents another heuristic approach to the special class of multiple model filters.

3.5 Summary

This chapter gave a brief overview on optimal Bayesian filters together with common implementations. For linear Gaussian systems, the well-known Kalman filter is the adequate choice. However, in most applications, the prerequisites of linear transition and measurement as well as Gaussian additive noise are often not fulfilled. Many different filter methods for specific classes of sub-problems have been developed over the last decades and only a handful were presented in this chapter. For nonlinear Gaussian systems, extended Kalman filter and unscented Kalman filter present common solutions while particle filters are able to handle any kind of distribution and transition functions. For multiple-model filtering, the GPB1 and GPB2 represent different variations of the assumed density filter. Depending on the problem at hand, the choice of filter algorithm is important and often a combination of different approaches leads to a successful filter method. For example, [153] combines an interacting multiple model filter and unscented Kalman filter for maneuvering target motion tracking with nonlinear Gaussian motion models. The multi-hypothesis tracking approach presented in [186] combines an extended Kalman filter with the strategy of pruning. The model presented in Chapter 5 has multiple linear transition models and a non-gaussian measurement model. Since multiple hypotheses are tracked, a second order generalized pseudo-Bayesian estimator filter comprised of a Kalman filter-like prediction, a sampling-based correction, and an additional project-step is implemented. Thus, it is possible to combine the different approaches from the large field of Bayesian filtering and find a suitable solution for one's problem at hand.

4 Looking In and Looking Out

The preceding Chapters 2 and 3 introduced the theoretical foundations for this thesis. The present chapter now introduces the broader scope of driver monitoring in which the contributions of this thesis are to be classified. Over the last decade, there has been a growing research effort to monitor the driver in order to understand their actions and state and to further improve driver assistance systems. Only by understanding the driver, their true need for assistance in the driving task becomes evident. As stated at the beginning, driver monitoring will be one of the topics of the future which is not only shown by the increasing amount of research publications shown in Fig. 4.1. It is also reflected by the intention of the Euro NCAP 2025 roadmap to incorporate driver monitoring within the passenger safety testing scheme by 2024 [43]. This proposition will probably lead to a large coverage of vehicles equipped with driver monitoring systems. Such systems can target many different objectives which are addressed in the following Section 4.1. One research focus in this area is the enhancement of the vehicle's situation interpretation by the driver's awareness in order to gain a more holistic representation of the present driving situation. Simply speaking, the vehicle is not only looking at its surrounding environment, but also at what is happening inside, i. e. in the vehicle's interior, and relates the gathered information. This way, future ADAS have the potential to adapt their warning and intervention strategies just like a good co-pilot. They also only point out potential risks when they have the feeling that the driver might be unaware of them. In Section 4.2.1, a literature review on current related work in the field of driver's gaze fusion with the environment perception of ADAS is given. Subsequently, in Section 4.2.2, challenges of the fusion process are pointed out and discussed. Many thoughts and arguments in this chapter have been used in the author's publications [158–160] and are presented here in a more general context.

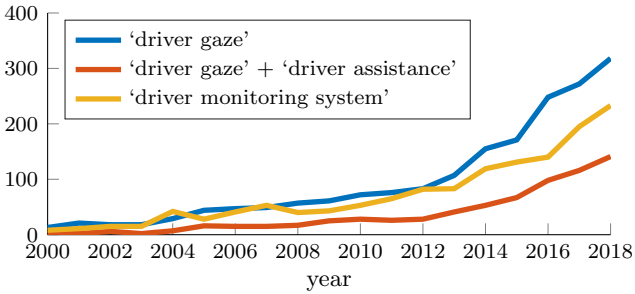


Figure 4.1: Number of publications found in the year 2019 in Google Scholar per year from 2000 until 2018 for the search terms "driver gaze", "driver gaze"+"driver assistance", and "driver monitoring system".

4.1 Facets of Driver Monitoring

The motivation for driver monitoring is manifold and mainly driven by the expectation of accident avoidance. Indeed, there are numerous imaginable applications and extensions of current ADAS, and in the future also automated driving (AD), that can benefit from a more thorough understanding of the driver and the passengers in the car. An overview on current research regarding human behavior in the traffic domain in general is surveyed in [131]. It is argued that in the age of (partially) self-driving cars, observing human behavior is not only about driver and passenger monitoring but also about humans around the vehicle such as pedestrians or inside of surround vehicles. In this work, the focus is on the ego-vehicle driver. Even in this domain, driver monitoring has many facets [94, 131] which themselves consist of a multitude of potential sub-problems. A review of the complete space of driver monitoring is beyond the scope of this thesis but the overview over different research topics within this broader domain shown in Fig. 4.2 gives an idea of where the contributions of this thesis are located. This clustering is not intended to be exhaustive, so for further information, the interested reader is referred to the given references, which, if possible, include survey papers of the respective research topic.

In this thesis, the driver's gaze is fused with the vehicle's environment perception trying to detect visual fixations on other road users and traffic relevant objects. The topic is thus located at the boundary of the areas of driver gaze direction estimation, attention estimation and environment perception, which is often associated with the concept of situation awareness.

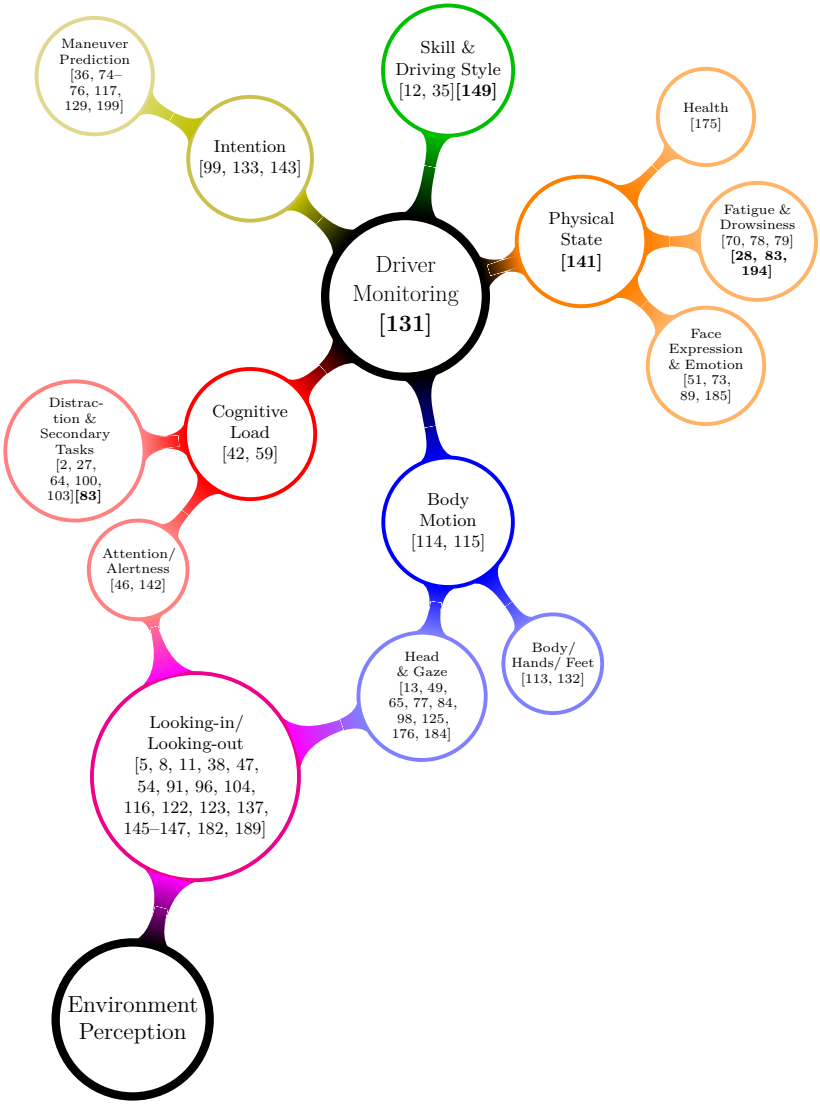


Figure 4.2: Different facets of the large area of driver monitoring including related research papers. Survey papers are set in bold and separate brackets. The contributions of this thesis are within the area of Looking-in/Looking-out (LiLo). Interdependencies between topics are not indicated and works with overlapping contributions are sorted where they fit best.

Situation awareness is defined by Endsley as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future"[41]. According to this definition, situation awareness consists of the three levels of perception, understanding and prediction of the situation [152]. While the latter two levels are difficult to measure by remote sensors, the first level of human perception can be modeled based on the observation of visual attention in relation to the situation (cf. Chapter 2). The need for the joint and combined consideration of the vehicle's interior and exterior was first formulated by Apostoloff and Zelinsky in 2002 [7] and 2004 [8] under the catch phrase *Vision In and Out of Vehicles*. The idea was taken up by Trivedi et al. [189] in 2007 as *Looking-in and Looking-out of a vehicle* (LiLo) and since then numerous publications in the area appeared. A more in-depth review is presented in the following section. Since the estimation of the driver's situation awareness always includes the vehicle's surround representation, the cluster topic of environment perception is included here in the overview as second root node. However, even though driver monitoring systems can benefit from a precise and thorough scene perception and interpretation, no explicit references are given here since the research topic itself is much larger than driver monitoring.

Besides this rather new research branch of Looking-in and Looking-out of a vehicle, obvious intentions of driver monitoring systems are fatigue or distraction detection and driver state estimation. This driver state estimate can consist of different aspects including health parameters, emotion, and body motion. But also individualized and adaptive ADAS present possible use cases for increased safety which is why also the driver's intention as well as their driving style and skills are of interest.

4.2 Fusing Driver and Situation Information

4.2.1 Literature Review

One of the future aspects of ADAS will be a comprehensive situational understanding involving both driver and environment [87]. Since the human gaze is the most important indicator of the driver's situation awareness as indicated in Fig. 4.3, driver gaze motion has been investigated for decades in driving simulators as well as real driving. As mentioned, research activities on driver gaze estimation and its matching to the surrounding environment have been growing in recent years and can be roughly classified into two

Figure 4.3: Human gaze is the most important indicator of what the driver has perceived and what they are aware of.

topics. In the first and more recent, the driver's focus of attention is extracted as region in the surround camera image and then these attention maps are used as features to build learning models that create saliency maps. Since these saliency maps are learned from real world observations of driver attention, they are called to attend like humans [134, 180, 181, 196]. The idea behind these attention models is that also cars do not need to concentrate on all objects in the sensors' view for new types of ADAS and AD and that the relevant objects can be learned from human viewing behavior. This concept has already been mentioned in Section 2.3.2. The motivation of this work, however, is within the second aspect.

The second category comprises works that extract the driver's focus of attention in the surrounding environment. E. g., in case of Level 3 of AD, the knowledge whether the driver is aware of situation-relevant objects, e. g. those with minimal time to collision or those with minimal distance, can shorten the handover time or can enhance the safety of the hand-over process by explicitly making the driver aware of them. But also Level 0 to Level 2 systems can benefit from a driver's situation awareness estimate that includes the information of seen and unseen potential traffic hazards as the driver could be warned at an early stage and lower risk level, e. g. against unseen crossing traffic participants or vehicles merging into the

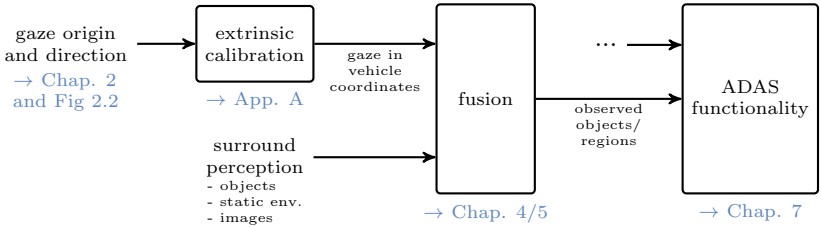


Figure 4.4: Fusion approach in this work: For mapping gaze with the surround perception, all information needs to be represented in one shared coordinate system. In this work, gaze is directly given in vehicle coordinates. The fusion block computes an estimate of the current gaze target and thus of the seen objects. It can incorporate a multitude of ideas a few of which are presented in this chapter. The subsequent chapter then presents the author's approach to the gaze/object fusion. The gaze target estimation can be used additionally to common information as input of novel ADAS functionalities.

lane. From an engineering perspective, in both cases the task is to infer the sequence of the driver's visual attention given gaze direction and exterior sensor measurements, i. e. "an online analysis of the driver's scanpath with respect to the entities that appear on the visual scene" [176]. A coarse overview on the approach in this work is presented in Fig. 4.4. Gaze is given in vehicle coordinates and can be mapped to the entities in the driving scene.

In the following, it is focused on applications which use information of the outside world and integrate gaze or at least head motion features trying to answer the question where and/or at what specifically the driver is looking during vehicle operation or what they are aware of. The general goal is to develop estimation strategies which are more precise than a simple "eyes-on-road" estimator. For this, it is commonly assumed that the gaze direction corresponds to the driver's visual attention and that the 3D point of regard, i.e. the intersection of the line of sight with the scene, is attentively perceived by the driver. In case of determining also gaze target instances, this is also referred to as *object of fixation detection* [116] or *gaze target tracking* [158].

Works on looking-in and looking-out are performed in many different ways with different setups and configurations, often custom made for specific purposes. Nevertheless, all of them need some kind of visual driver monitoring and some kind of scene perception. Current research effort can

mainly be characterized by three different criteria, first the research goal, i. e. what to achieve, second the data fusion space and third the applied methods. A few different approaches to Looking-in and Looking-out of a vehicle which are also discussed in the following are depicted in Fig. 4.5.

Research Target

Distinguishing by the research target, [8] and [183] estimate the attentive state of the driver. These works incorporate gaze either by using gaze zone estimation inside the vehicle [184], i. e. mapping the gaze to predefined areas in the vehicle cabin (e. g. left, right, road, ...), or outside the vehicle [8, 123], i. e. mapping the gaze to predefined areas relative to the vehicle coordinate system. With these approaches, however, it is not feasible to determine whether the driver has actually seen, i. e. visually fixated[25], objects around them. Approaches to estimate the awareness of objects without explicitly mapping gaze data to objects are presented by [38, 123, 147]. Interestingly, in [38], only interior gaze zone estimation is performed which could be associated to a certain exterior image region by the use of an omnidirectional camera on the vehicle's roof. In contrast, other works determine the driver's focus of attention in the surrounding environment as precise as possible either as point [91, 176, 178] or region [5, 38, 96, 159, 200], however, without explicit mapping to objects or gaze targets. The estimation of the driver's focus of attention on an instance-level, i. e. also performing a mapping to the objects in the current field of view, is done by [11, 116, 158, 182].

Data Fusion Space

The choice of sensors is often based on the preconditions or the motivation and goal of a research project. Looking inside, at the driver, or the driver's eyes is always performed with cameras. The surround perception can be achieved with common vehicle sensors such as radar, lidar, camera and of course a combination of those. Furthermore, wearable gaze tracking devices play a special role as the environment perception does not come from the vehicle but is provided by the scene camera of the eye tracking device. They are used e. g. in [5, 116, 176, 182]. One exception is [96] where the measured gaze direction is transformed into the vehicle coordinate system by means of calibration markers in the vehicle's interior visible in the device's scene

[25] "Seen" is used here in the sense of visually fixated. The derivation of actual perception is impossible without further observations of the driver's actions, e. g. evading a pedestrian or coming to a stop at a pedestrian crossing.

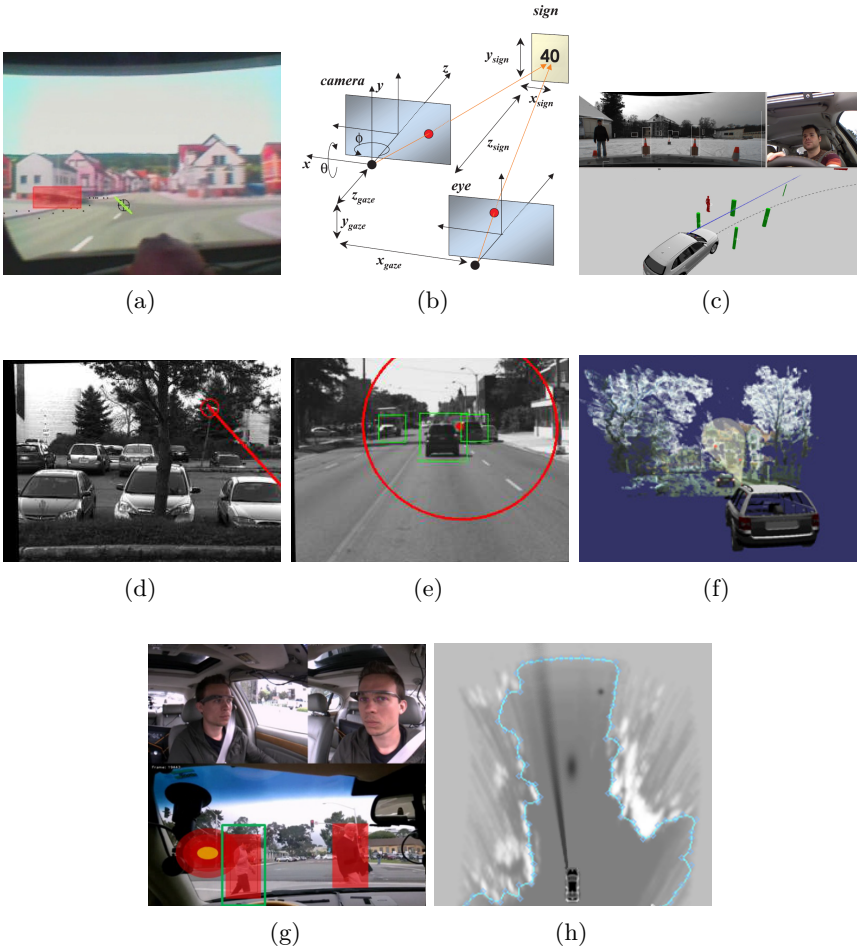


Figure 4.5: Different fusion approaches. (a) Fixation detection on objects in simulator setup [84], (b) Scene camera and eye interpreted as two-camera system [47], (c) Gaze in 3D vehicle coordinates with angular threshold [11], (d) Gaze in image space with depth from stereo [91], (e) Object detection within gaze region [200], (f) Gaze from wearable head-eye-tracking system combined with 3D stereo point cloud [96], (g) Gaze from wearable head-eye-tracking system with pedestrian detection [182], (h) Gaze in vehicle space with probabilistic modeling (author's approach) [158]. Images taken from the respective mentioned sources.

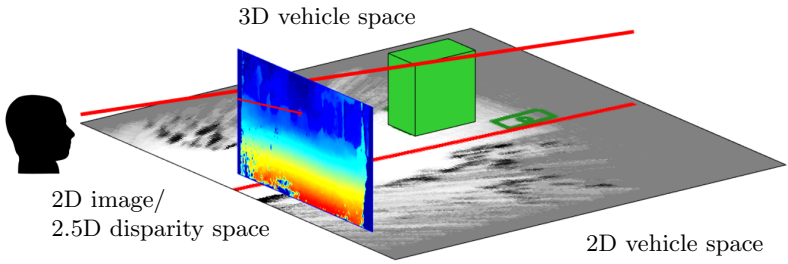


Figure 4.6: Coordinate systems for sensor fusion. Road users are schematically drawn as cubes/rectangles. Gaze is drawn as red line in each fusion space.

camera. While wearable devices in general provide higher accuracy and precision, remote systems can often be better integrated in the vehicle's sensor and network setup. Also with regards to series development of a new ADAS, remote eye-tracking systems are preferred.

Using interior and exterior measurements for gaze target estimation, either by modeling awareness of objects or explicit mapping of gaze to objects, an adequate fusion space is required. Depending on the sensor setup, there are different possibilities. Systems using 3D-lidar or stereo camera sensors can operate in the 3D vehicle coordinate system whereas 2D lidar based systems only have a 2D representation of the vehicle's surrounding. Modern automotive radar systems can to a limited degree also determine the elevation of a target which is why it's readings can be seen as 2.5D representation in vehicle coordinates. Mono camera systems are dependent on epipolar geometry and fusion in the 2D image coordinate system[26]. The 2.5D image space is used if fusion is done in image space including disparity information from a stereo camera system. All three options are present in the current literature and are depicted in Fig. 4.6.

A 3D system has the advantage of full spatial information due to a low level of abstraction. However, in the existence of noise, it is difficult to determine the point of regard robustly. In [11, 137], simple threshold based models are used to map gaze to objects in the 3D space. In [38, 123, 147, 158, 159], data is fused in the 2D surround space. In 2D, the point of

[26]Modern algorithms are able to reconstruct the complete 3D scene by using the motion between two image frames and additional information to determine the remaining degree of freedom which is the scale of the scene. These approaches are called also calles *structure from motion* (SfM) [111].

regard can be easily and more robustly determined, however, due to the lost height information, uncertainty in the result arises, e. g. the driver could be looking at the speedometer or a hanging traffic light instead of the road. The third and most common way in literature is to fuse information in the 2D or 2.5D image space by reprojecting the gaze ray or gaze region into the image [5, 91, 96, 177, 178, 183, 200] which is very intuitive and interpretable at first glance. However, normally a stereo camera is needed to obtain the scene's depth and to determine the point of regard more reliably. Also, to incorporate information from other sensors, this information must either be transformed into image coordinates, which can be very costly, or the point of regard needs to be transformed into vehicle coordinates, which can be error prone. Nevertheless, the strengths of camera sensors, i. e. the rich information in each image, can be used so that additional model assumptions can be incorporated. The gaze target can also be inferred in image space given a suitable object tracking algorithm.

Matching Approach

Finally, current work on gaze target estimation with explicit gaze/object mapping can be categorized by the methodology of the mapping process. The lack of ground truth information about awareness which was identified as one major problem in [96, 158] and which will be addressed in Chapter 6 is probably the reason why most works in literature dealing with gaze target estimation are based on rather simple assumptions of distance thresholds [11, 38, 54, 96, 122, 137]. I. e. an object or area of interest counts as seen, if the distance between target and the point of regard or the line of sight, respectively, falls below a predefined threshold. Only in [11], the matching threshold model is enhanced by a minimum fixation time and in [38] it is conceded that the driver's gaze motion cannot be reliably explained by exterior object occurrence alone. In the work of [198, 200], the threshold is computed the other way round. After estimating the area of attention, only this area is searched for objects. The thresholds in the mentioned approaches acknowledge the fact that human vision is not simply a single gaze ray. In contrast to such threshold models, in [91], the point of regard is simply modeled as point resulting from the reprojected gaze ray fused with the scene's depth image. Such simple intersection model can as well be thought of in 2D vehicle coordinates and presents a special case of the threshold model. In the subsequent section, a closer look is taken on these two approaches.

Nonetheless, the large amount of simulator and real world driving studies

have revealed some commonly observable characteristics of gaze behavior [98], cf. Chapter 2. Among those are expressible rules such that fixations fall on task-relevant objects and locations. An attempt to include such insights was taken in [182], where the gaze target is determined according to defined potential targets[27]. [147] models the interaction and awareness between driver and pedestrians probabilistically, however only coarse head direction cues are applied, thus no explicit mapping is performed. Besides these rather top-down gaze strategies, also bottom-up gaze behavior characteristics like saccadic behavior, fixation duration and saccade speeds can be incorporated. The first attempt to incorporate psychophysical cues in a dynamic filter to estimate the point of regard and infer the gaze target from it was done by the author in [159]. However, the applied grid-based approach has several limitations such as restricted virtual sensor range, high computational cost and necessary postprocessing of matching the filter result with the objects.

In the following, the different fusion spaces and basic geometric models are investigated more in detail before the proposed approach is outlined.

4.2.2 Point of Regard and Gaze Target Estimation

In Chapter 2, the basics of human gaze behavior have been introduced. According to the current understanding, visual information intake of the surrounding scene, e. g. traffic signs, pedestrians or other traffic participants, requires foveal vision, i. e. explicit fixations on the object or area of interest. In a nutshell, "this means, that to have seen a traffic hazard, the driver must have fixated it" [84]. Derived from this, the actual question for this work is how to determine visual fixations on objects. This motivation has two aspects, namely the detection of a continuous time span and the determination of the specific target in the traffic scene that the gaze falls on.

Even though perception can also occur from the peripheral view, humans tend to direct their gaze towards regions which have caught their attention. In [47], a simple experiment was performed to obtain a rule of thumb for the driver's ability of speed sign recognition. Using a desktop PC, different test subjects were asked to fix their gaze on a stationary cross on the screen and at the same time to read the number on a speed sign presented to them at different positions of the screen. In accordance with the concept of

[27] This is conceptually similar to the maximum likelihood estimate of the tracking model in the next chapter, but without a hypothesis for the case that the driver is not looking at one of the given hypotheses. That way, the gaze target is always the one closest to the gaze location regardless of the distance between them.

foveated vision, reliable discrimination ability drops off at over 4° from the fixation point supporting the modeling assumptions in the next chapter.

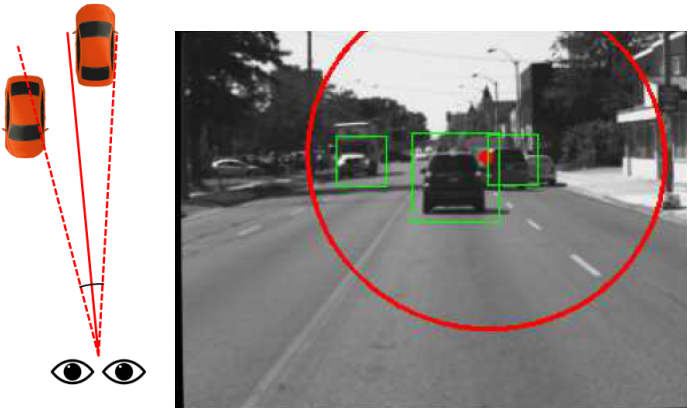
Another difficulty is given by the fact, that human perception cannot be measured. Even if some tasks can be performed with peripheral perception only, the main source of information is foveal vision. Also using eye-tracking, the only tractable quantity is foveally received information [139] together with fixation duration and from this, perception can be assumed. As just stated, perception in the peripheral field of view is possible, but also merely measurable. It is thus reasonable, to count objects in the periphery as not attentively perceived and concentrate on the relationships that actually can, to a certain extent, be observed.

Due to the desire to determine actual fixations on objects, it is important to use gaze direction measurements and not only the head direction even though head direction is more robust to compute and presents a good prior for the gaze direction. In [49], it was shown that drivers exhibit different viewing behavior which they called owl and lizard. Some test subjects exhibit stronger head motion (owl) while others merely turn their head when changing gaze direction (lizard). Thus, a precise gaze tracking module is inevitable for the proclaimed target. The sensor setup of the test vehicle of this work consisting of a multi-camera head-eye-tracking system with IR illumination combined with series sensors from the ADAS domain for surround perception presents one possibility of how to realize fusion of gaze and environment in the vehicle. All sensors are fully calibrated (cf. Appendix A) and time-synchronized. Based on this configuration, different ways of fusion are thinkable. In the following, a few challenges, advantages, and disadvantages of each are outlined.

Intersection and Threshold Models

The simplest approach to point of regard computation and gaze target estimation are basic geometric intersection and threshold models as often found in the literature. In principle, they can be realized in any coordinate system as indicated in Fig. 4.7. The former simply compute the intersection of the gaze ray with the surrounding scene resulting in an estimate of the point of regard. It is quite intuitive that with a noisy gaze or scene measurement, intersection models are not very robust as the fixated gaze target might not be hit by the gaze ray in every time step during a visual fixation. For increased robustness, threshold models extend intersection models by the assumption that human gaze can be modeled by a cone with a certain opening angle. The opening angle of the cone around the gaze

(a) Geometric intersection: the point of regard is given as the point where the gaze ray intersects the scene. If the gaze ray misses the object (dashed line), it is considered as not fixated. Example in image space taken from [91].



(b) Geometric threshold: All entities in the scene, which are within the visual attention area (dashed lines/red circle) are considered perceived. Example in image space taken from [200].

Figure 4.7: Schematic images for intersection and threshold models with examples. The measured gaze direction is drawn as red line and the icon of the two eyes depicts the location of the gaze origin, i. e. the driver's eyes.

direction vector is often set to about 10° to 13° which corresponds to the parafovea region on the cornea of the human eye[28]. For the detection of fixations in real driving setups, most models indeed use broad tolerance thresholds for the gaze measurements for increased robustness [11, 96, 137, 200]. As stated, this is mainly because remote eye tracking systems in automotive applications often do not reach the necessary precision to rely solely on the single point measurements themselves [156]. However, threshold models do not solve the problem of erroneous gaze measurements! The modeling of gaze as a cone only accounts for the fact that human vision is not a strict gaze ray but does not account for erroneous gaze direction measurements. The intention is to have a generous threshold so that an object that has been seen also counts as seen. But given a noisy gaze direction measurement, an object close to the threshold could now be counted as seen even though it was missed solely because it slipped into the gaze cone. Thus, a threshold model reduces the number of false negatives but at the cost of a potentially increased number of false positives. In contrast to intersection models, threshold models additionally come with the drawback that they do not estimate visual fixations, i. e. they do not make decisions on the current object of fixation of the driver without further ado. They simply count everything as seen that falls into the modeled gaze region with a sharp boundary. Therefore, postprocessing steps are inevitable in both cases to estimate at which object the driver might has been looking.

Example using Simple Geometric Models In order to visualize the just mentioned drawbacks of simple intersection and threshold models, they have been implemented in 2D vehicle space as depicted in Fig. 4.7. For the intersection model, the gaze target is determined as the entity which first, i. e. with the shortest distance, intersects the gaze ray and for the threshold model, all entities within the gaze cone (opening angle of 12°) equally count as fixated. Both models are tested on a short urban driving sequence shown at the top of Fig. 4.8 which is also used in the evaluation in the next chapter. The scene shows an oncoming car and a biker in front of the ego-vehicle who intends to turn left, thus having a high probability of crossing the lane. The driver is thus also expected to look at the cyclist to visually secure the situation. Furthermore, the object list contains a false detection in front of the building. The plots in Fig. 4.8

[28] The region of highest acuity is the fovea centralis with about 2° opening angle with the surrounding parafovea of approximately 10° [39], cf. Section 2.1.1.

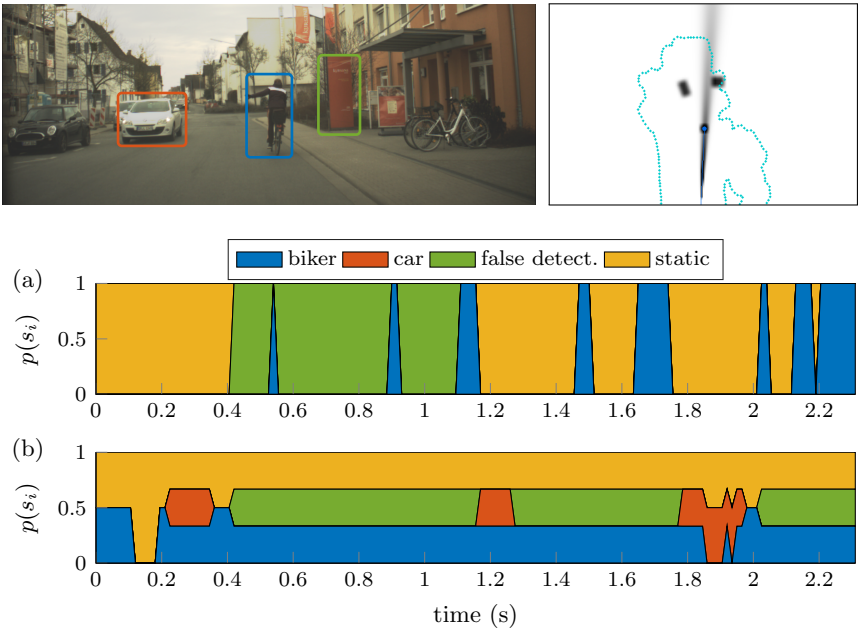


Figure 4.8: Exemplary driving scene in urban traffic shown on the top left. The birdseye view on the top right shows the vehicle's environment perception consisting of three dynamic objects, the free space boundary and the gaze measurement. Course of the estimated gaze target of a simple (a) intersection and (b) threshold model. The "static" target is given by the free space boundary.

show an area-diagram over time. The x-axis describes the time course of the situation. Based on the notation of the subsequent chapter, the y-axis denotes the probability $p(s_i)$ of the gaze targets, where one s_i denotes one potential target in the scene. At each time slice, the gaze target is the object whose color has the largest height in the plot. For the two simple geometric models, this obviously implicates that the probability of a target is either 1 for the intersection model or $\frac{1}{m}$ for the threshold model where m describes the number of objects within the gaze cone including the static environment.

As stated before, simple intersection models are not particularly robust. The model fails to detect a continuous timespan where the driver's gaze falls onto the biker in the scene. Even though the biker is already quite close to the ego-vehicle, its size is too small in combination with the

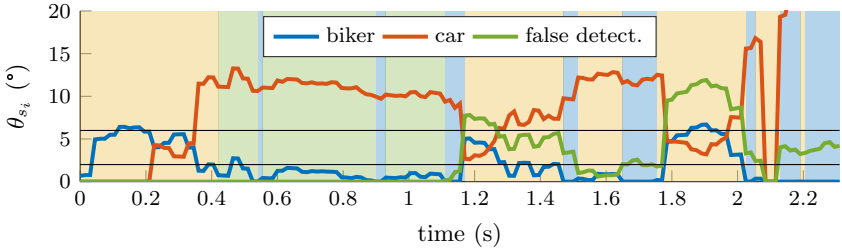


Figure 4.9: Angular deviation between the gaze heading and the direction of each object in the scene. The colors in the background are the respective gaze targets from Fig. 4.8a.

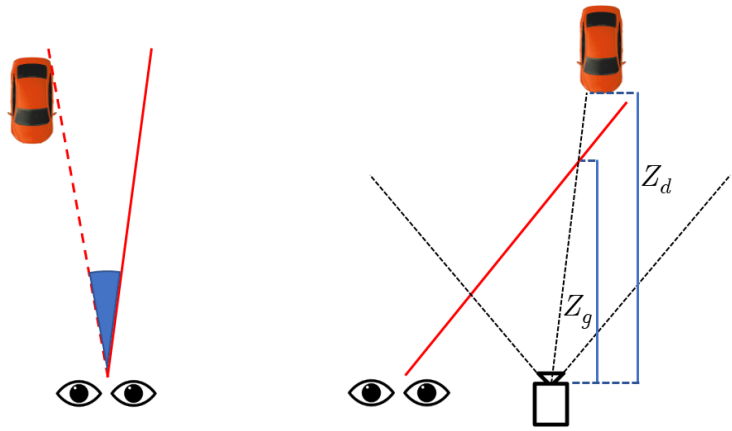
gaze measurement accuracy in order to be reliably detected. On the other hand, threshold models try to account for the parafoveal visual gaze region without determining one specific gaze target. As can be seen, often several objects fall into the modeled region of attention. This is further emphasized in Fig. 4.9 where the angular deviation of the gaze heading to the closest objects' edges are plotted. The two horizontal black lines denote 2° (approximately the area of sharp view) and 6° (angle assumed in the threshold model) deviations. While the intersection model requires the angular distance of the gaze target to be zero, the threshold model considers all potential targets whose angular distances are below the threshold, i. e. here the 6° line. Yet, taking the anatomy of the human eye as well as information intake models and measurement errors into account, fixations are expected to fall only one gaze target. At the same time, this target might not be hit by the gaze ray. This is indicated by the 2° threshold, suggesting that the biker is fixated by the driver for large parts of the given scene. For both basic geometric models, additional postprocessing steps are necessary to infer which objects the driver might have fixated and thus seen. This simple example shows that basic geometric intersection models are not yet precise enough to determine visual fixations on objects in driving scenes. Since the intersection model can be regarded as threshold model with opening angle of 0°, the parameter trades off robustness versus sensitivity in the model.

Fusion Spaces

Vehicle Space If gaze is given in vehicle coordinates, fusion in vehicle coordinates is the easiest, most straightforward and intuitive way to combine gaze and scene. The fusion can be performed in 2D as well as in 3D coordinates and the just presented intersection and threshold models are straightforward to implement (cf. Fig. 4.7). For a threshold model, the angle between the gaze ray and an object in the scene can be easily calculated. The threshold value is then directly applied to the angular deviation making the threshold much more intuitive than when fusing the data in image space as will become clear right away. This is also depicted in Fig. 4.10a. One requirement of this approach is that it necessitates a 2D or 3D representation of the surrounding scene in vehicle coordinates which normally requires radar or lidar sensors. Those provided, a reliable object list and a reasonable representation of the static scene are advantageous. However, this requires a good understanding and interpretation of the sensor readings which might not always be given.

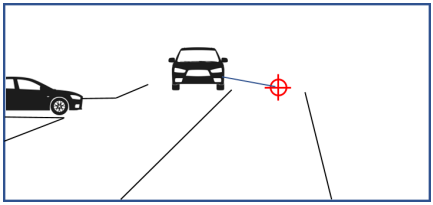
Image Space As already stated, fusion in image space is the most widely used approach in the literature. This is mainly due to the very intuitive interpretability of the result for a human observer. However, since the decision on what the driver really looked at is not straightforward, the result in image space often only consists of a single point of regard, a delimited region in the image or a simple heatmap without a statement of what the driver looks at. Thus, the result often answers the question of where the driver looks in the image but not on which object. This information, however, is not useful if it is not known what scene information is present within the detected image region. It is therefore necessary to combine the obtained image region or pixel-precise point of regard with semantic information. This information can be given by bounding boxes of target objects, e. g. road users [200], traffic signs [198] or traffic lights, or also a complete sematic segmentation of the image [134, 160]. In order to compute an estimate of which object is fixated over time, also a robust object tracking module in image space is required.

For a setup with remote head eye tracking device and a vehicle mounted camera, the precise extrinsic calibration of the sensors is necessary. This calibration provided, the gaze ray can be projected into the image of the



(a) Angular distance function in vehicle coordiante system.

(b) Distance when gaze ray is reprojected in stereo image given as difference of gaze and scene depth.



(c) Euclidean pixel distance in image. The point of regard might not always be given or needs to be computed beforehand.

Figure 4.10: Distance functions behind different approaches.

vehicle's camera by means of the reprojection equation [111]

$$\bar{\boldsymbol{x}}_g = \frac{1}{Z_g} \boldsymbol{K} \boldsymbol{\Pi}_0 \underbrace{\begin{bmatrix} \boldsymbol{R}_{cv} & \boldsymbol{t}_{cv} \\ 0 & 1 \end{bmatrix} \bar{\boldsymbol{X}}_{g,v}}_{\bar{\boldsymbol{X}}_{g,c}} . \tag{4.1}$$

In the above equation, homogeneous coordinates[29] are used to project the homogeneous representation $\bar{\boldsymbol{X}}_{g,v} = [\boldsymbol{X}_{g,v}^\top, 1]^\top \in \mathbb{R}^4$ of the 3D point

[29] The strict differentiation of vectors and matrices by means of lower case and capital letters is broken here with homogeneous coordinates. According to the usual practice in the literature of projective geometry, points of the 3D space are denoted with capital letters and of the 2D image space with lower case letters.

$\boldsymbol{X}_{g,v} = [X_{g,v}, Y_{g,v}, Z_{g,v}]^\top \in \mathbb{R}^3$ in vehicle coordinates onto the image plane. First, $\bar{\boldsymbol{X}}_{g,v}$ is transformed into camera coordinates $\bar{\boldsymbol{X}}_{g,c}$ by means of a special Euclidean transformation with rotation matrix $\boldsymbol{R}_{cv} \in \mathbb{R}^{3\times3}$ and translation vector $\boldsymbol{t}_{cv} \in \mathbb{R}^3$ describing the geometric relations between the two coordinate systems. The actual reprojection onto the image plane is then performed with the *canonical projection matrix* $\boldsymbol{\Pi}_0$ and the calibration matrix $\boldsymbol{K}$ yielding the homogeneous representation $\bar{\boldsymbol{x}}_g = [x_g, y_g, 1]^\top$ of the image point $\boldsymbol{x}_g$ in pixel. The reprojection of the complete gaze ray is obtained by sampling the gaze ray. From this reprojected gaze ray, a 2D point of regard is obtained from the minimization of the depth error between the gaze ray's depth and the scene depth at the respective pixel position

$$\boldsymbol{x}_{\mathrm{PoR}} = \underset{\boldsymbol{x}_g}{\operatorname{argmin}} |Z_g - Z_d|, \tag{4.2}$$

where Z_g is the depth of $\boldsymbol{X}_{g,c}$ in camera coordinates and Z_d is the depth of the scene at $\boldsymbol{x}_g$ obtained from the stereo disparity map as indicated in Fig. 4.10b. The corresponding 3D point of regard $\boldsymbol{X}_{\mathrm{PoR}}$ is obtained as the 3D point which was used to compute the 2D image point $\boldsymbol{x}_{\mathrm{PoR}}$. This approach corresponds to a simple geometric intersection model and has been proposed by [91]. The whole procedure is exemplarily depicted in Fig. 4.11. First, the gaze ray is reprojected and with the depth map of the stereo image, the 2D estimate $\boldsymbol{x}_{\mathrm{PoR}}$ is obtained [91]. The estimated gaze point can subsequently be combined with the semantic information at this location [134, 160]. The distance function which is minimized in this approach is quite counterintuitive. Instead of minimizing the yaw and pitch angles between gaze ray and scene, the depth in the direction of the ray connecting the camera origin and the 3D world point is minimized (cf. Fig. 4.10b). Due to the characteristics of projective geometry, the computed depth distance can be quite large when looking at a target at far distance and therefore closer parts of the scene could be favored. For this reason, it could be beneficial to use a relative error

$$\boldsymbol{x}_{\mathrm{PoR}} = \underset{\boldsymbol{x}_g}{\operatorname{argmin}} \frac{|Z_g - Z_d|}{Z_g} \tag{4.3}$$

since normalization with the depth has also proven helpful in the feature selection of visual odometry [24]. The minimization of the disparity deviation instead of the depth difference is not recommended as the disparity values are close to zero for far objects. Another drawback of this method is the bias towards the side at which the gaze ray enters the image. Due

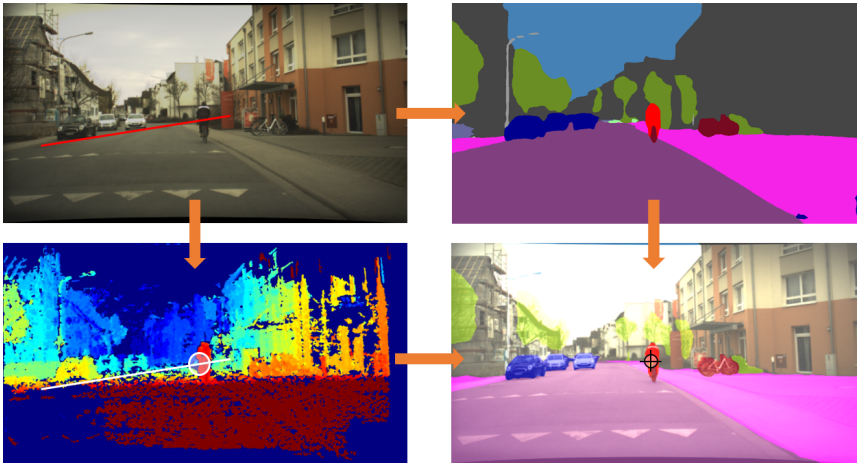


Figure 4.11: Exemplary fusion of gaze and environment in image space. From the reprojected gaze ray, the point of regard is found as the point with minimum depth distance [91]. The information of "what" the driver looks at, is given by the semantic class at the pixel position [134, 160].

to the configuration of a left steered vehicle and a camera in the middle of the windshield, the reprojected gaze ray enters the image always from the left side. And since underestimating depth leads to a larger pixel error than overestimating depth, the computed point of regard will more likely lie on the left side of the true point of regard than on the right side[30]. Even though the specific direction of the bias comes from the hardware configuration and can be altered, a bias towards one direction will remain.

In case that no precise calibration or no stereo image is available, heuristic approaches such as the definition of a common depth or just the point in infinity can be used. In [54], a remote eye-tracker is used which is normally calibrated together with a screen providing gaze coordinates on the screen. Performing a rather simple calibration with a forward facing webcam and the assumption of a "virtual screen", the gaze point is directly given in image coordinates. Provided the fixation point in the exterior camera's image, the pixel distance to predefined areas and objects of interest can be computed as indicated in Fig. 4.10c. It is argued here that for this approach, the assumed size and resolution of the virtual screen must match the viewing field and resolution of the webcam. Furthermore, it is important

[30] This characteristic can be observed in the evaluation of Chapter 6.

to place the exterior camera in line with the driver's head and as close to it as possible. Even then, some geometrical effects such as rotation or parallax are not completely considered. The resulted error measure, the euclidean pixel distance, is similarly intuitive as the angular deviation in the vehicle fusion space. Nevertheless, in order to evaluate the pixel distance, an estimate of the fixation point is necessary which might not be available for other specific sensor setups or it results from the biased method outlined above.

4.2.3 Discussion & Proposed Approach

The above description of different sensor setups, fusion spaces, and gaze-object matching methods shows that a multitude of approaches have been developed in the last two decades to combine information about the driver's gaze with the surround environment. In most cases, it is argued that the point of regard or object-of-fixation detection is feasible with the respective specific approach. However, if the intrinsic and extrinsic calibration of different sensors is done well and if the estimation of gaze and the tracking of road users is precise, the feasibility follows logically and naturally from such a high quality setup. Yet, the considerations in this chapter on fusion spaces, matching approaches, and applied sensors also suggest that certain configurations might be favorable over others. Additionally, it is important to note that every setup has its limitations and that the fusion results might differ significantly in challenging scenarios as exemplarily suggested in Fig. 4.12. Of course this question also depends on the specific goal[31]. However, no ground truth data[32] exists with which it would be possible to actually evaluate different fusion methods and to find out preferable spaces and algorithms among those used in the literature. Especially for object-of-fixation detection, many methods are not directly suitable without proper postprocessing since they only answer the question "where" the driver is looking but not "at what". This follows directly from the usage of threshold models which do not perform an estimation of visual fixations on specific objects. Intersection models would be able to do so, however, they are less robust due to limited precision of the gaze direction measurements

[31] The question whether the driver has seen an object might be well inferred in 2D/3D vehicle space but the question whether the driver has seen a traffic light might be better answered in image space.

[32] Chapter 6 is concerned with this question and at least for the specific sensor configuration of this work, a reference dataset was recorded in order to compare different methods.

Figure 4.12: Scene in urban scenario (zoomed view). There will always remain difficulties to reliably detect the driver's gaze target if it is occluded, small, far away, or highly unprobable like the biker in front of the black car.

and thus require postprocessing as well. Furthermore, almost all methods in the literature apply geometric approaches which can be considered as straightforward *measurements*.

The goal of this work is to robustly classify objects as attentively visually fixated by the driver by jointly analyzing the temporal process of the driver's gaze and the environment including spatial relations. This information can be used to enhance the observed objects by a driver awareness attribute. For this, a method in 2D vehicle space is aimed at. Fig. 4.5h shows a visualization of the approach. Within this frame, not only an angular distance function can be used which reflects human gaze behavior. The analysis in the 2D vehicle coordinate system is in many cases also sufficient since normally the gaze pitch angle is quite small when the driver's gaze falls within the eyes-on-road region. By applying a probabilistic tracking model instead of a plain geometric approach, human viewing behavior characteristics are incorporated into the model. Furthermore, in contrast to threshold models with sharp boundaries of "seen" and "unseen" regions, a softened gaze cone accounts for the foveal area of sharp vision. Most importantly, the approach proposed in the subsequent Chapter 5 directly yields a probabilistic estimate of the gaze target on object level. This not only avoids the postprocessing step of matching the gaze point to a specific object, it also directly yields the fixation duration and provides a confidence how well scene and gaze match.

4.3 Conclusion

This chapter first led from the different fields within the large area of driver monitoring and driver state estimation towards the topic of gaze target estimation. After the presentation of related work, two different main challenges of how to bring the interior and exterior world together have been discussed in more detail. First, different data fusion spaces used in the literature have been presented and their characteristics compared. Second, the most basic geometric approaches to determine the driver's gaze target and their drawbacks have been discussed using an example scene. On the basis of the considerations in this chapter, a special tracking framework that incorporates characteristics of human gaze behavior and that considers the limitations of the surround perception is introduced in the following Chapter 5. Subsequently, the newly presented algorithm is evaluated against two base line models named in this chapter, namely a simple intersection model in vehicle and in image space.

5 Gaze Target Tracking

The preceding chapter discussed several approaches and different aspects of how to estimate the point of regard and the gaze target. The present chapter now presents the author's approach to this task consisting of a probabilistic filter description. First ideas such as the gaze model or the incorporation of human gaze behavior characteristics have been published by the author in [159]. This first model, however, had some drawbacks from which the subsequent model emerged. The new filtering approach, which is described in detail in the present chapter, was published in [158]. A few refinements have been made in the inference step, which are explicitly pointed out here. The proposition of the gaze target tracking algorithm presents one main contribution of this thesis.

The chapter is structured as follows. First, a short motivation is given, which extends the previous chapter by the specific reasons for the use of a probabilistic filter. Afterwards, the Multi-Hypothesis Multi-Model filter approach is derived and explained in detail. The advantage of the new approach is qualitatively shown on a few real world sequences.

5.1 Introduction and Motivation

5.1.1 Probabilistic Description of Gaze and Environment

The lack of ground truth information about driver awareness and cognitive perception is probably the reason why most works in literature dealing with gaze target estimation are based on rather simple assumptions of distance thresholds [11, 38, 54, 96, 122, 137, 198] as outlined in the previous chapter. As was also shown, simple geometric models suffer from the basic trade-off between robustness and sensitivity. Thus, for a specific application, it is often easier to choose a robust model instead of developing a more sophisticated model whose advantages are difficult to evaluate. This might be the reason why, up to now, the actual use of behavioral rules as model knowledge for a joint description of driver's gaze behavior and surrounding

environment to extract the gaze target has been lacking in the literature. And this even though Bayesian filtering and inference has proven to improve a state estimate given noisy measurements in numerous applications. It has evolved to an accepted tool for state estimation in many domains. A possible exemplary framework for a complete joint probabilistic description of the gaze target, human gaze and the surrounding environment is motivated in Fig. 5.1. On the lower filter levels, common frameworks for each of the sensor measurements can be applied. For traffic participants, the task consists of multi-object tracking (MOT) for which numerous approaches exist [154, 166] (blue box). Also the surrounding static scene can be described and tracked in a probabilistic manner [154] (purple box). Due to the specific characteristics of gaze motion, the gaze direction (green box) is generally not filtered by a Bayesian network. Current estimation methods can be found in [57], however the measurements can still be interpreted as stochastic and can be stabilized by head pose estimates which rely on current tracking techniques [126]. The filter on the higher level marked by the orange box models the time course of the driver's gaze target via a Multi-Hypothesis Multi-Model (MHMM) approach. It marks one of the core contributions of this work and its goal is to explicitly infer the gaze target.

Multiple Models and Multiple Hypotheses

A multi-model (MM) filter tracks different models (e. g. for motion) for one object/hypothesis/agent and determines the predominant transition model based on how well the model predictions fit the sensor measurement [14]. For this application, the interacting multiple model filter introduced in Chapter 3 is often favored as it is computationally less demanding than the second order generalized pseudo-Bayesian estimator filter while being almost as accurate.

Similarly, multi-hypothesis (MH) tracking represents the distribution over all possible hypotheses and determines the weights for each hypothesis based on how well the hypothesis predictions fit the sensor measurement [48]. At first, the multi-hypothesis tracking sounds very similar to multi-model tracking[33]. However, multi-hypothesis tracking is an approach to solve the data association problem within multi-object tracking [19]. The different hypotheses are represented by the measurements which need to be assigned to the multiple tracked objects. For this, the probability of each

[33]They are similar in the sense that in both cases, the number of filter components would grow exponentially (cf. Fig. 3.4) without appropriate approximation scheme.

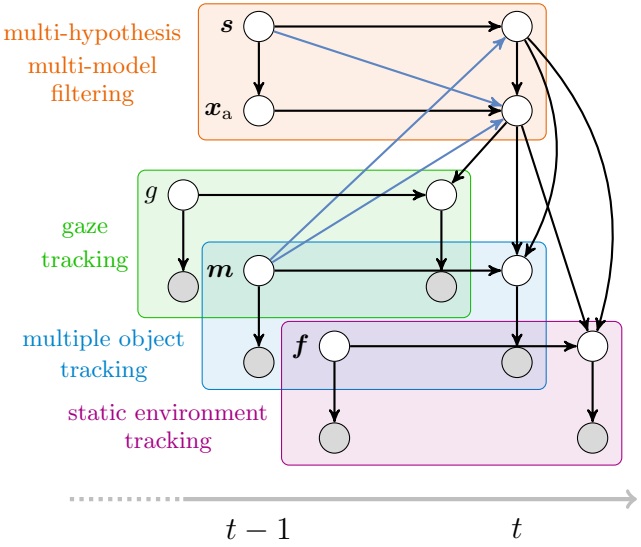


Figure 5.1: Graphical model of a complete joint probabilistic description of the gaze target, human gaze and the surrounding environment. The graphical model shows a combination of the proposed Multi-Hypothesis Multi-Model filter with gaze target switching variable $\boldsymbol{s}$ and area of attention variable $\boldsymbol{x}_\mathrm{a}$ together with stochastic representations of gaze g, multiple objects $\boldsymbol{m}$ and static scene $\boldsymbol{f}$, each represented by a hidden Markov model. The variables are explained in Section 5.2. Additional dependence assumptions introduced in this chapter are drawn in blue. Arrows to lower filter level at $t-1$ are omitted for better view.

measurement belonging to one of the objects is evaluated and hypotheses with low probabilities are discarded (*pruning*) in order to keep the filter tractable [186].

In this work, the multiple hypotheses of the gaze target are formed by the multiple objects obtained from the vehicle's object list. This means, that for one "agent", i. e. the gaze target, several hypotheses exist and their weights are evaluated through the association of the hypothesis predictions with the gaze measurement. Thus, no data association problem in the classical sense is present where it is unknown which measurement belongs to which object but it is tried to determine to which target the current gaze belongs. Since the gaze target can change over time, motion models for fixations and saccades are formulated. This combination of multiple hypotheses with multiple models can be well tackled by a second order

generalized pseudo-Bayesian estimator filter which represents the posterior as mixture model (see Section 5.2.2). Each mixture component of the posterior thus corresponds to one hypothesis for the gaze target. This representation is not possible with a first order generalized pseudo-Bayesian estimator or interacting multiple model filter approach which would either represent the current belief as one component or create a mixture state of the current location of the visual attention.

Advantages

The developed MHMM tracking has several advantages over basic geometric methods introduced in the previous chapter. Most importantly, the result of the proposed approach explicitly estimates for each object the probability of being the current gaze target. Since the filter output directly models "how much" visual attention each object gets compared to others, no postprocessing is necessary which assigns the point of regard to an entity in the scene. Continuous time spans of identical gaze targets describe fixations or smooth pursuits which serve as indicators for the driver's perception of the respective object. Also, although the algorithm is designed for the tracking of the gaze target, the filter output can also be interpreted as area of attention estimation in 2D vehicle space, the reprojection provides a robust attention estimate in image space and the course of gaze targets can be seen as rough fixation/saccade estimation. Since object list and free space can in principle be obtained from radar, lidar, stereo camera or a combination of the previous, the tracking approach remains sensor independent just like basic geometric models.

Analogy from Different Domain The idea behind this MHMM approach can be explained very vividly with the analogies to the work of [53]. There, an autonomous robot, part of a human-robot soccer team, is supposed to track the ball. There are different actions on the ball (kicking, dribbling and free motion) resulting in multiple motion models. Furthermore, the ball could be in the possession of any team member resulting in multiple hypotheses that need to be tracked. This transfers to the present model in a surprising way: The robot (the vehicle) tracks the driver's gaze target considering different motion models (fixations and saccades), while multiple objects (hypotheses) can act on the target (attract the gaze).

5.1.2 Human Gaze Behavior Model Knowledge

In Chapter 2, the basics of anatomy of human eyes as well as of human viewing behavior in general and while driving have been introduced. Some of these insights can be used as model knowledge for the filter algorithm. Firstly, assumptions derived from top-down (task-related) factors include:

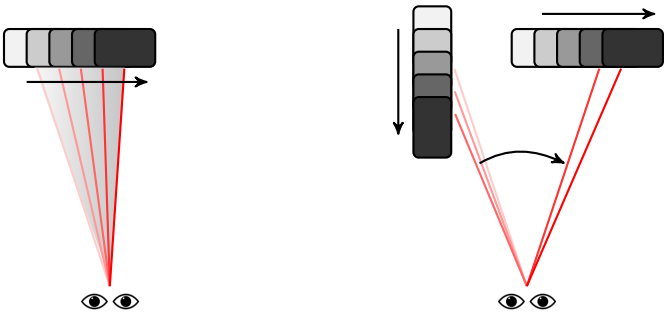
- There always exists a gaze target, i. e. something that the driver is looking at and this gaze target can change.
- These gaze targets are mainly task relevant objects, i. e. other traffic participants such as oncoming traffic, merging traffic but also vulnerable road users such as bikers and pedestrians that the driver is expected to not have a collision with. The driver can also look at task relevant regions inside the vehicle such as the speedometer, yet, these glances are discarded within the scope of this work.
- During the time that a driver is not looking at such a target, they can exhibit free viewing behavior, e. g. look at buildings, advertising, etc. This means, that an alternative hypothesis needs to be provided.

Since task relevance also means that the driver is mainly looking forward i. e. in the eyes-on-road region[34], this thesis is also about how to obtain more precise, road user-related knowledge gain.

Secondly, modeling assumptions derived from basic gaze motion characteristics and anatomy include:

- The fovea is the region of highest acuity on the retina of the human eye. The area of sharp view corresponds to approximately the central 2° of vision.
- Gaze consists of fixations, i. e. periods of relative stability, and saccades, i. e. rapid jerky movements. Both have characteristic durations.
- Saccades have characteristic speeds and thus, the angular distance of two consecutive gaze rays is limited between two time steps.
- The term fixation is used here as in the naturalistic eye movement literature so that smooth pursuits and fixations belong to the same category of eye motion stabilizing the visual target on the retina.

[34] It has been observed that a simple center-bias, i. e. the average gaze direction distribution, is a great predictor for gaze location prediction [180].



(a) Gaze moving consistently with the current target.

(b) Gaze jumping from one target to another.

Figure 5.2: The motivation for the proposed MHMM tracking approach is the dynamic behavior of human gaze. In case of fixations and smooth pursuits, gaze follows the current object of visual attention. Over time, the gaze target changes. In case of a gaze jump to another object, the model should be able to follow the highly nonlinear dynamics.

> This means that during a fixation or a smooth pursuit, the driver is tracking an object with their gaze.

Putting these assumptions together, the gaze ray and the gaze target move consistently during a fixation or a smooth pursuit, meaning that their measurements should overlap for a measurable time span. This motivation is schematically depicted in Fig. 5.2. “Small” measurement errors, either of gaze or object, should be compensated by the filter if object and gaze are close to each other and if the course of the gaze suggests a fixation. Between periods of fixations, jumps can guide the gaze from one visual target to the next. The designed filter approach should be able to handle this dynamic behavior.

5.2 Multi-Hypothesis Multi-Model Gaze Target Tracking

5.2.1 System Overview

The superordinate goal when estimating the gaze target with a probabilistic filter is to determine the full joint posterior probability density of the driver’s

gaze, the surrounding objects and the non-observable variable on which object the driver has their visual attention. The network motivated in Fig. 5.1 describes one possible manifestation of the interplay between the driver's gaze and potential gaze targets. In case of exact and complete inference, the lower level stochastic variables are influenced by the estimate of the current target and its area which are estimated by the filter layer on top. This joint consideration can lead to an information gain in both domains. On the one hand, the fixated object's position could be improved by the additional knowledge of the driver's gaze direction and on the other hand, the gaze direction measurement could be corrected. It is up to the MHMM filter on the upper level to track the area of attention with suitable models given the information of the driver and the surrounding environment. Yet, in this thesis, the lower level filters are modeled as independent from the higher level filter as object data is generally affected by rather low measurement noise and also no exact and precise confidence bounds on the gaze measurement are known, meaning that a correction of the objects' positions via the gaze observation is difficult. Using this approximate inference, the joint probability factorizes into separate filter models for each sub-problem and the result of the higher filter is not propagated back down to the lower filters which makes the lower state estimates behave like likelihoods, i. e. like measurements. Thus, the simplified version shown in Fig. 5.3 is presented with the focus on the task of gaze target tracking. In the following, the different parts of the model, i. e. state space representation as well as measurement and transition models are explained in detail.

5.2.2 Model Description

Representation of the Posterior Distribution

The goal of the proposed filter model is to determine the joint a-posteriori distribution $p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t})$[35] for the current time t. This joint distribution is modeled as a mixture of Gaussians according to equation (3.33b). Each mixture component thereby represents one of $n+1$ potential targets as schematically depicted in Fig. 5.4. The switching variable $\boldsymbol{s} \in \mathbb{R}^{n+1}, s_i \in \{0,1\}, \sum_i s_i = 1$ is a $(n+1)$-dimensional binary random variable having a 1-of-$(n+1)$ representation in which only one s_i is equal to one representing the current gaze target and the rest of the elements

[35] The notation $(\{\cdot,\cdot\}^{1:t})$ is used as a short, yet comprehensive way to express that several variables share a common time index, usually denoted by $(\cdot^{1:t}, \cdot^{1:t})$.

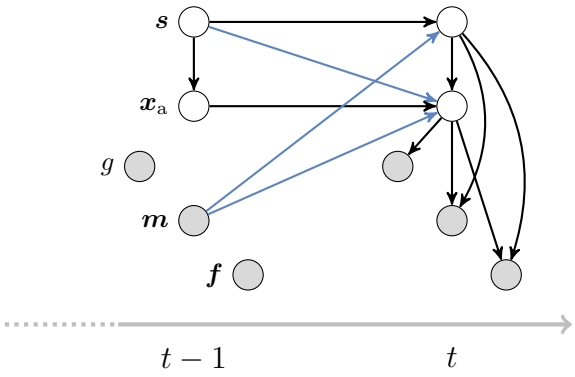


Figure 5.3: Simplified version of Multi-Hypothesis Multi-Model filter. $g, \boldsymbol{m}, \boldsymbol{f}$ denote the measurements of gaze, objects and free space. Compared to the standard version of a jump Markov system in Fig. 3.3, additional dependence connections are used which are drawn in blue. The dependencies from $\boldsymbol{m}^{t-1}$ and $\boldsymbol{s}^{t-1}$ to $\boldsymbol{x}_\mathrm{a}^t$ are introduced together with the motion transition models in Section 5.2.2. The arrow from $\boldsymbol{m}^{t-1}$ to $\boldsymbol{s}^t$ is explained in Section 5.2.3. Arrows to measurements at $t-1$ are again omitted for better view.

are zero. Just like in Chapter 3, the notation $s_i^t = 1$ is abbreviated as s_i^t in the following when used in the argument of probabilities. The dimension n is given by the number of objects in the object list and the additional dimension holds the hypothesis of the static environment. Each potential gaze target i has a certain probability of being the current target of attention denoted with $p(s_i) = \pi_i$ and a certain spatial distribution $p(\boldsymbol{x}_\mathrm{a}|s_i)$ described by a single normal distribution $\mathcal{N}(\boldsymbol{x}_\mathrm{a}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with $\boldsymbol{x}_\mathrm{a}, \boldsymbol{\mu}_i \in \mathbb{R}^2$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{2\times 2}$. The random variable $\boldsymbol{x}_\mathrm{a}$ denotes the area of attention in the 2D vehicle coordinate system. The joint distribution is then given as weighted combination of the individual normal distributions

$$p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t}) = \prod_{i=1}^{n+1} (\pi_i^t)^{s_i} \mathcal{N}(\boldsymbol{x}_\mathrm{a}^t | \boldsymbol{\mu}_i^t, \boldsymbol{\Sigma}_i^t)^{s_i}. \tag{5.1}$$

The measurement $g = \alpha \in \mathbb{R}$ describes the yaw angle[36] of gaze measured by the head eye tracking system. $\boldsymbol{m} \in \mathbb{R}^{5n}$ is an ordered, concatenated object list of variable length containing the position (x, y), heading (φ)

[36]While g describes the observed stochastic variable of gaze direction, α denotes the measured yaw angle. In this context, they can be used interchangeably.

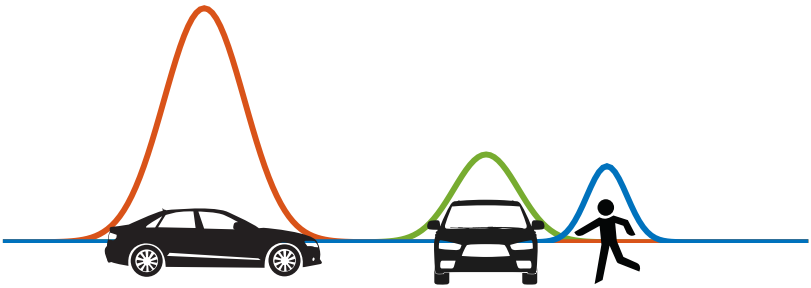


Figure 5.4: The posterior distribution is modeled as Gaussian mixture: Each mixture component represents one potential gaze target. The corresponding weight describes the probability of the respective object of being the current target.

and relative velocity $(v_{\mathrm{rel},x}, v_{\mathrm{rel},y})$ of the dynamic objects captured by the ego-vehicle's radar and tracking system. Finally, the static measurement $\boldsymbol{f} = \{\boldsymbol{x}_\mathrm{P}\}, \boldsymbol{x}_\mathrm{P} \in \mathbb{R}^2$ is the set of control points of the B-spline describing the free space area (cf. Section 5.2.2, [154]). The full Bayesian filter equation is given according to equation (3.7b) by

$$p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t}) \sim \underbrace{\ell(\{g, \boldsymbol{m}, \boldsymbol{f}\}^t | \{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t)}_{\text{Emission}} \underbrace{p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1})}_{\text{Prediction}}. \quad (5.2)$$

The prediction is analyzed first followed by the emission or measurement likelihood[37] and the inference step.

Prior

The prediction is realized by multiplying the transition probability with the old posterior and marginalizing over all possible states of the previous time step (cf. equation (3.25)). It is thus given as

$$p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1}) = \sum_{\boldsymbol{s}^{t-1}} \int \underbrace{p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}, \boldsymbol{m}\}^{t-1})}_{\text{Transition}} \underbrace{p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^{t-1} | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1})}_{\text{Old posterior}} d\boldsymbol{x}_\mathrm{a}^{t-1}. \quad (5.3)$$

[37] Likelihoods are here explicitly denoted by $\ell(\cdot)$ to distinguish measurement models from a valid pdf.

The fact that the transition probability contains the object measurements follows from the introduced connection in the graphical model which is a modeling aspect that differs from the standard jump Markov system structure. These measurements are used as control input which is explained subsequently[38]. The free space measurement of the previous time step is not included since it is not used in the modeling which will become clear in the following transition description as well.

Each possible transition of the driver's gaze target from one time step to the next needs to be defined with the following cases to occur: the driver's gaze can stay on a specific object $i \in 1, ..., n$, it can jump from object i to object $j \in 1, ..., n$, or to the static environment $j = n + 1$. Furthermore, also the static environment could be the current gaze target $i = n + 1$ and gaze can either stay or jump to another static location, in both cases $j = n + 1$. Or, alternatively, it can jump to another object $j \in 1, ..., n$. Due to the filter structure in Fig. 5.3 and the representation of the old posterior as mixture of Gaussians, the transition is practically formulated as a spatial transition and a scalar transition model of the switching state and can thus be further factorized into

$$p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}, \boldsymbol{m}\}^{t-1}) = p(\boldsymbol{x}_\mathrm{a}^t | \boldsymbol{s}^t, \{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}, \boldsymbol{m}\}^{t-1}) p(\boldsymbol{s}^t | \boldsymbol{s}^{t-1}). \tag{5.4}$$

Transition of the Switching State Variable Normally, the transition weights of the switching variable $p(\boldsymbol{s}^t | \boldsymbol{s}^{t-1})$ are constant filter design parameters that can be captured in one matrix $\boldsymbol{\Pi}$ with row sum 1 where each entry p_{ij} describes the probability that the gaze target switches from target i to target j. For a shorter notation of terms, the abbreviation by the indices "ij" means that for the indexed term $s_i^{t-1} = 1, s_j^t = 1$ holds. For now, it is assumed that the transition density can thus be formulated as

$$p(\boldsymbol{s}^t | \boldsymbol{s}^{t-1}) = \prod_{j=1}^{n+1} \prod_{i=1}^{n+1} p_{ij}^{s_i^{t-1} s_j^t}, \tag{5.5a}$$

$$p(s_j^t | s_i^{t-1}) = p_{ij}. \tag{5.5b}$$

Later, in Section 5.2.3, this transition density is extended by time-variant components as well as a spatial dependency on $\boldsymbol{m}^{t-1}$.

[38] Generally, position and velocity are estimated in object tracking. Here, as depicted in Fig. 5.3, directly the measurements are used.

Motion Transition Models (Spatial Prediction) The transition of the spatial distribution of attention $p(\boldsymbol{x}_\mathrm{a}^t|\boldsymbol{s}^t, \{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}, \boldsymbol{m}\}^{t-1})$ incorporates assumptions of gaze behavior, namely that gaze undergoes either a fixation or a saccade. Smooth pursuits are thereby treated just as fixations. By allowing the gaze to jump between targets, saccades are implicitly included in the filter model. These two aspects represent one main part of the proposed filter.

The spatial transition for each possible combination of gaze targets is modeled as linear dynamic process

$$\boldsymbol{x}_\mathrm{a}^t = \boldsymbol{x}_\mathrm{a}^{t-1} + \boldsymbol{u}_{ij}^{t-1} + \boldsymbol{\gamma}_{ij}, \qquad \boldsymbol{\gamma}_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{ij}) \tag{5.6}$$

with control input $\boldsymbol{u}_{ij}^{t-1}$ and process noise $\boldsymbol{\gamma}_{ij}$ determined by the covariance matrix $\boldsymbol{\Gamma}_{ij}$. As the indices suggest, control input and process noise are chosen dependent on the specific combination of i and j.

If attention stays at one object, describing a fixation or a smooth pursuit, the driver's visual attention should have the same motion dynamics as the target. Thus, for $i{=}j$, the control input is modeled by the relative motion of the target to the driver $\boldsymbol{u}_{ij}^{t-1} = \boldsymbol{v}_{i,\mathrm{rel}}^{t-1}\Delta t$, with relative velocity $\boldsymbol{v}_{i,\mathrm{rel}}$ of object i and gaze sampling time Δt.

At the same time, if the driver's visual attention is jumping to another object (saccade), the predicted distribution should be at the location of that object. For $i \neq j$ and gaze jumps onto dynamic objects, the control input is therefore modeled as $\boldsymbol{u}_{ij}^{t-1} = \boldsymbol{\mu}_{oj}^{t-1} - \boldsymbol{\mu}_i^{t-1} + \boldsymbol{v}_{j,\mathrm{rel}}^{t-1}\Delta t$. Here, $\boldsymbol{\mu}_{oj}^{t-1}$ is the last measured position of object j. Thereby, the predicted mean lies at the predicted object position. Here, in the model for the spatial transitions, the dependency of $\boldsymbol{x}_\mathrm{a}^t$ on $\boldsymbol{s}^{t-1}$ and $\boldsymbol{m}^{t-1}$ in the graphical model becomes clear.

For a fixation on the static scene, i. e. $i{=}j{=}n+1$, no relative motion is considered, thus $\boldsymbol{u}_{i,j}^{t-1} = \mathbf{0}$. But since the driver's gaze could jump from the current position on the free space spline to almost any other position on the spline, this case must also cover saccades. Therefore, the process noise covariance given by the matrix $\boldsymbol{\Gamma}_{i,j}$ is chosen larger.

And finally, gaze jumps to the static environment ($i \in 1...n$ and $j = n{+}1$) need to be considered separately. Since it is not possible to determine where the gaze will jump onto the free space spline, the current location estimate of the static target $\boldsymbol{\mu}_j^{t-1}$ serves as basis for the prediction. Thus the control input is given by $\boldsymbol{u}_{ij}^{t-1} = \boldsymbol{\mu}_j^{t-1} - \boldsymbol{\mu}_i^{t-1}$.

Fig. 5.5 schematically shows the idea how the different spatial components contribute to the predicted distribution. The resulting prediction of where

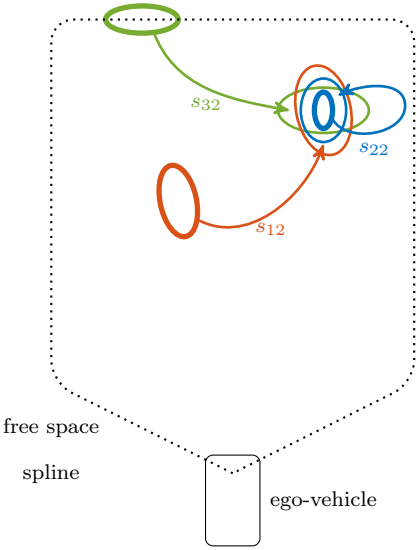


Figure 5.5: Creation of the prior: each transition $i \to j$ is modeled by a switching weight component p_{ij} and a motion transition model. Here, the schematic depiction shows the transitions $i \to 2$ for $n = 2$ objects and a static free space.

the potential gaze targets and the gaze are expected is constrained to a small and specific area. Written in the form of a probability (cf. equations (3.11) and (3.34)), the mode matched spatial transition

$$p(\boldsymbol{x}_{\mathrm{a}}^{t} | s_j^t, \{s_i, \boldsymbol{x}_{\mathrm{a}}, \boldsymbol{m}\}^{t-1}) = \mathcal{N}(\boldsymbol{x}_{\mathrm{a}}^{t} | \boldsymbol{x}_{\mathrm{a}}^{t-1} + \boldsymbol{u}_{ij}^{t-1}, \boldsymbol{\Gamma}_{ij}). \tag{5.7}$$

is used in the evaluation of the prior in the next step.

Evaluating the Prior Now that all terms of the transition have been modeled and defined, the prior is evaluated. Just like in equation (3.30) in Section 3.4, it is sufficient to consider each object j separately in the filter step and thus also in the formulation of the prior. Thus, by inserting (5.1) for the old posterior, (5.5a) and (5.7), the prior (5.3) for a given s_j^t can be

written and reorganized as

$$
\begin{aligned}
&p(\{s_j, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1}) \\
&= \sum_{\boldsymbol{s}^{t-1}} \int p(s_j^t | \boldsymbol{s}^{t-1}) p(\boldsymbol{x}_\mathrm{a}^t | s_j^t, \{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}, \boldsymbol{m}\}^{t-1}) \\
&\qquad\qquad p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^{t-1} | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1}) d\boldsymbol{x}_\mathrm{a}^{t-1} \\
&= \sum_{\boldsymbol{s}^{t-1}} \int \prod_{i=1}^{n+1} p_{ij}^{s_i} \mathcal{N}(\boldsymbol{x}_\mathrm{a}^t | \boldsymbol{x}_\mathrm{a}^{t-1} + \boldsymbol{u}_{ij}^{t-1}, \boldsymbol{\Gamma}_{ij})^{s_i} \pi_i^{s_i} \mathcal{N}(\boldsymbol{x}_\mathrm{a}^{t-1} | \boldsymbol{\mu}_i^{t-1}, \boldsymbol{\Sigma}_i^{t-1})^{s_i} d\boldsymbol{x}_\mathrm{a}^{t-1} \\
&= \sum_{i=1}^{n+1} \int p_{ij} \mathcal{N}(\boldsymbol{x}_\mathrm{a}^t | \boldsymbol{x}_\mathrm{a}^{t-1} + \boldsymbol{u}_{ij}^{t-1}, \boldsymbol{\Gamma}_{ij}) \pi_i \mathcal{N}(\boldsymbol{x}_\mathrm{a}^{t-1} | \boldsymbol{\mu}_i^{t-1}, \boldsymbol{\Sigma}_i^{t-1}) d\boldsymbol{x}_\mathrm{a}^{t-1}. \quad (5.8)
\end{aligned}
$$

The integral over the multiplication of two Gaussians is again a Gaussian and thus, the mode matched prior can be written as

$$
\begin{aligned}
&p(\{s_j, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1}) = \\
&\qquad\qquad \sum_{i=1}^{n+1} p_{ij} \pi_i \mathcal{N}(\boldsymbol{x}_\mathrm{a}^t | \boldsymbol{\mu}_i^{t-1} + \boldsymbol{u}_{ij}^{t-1}, \boldsymbol{\Sigma}_i^{t-1} + \boldsymbol{\Gamma}_{ij}). \quad (5.9)
\end{aligned}
$$

As expected from a multi-model filter approach, it is visible that the prediction for one given gaze target object j is composed of $n+1$ different components as shown in Fig. 5.5. Thus, a total of $(n+1)^2$ filters are needed to compute all possible configurations given the previous and current gaze target. Within the inference step described after the measurement models, the different components for one potential gaze target are collapsed to one single component per object so that the filter approach stays tractable.

Measurement Models

The measurements of gaze, dynamic objects and static environment are assumed to occur independently from each other. For gaze, this is quite intuitive since the information is taken from different sensors than the outer environment. For the other two, except for few artifacts, dynamic objects are removed from the occupancy gridmap from which the free space originates prior to the free space spline computation. Thus the likelihood factorizes to

$$
\ell(g, \boldsymbol{m}, \boldsymbol{f} | \boldsymbol{s}, \boldsymbol{x}_\mathrm{a}) = \ell(g | \boldsymbol{s}, \boldsymbol{x}_\mathrm{a}) \ell(\boldsymbol{m} | \boldsymbol{s}, \boldsymbol{x}_\mathrm{a}) \ell(\boldsymbol{f} | \boldsymbol{s}, \boldsymbol{x}_\mathrm{a}). \quad (5.10)
$$

For simplicity, time indices are omitted since the measurements are incorporated at the same and current time t.

Gaze Model The gaze measurement applies to all potential gaze targets and is thus independent of $\boldsymbol{s}$. Furthermore, the proposed gaze model incorporates the following concepts:

- The area of sharp vision (foveal region) is within a few degrees around the line of sight.
- The likelihood is higher at locations closer to the measured line of sight in terms of angular distance.
- The probability decreases with the distance of an object to the driver favoring objects in the driver's vicinity.

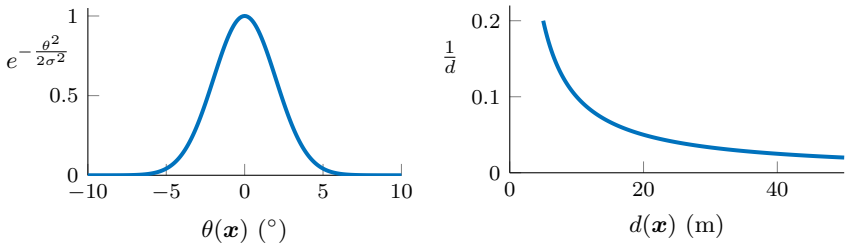
Thus, in order for the model to assign a high likelihood to positions close to the gaze ray, the measurement model for the gaze direction g given a point $\boldsymbol{x} = (x, y)^\top \in \mathbb{R}^2$ in the plane is formulated as

$$g = \arctan\left(\frac{y - y_\mathrm{o}}{x - x_\mathrm{o}}\right) + \nu, \qquad \nu \sim \mathcal{N}(0, \sigma), \tag{5.11}$$

where $\boldsymbol{x}_\mathrm{o}$ is the gaze origin, i. e. the location of the drivers eyes and where the parameter $\sigma \in \mathbb{R}, \sigma > 0$ controls the opening angle of the likelihood "cone". For simplicity, measurement noise and the deviation of the line of sight to the visual axis are combined in the single parameter σ. For the formulation of the likelihood term, the angular difference between the gaze direction g and the point's direction $\measuredangle \boldsymbol{x} = \arctan\left(\frac{y-y_\mathrm{o}}{x-x_\mathrm{o}}\right)$ is denoted with θ. In order to realize a smaller likelihood for objects which are farther away, the likelihood term for $\boldsymbol{x}$ at distance $d(\boldsymbol{x})$ to the gaze origin is weighted with $\frac{1}{d}$. The reason for this weighting is given by the fact that an object in the line of sight but close to the driver would otherwise obtain a smaller likelihood function than an object of the same size farther away. This is, because the gaze measurement model is defined in polar coordinates and the closer object extends over a larger angular area. In terms of probabilities, this means a larger uncertainty which is compensated by the weighting factor.

Putting everything together, the likelihood is formulated as

$$\ell(g|\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}) = \ell(g|\boldsymbol{x}_\mathrm{a}) = \frac{1}{d}\mathcal{N}(g|\measuredangle \boldsymbol{x}, \sigma) \sim \frac{1}{d} e^{-\frac{\theta^2}{2\sigma^2}}. \tag{5.12}$$



(a) Gaze likelihood decrease over the angular distance to the measured gaze ray.

(b) Gaze likelihood decrease over the distance to the driver's eyes.

Figure 5.6: Course of the gaze likelihood in each direction.

The course of the model in angular as well as radial direction is shown in Fig. 5.6 and a gray-scale example image of the likelihood is shown in Fig. 5.7a.

A quite similar approach is used in [104] where objects farther away from the gaze ray are less probable to be seen. However, in that work the opening angle is about 10-times larger and thus covering the total field of view while here only the foveal region is considered. This has interesting implications: the same gaze model can be used to model peripheral view and fixations depending on the parameter choice. But only for modeling fixations, it is reasonable to also apply a gaze target tracking approach.

Object Likelihood Each object j[39] is represented by a weighted Gaussian $\tilde{\mathcal{N}}(\boldsymbol{x}_\mathrm{a}|\boldsymbol{\mu}_{oj}, \boldsymbol{\Sigma}_{oj})$ with mean at the object's position. The covariance matrix' eigenvalues are chosen such that the major axes of the 90 % covariance ellipse correspond to width and length of the object and the ellipse's orientation corresponds to the object's heading. Being a likelihood, $\tilde{\mathcal{N}}$ describes a scaled normal distribution such that the maximum at the mean is equal to one. In Fig. 5.7b, an exemplary scene with three objects of different sizes is shown. In formulas, the likelihood for all objects can be written as

$$\ell(\boldsymbol{m}|\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}) = \prod_{j=1}^{n} \tilde{\mathcal{N}}(\boldsymbol{x}_\mathrm{a}|\boldsymbol{\mu}_{oj}, \boldsymbol{\Sigma}_{oj})^{s_j}. \tag{5.13}$$

[39] Just like in Section 3.4, index j is used at the current time step while i is used at the previous time step. That way, the notation of some terms can be abbreviated.

Free Space Spline Curve As already stated, an alternative hypothesis of where the driver might be looking must be provided for the case that the gaze target is not among the road users in the object list. The likelihood for the static environment is obtained from the control points $\boldsymbol{x}_\mathrm{P}$ of a B-spline representation of free space [153, 154]. Since this spline curve can neither be formulated in closed form nor does it come in Gaussian shape, samples from the contour are taken: first, points on the spline $\boldsymbol{r}(s) = (x(s), y(s))^\top$ between the control points are sampled with the coordinate functions $x(s)$ and $y(s)$[40] and curve parameter s. It is then linearly interpolated and sampled between relevant spline points $\boldsymbol{r}(s^\star)$ with a predefined distance d_fs which approximates the spline contour in the gaze region sufficiently yielding a set of sample points $\boldsymbol{x}_\mathrm{fs}$. Since sampling is performed in the inference step due to the non-Gaussian gaze likelihood model anyway, this poses no drawback. In formulas, the likelihood can be written as

$$\ell(\boldsymbol{f}|\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}) = f_\mathrm{f}(\boldsymbol{x}_\mathrm{fs}, \boldsymbol{x}_\mathrm{a})^{s_{n+1}} \tag{5.14}$$

$$f_\mathrm{f}(\boldsymbol{x}_\mathrm{fs}, \boldsymbol{x}_\mathrm{a}) = \begin{cases} \rho & \text{if } \boldsymbol{x}_\mathrm{a} \in \{\boldsymbol{x}_\mathrm{fs}\}, \\ \epsilon & \text{otherwise, with } \rho \gg \epsilon > 0, \end{cases} \tag{5.15}$$

so that $\ell(\boldsymbol{f}|s_{n+1} = 0, \boldsymbol{x}_\mathrm{a}) = 1$ which means that the likelihood does not affect the modeled probabilities of the dynamic targets. On the other hand, ϵ is chosen very small so that in the sampling process in the inference step only the points $\boldsymbol{x}_\mathrm{fs}$ need to be taken into account while other locations can be neglected. An exemplary image of a free space spline overlaid over the occupancy grid map is shown in Fig. 5.7c.

As of now, all terms of the filter's transition and emission have been defined. In the following section, the inference step is explained, together with the measures to keep the algorithm tractable.

Inference

The total inference step for the filter in Fig. 5.3

$$p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t}) \sim \ell(g, \boldsymbol{m}, \boldsymbol{f}|\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}) p(\{\boldsymbol{s}, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1}) \tag{5.16}$$

[40] Detailed information on the free space spline, its generation and mathematical description can be found in the dissertation of Schreier in [153].

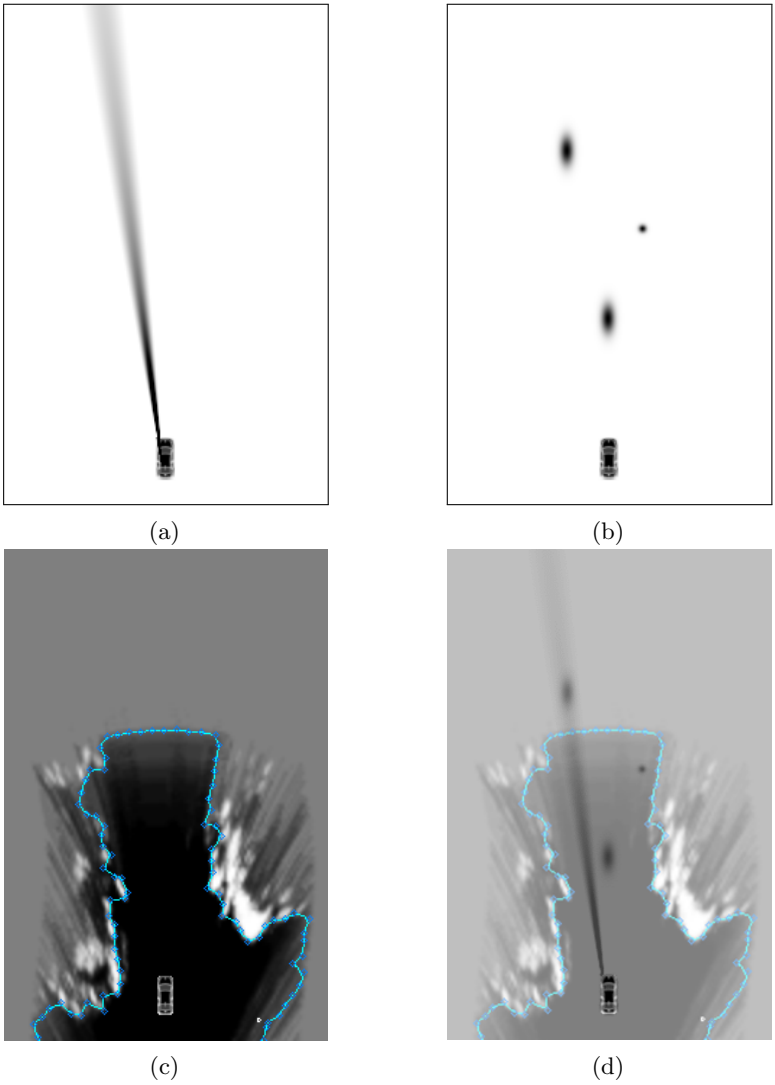


Figure 5.7: Images of the likelihood models. (a) Gaze (darker areas describe higher likelihood); (b) Object likelihood for three dynamic objects; (c) Free space spline drawn on the occupancy gridmap; (d) Additive overlay of the measurement models. Even without any tracking, a one step estimation of the gaze target can be obtained from the multiplication of the gaze model with the measurements of objects and free space spline.

can now be written for a given s_j^t (cf. (3.28c) and (3.30)) as

$$p(\{s_j, \boldsymbol{x}_\mathrm{a}\}^t | \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t}) \sim \ell(g|\boldsymbol{x}_\mathrm{a})\ell(\boldsymbol{m}|s_j, \boldsymbol{x}_\mathrm{a})\ell(\boldsymbol{f}|s_j, \boldsymbol{x}_\mathrm{a}) \times \sum_{i=1}^{n+1} p_{ij}\pi_i \underbrace{\mathcal{N}(\boldsymbol{x}_\mathrm{a}^t | \boldsymbol{\mu}_i^{t-1} + \boldsymbol{u}_{ij}^{t-1}, \boldsymbol{\Sigma}_i^{t-1} + \boldsymbol{\Gamma}_{ij})}_{p(\boldsymbol{x}_\mathrm{a}^t | s_j, s_i, \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1})}. \quad (5.17)$$

In the above inference step (5.17),

1. the gaze and free space likelihoods $\ell(g|\boldsymbol{x}_\mathrm{a})$ and $\ell(\boldsymbol{f}|s_j, \boldsymbol{x}_\mathrm{a})$ are modeled such that a formulation of the filter in closed form is not possible[41] and
2. the prediction step yields $(n+1)$ sums of $(n+1)$ Gaussians (one sum for each target object j) which makes the filter intractable without suitable approximations.

Approximation of Single Components To encounter the first problem, each distribution resulting from a possible transition $i \to j$ is approximated by a Gaussian. By comparing (5.17) with (3.32), it can be seen that the mode matched filtering is normalized with the model likelihood Λ_{ij} from which the approximation

$$\begin{aligned} \ell(g|\boldsymbol{x}_\mathrm{a})\ell(\boldsymbol{m}|s_j, \boldsymbol{x}_\mathrm{a})\ell(\boldsymbol{f}|s_j, \boldsymbol{x}_\mathrm{a}) p(\boldsymbol{x}_\mathrm{a}^t | s_j, s_i, \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1}) & \\ = \Lambda_{ij} p(\boldsymbol{x}_\mathrm{a}^t | s_j, s_i, \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t}) & \quad (5.18) \\ \approx \Lambda_{ij} \mathcal{N}(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}) & \quad (5.19) \end{aligned}$$

follows. Mean and covariance of the fitted Gaussian are obtained from a set of K weighted sample points $\boldsymbol{x}_k$ and are given by maximum likelihood parameter estimation [136]

$$\boldsymbol{\mu}_{ij} = \frac{\sum_k \lambda_{ijk} \boldsymbol{x}_k}{\sum_k \lambda_{ijk}}, \quad (5.20)$$

$$\boldsymbol{\Sigma}_{ij} = \frac{\sum_k \lambda_{ijk} (\boldsymbol{x}_k - \boldsymbol{\mu}_{ij})(\boldsymbol{x}_k - \boldsymbol{\mu}_{ij})^\top}{\sum_k \lambda_{ijk}}, \quad (5.21)$$

with the sample weights λ_{ijk}. Given a set of uniformly distributed samples[42], the sample weights are obtained from the evaluation of the term

[41] Even if the formulation would be possible, e. g. in the absence of the free space, the mode matched predictions are not a conjugate prior to the gaze likelihood.

[42] It is this sampling step in the computation which differs from the author's publication [158].

$\ell(g|\boldsymbol{x}_\mathrm{a})\ell(\boldsymbol{m}|s_j,\boldsymbol{x}_\mathrm{a})\ell(\boldsymbol{f}|s_j,\boldsymbol{x}_\mathrm{a})p(\boldsymbol{x}_\mathrm{a}^t|s_j,s_i,\{g,\boldsymbol{m},\boldsymbol{f}\}^{1:t-1})$ at $\boldsymbol{x}_k$. The likelihood function Λ_{ij} is then approximated by the mean sample weight

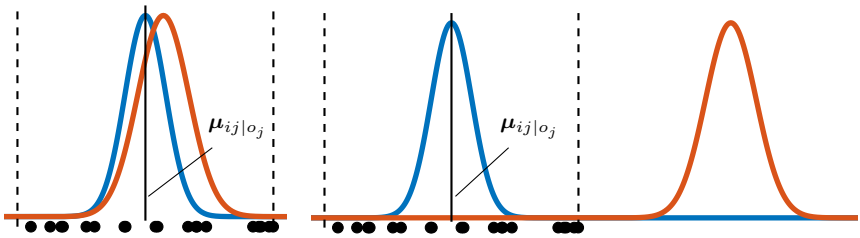
$$\Lambda_{ij} \approx \frac{1}{K}\sum_k \lambda_{ijk}. \tag{5.22}$$

Sample Weights and Model Likelihood for Objects For objects from the object list, the above term can easily be evaluated at sample points $\boldsymbol{x}_k$. Since the multiplication of the two normal distributions $\ell(\boldsymbol{m}|s_j,\boldsymbol{x}_\mathrm{a})p(\boldsymbol{x}_\mathrm{a}^t|s_j,s_i,\{g,\boldsymbol{m},\boldsymbol{f}\}^{1:t-1}) = z_\mathrm{c}\mathcal{N}(\boldsymbol{\mu}_{ij|o_j},\boldsymbol{\Sigma}_{ij|o_j})$[43] can be performed analytically yielding again a Gaussian, it is done prior to sampling. This multiplication of two Gaussians yields another weighting factor z_c expressing the similarity between the area of attention prediction and the measurement of the object position.

A strict uniform sampling procedure has a significant drawback if the investigated object is not the current gaze target. In order to obtain a more precise mean and variance $\boldsymbol{\mu}_{ij},\boldsymbol{\Sigma}_{ij}$, the sample points should cover a region including the peaks of both likelihood functions of object and gaze. In the mentioned case, this would result in a high number of required sample points since the distance between object and gaze ray is large. However, since the distance is large, the resulting respective model likelihood Λ_{ij} is quite small and thus the respective mixture component contributes only little to the final distribution. For this reason, it is argued that a uniform sampling in the object's vicinity as shown in Fig. 5.8 is sufficient for the present application: if the respective object is close to the gaze ray, the sample points also cover the gaze ray leading to a higher precision of the estimated mean and variance. But if the respective object is far from the gaze ray, its model likelihood will be small anyway which is why it would be unprofitable to increase the computation effort for an object without interest. Furthermore, even though it is an approximation, the resulting distribution components stay close to the respective object positions which is advantageous in the subsequent filter step.

Sample Weights and Model Likelihood for Free Space For the evaluation of the mode matched filtering for the free space the already

[43] The intermediate symbols for mean $\boldsymbol{\mu}_{ij|o_j}$ and covariance $\boldsymbol{\Sigma}_{ij|o_j}$ describe the normal distribution resulting from the multiplication of the mode matched filter prediction with the respective object likelihood. At this stage, the multiplication with the gaze likelihood is pending before reaching the new posterior.



(a) Measured gaze close to the prediction.

(b) Measured gaze far to the prediction.

Figure 5.8: Schematic view of the sampling step reduced to one dimension. The predicted area of attention updated with the object likelihood is schematically represented by the blue curve. The gaze likelihood is represented by the red curve and an exemplary set of sample points is drawn in black. In order to approximate the multiplication of the two curves with sampling, the samples need to cover a meaningful value range. If gaze and predicted area of attention are close (a), a sampling covering the two peaks approximates the result sufficiently. If gaze and predicted area of attention are not close by (b), then the sampling effort covering the main peaks with the samples would be high despite that the resulting model likelihood Λ_{ij} would be small and therefore also the probability of the respective object to be visually attended is comparably low. Therefore, the area from which samples are drawn is restricted to the 99.5 % confidence range around $\boldsymbol{\mu}_{ij|o_j}$ represented by the dashed lines.

introduced sample points $\boldsymbol{x}_k = \boldsymbol{x}_{\mathrm{fs}}$ are used. As a consequence, the sample points are already limited to a certain region in the 2D vehicle coordinate system. However, in contrast to dynamic objects, there exists no normal distribution which can be multiplied with the area of attention prediction and thus, also no weighting factor z_{c} exists, which expresses the similarity between prediction and free space measurement. Since neglecting this effect leads to a practically observable over-weighting of the static environment, the weighting factor is approximated by the mean sample prediction weight $z_{\mathrm{c}} = \frac{1}{K'} \sum_k p(\boldsymbol{x}^t_{\mathrm{fs},k} | s_j, s_i, \{g, \boldsymbol{m}, \boldsymbol{f}\}^{1:t-1})$, where K' is the number of used sample points for the static hypothesis.

Approximation of Gaussian Mixtures by GPB II In order to keep the filter tractable, the second problem of an exponentially growing number of mixture components is tackled by approximating the sum of

Gaussians per target with one single Gaussian according to the second order generalized pseudo-Bayesian estimator introduced in Section 3.4.2. First, the merging probabilities

$$\pi_{ij} = \frac{\Lambda_{ij} p_{ij} \pi_i^{t-1}}{\sum_{i_1}^{n+1} \Lambda_{ij} p_{ij} \pi_i^{t-1}} \tag{5.23}$$

are computed according to (3.44). The merging probabilities are then used to calculate mean and variance of the merged Gaussian via minimization of the KL-Divergence which are obtained from (3.24a) and (3.24b) as

$$\boldsymbol{\mu}_j^t = \sum_i^{n+1} \pi_{ij} \boldsymbol{\mu}_{ij} \tag{5.24}$$

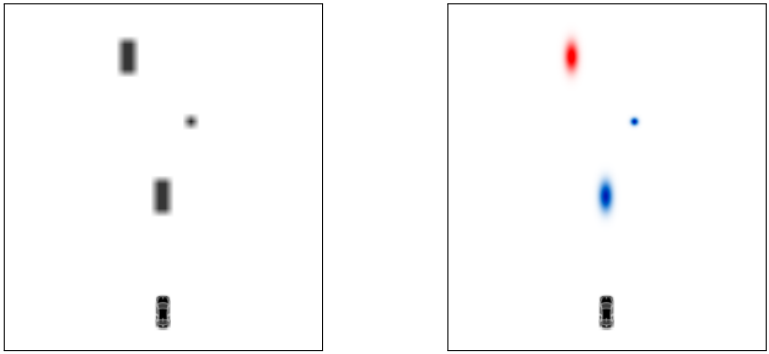
$$\boldsymbol{\Sigma}_j^t = \sum_i^{n+1} \pi_{ij} (\boldsymbol{\Sigma}_{ij} + (\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_j^t)(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_j^t)^\top). \tag{5.25}$$

This is legitimate since for each object the different means (fixation of object j and saccade to object j) are close to each other. Finally, the mode probabilities are updated according to the corrected belief

$$p(s_j) = \pi_j = \frac{\sum_i \pi_{ij}}{\sum_j \sum_i \pi_{ij}}. \tag{5.26}$$

Reinitialization and Reduced Sampling Effort

As can be seen by inserting the respective control input $\boldsymbol{u}_{ij}$ into (5.9), the same predicted mean is obtained for all i given one fixed j (except $i = j$, which needs to be sampled separately). Only the predicted variances $\boldsymbol{\Sigma}_i + \boldsymbol{\Gamma}_{ij}$ differ as they depend on the previous uncertainty $\boldsymbol{\Sigma}_i^{t-1}$. However, it is counterintuitive that the variance of the gaze location after a saccade depends on the previous uncertainty. Practically speaking, the driver's gaze can in principle jump to any location in the scene. For the estimation of the gaze target and location directly after a saccade, the uncertainty of the visual focus of attention before the saccade is of no importance. In contrast, rather the size and shape of the new object of fixation influences the uncertainty of where the driver is looking. Transferred to the tracking model, the predicted variances are instead newly initialized according to the sizes and orientations of the respective objects plus some process noise as shown in Fig. 5.9. The initialization covariance is denoted with



(a) Scene with three road users. (b) Prediction with reinitialization.

Figure 5.9: Spatial gaze prediction given that the gaze target (red) is known from the previous step. The predicted uncertainty in case of saccades (blue) does not depend on the previous gaze target, but on the size and orientation of the predicted targets.

$\boldsymbol{P}_j = \boldsymbol{\Sigma}_{oj} + \boldsymbol{\Gamma}_{ij}$ and it only depends on the object j[44]. Like this, the spatial dependency on the previous state vanishes in case of saccades and for all jumps from any object i to another object j, $i \neq j$, it suffices to sample only once instead for each i separately. Fixations on targets ($i = j$) are sampled separately as the predicted mean originates from the previous estimate of the area of attention. The sampling effort can thus be practically reduced from complexity $O(n^2)$ to $O(n)$.

In the evaluation section, it is exemplarily shown that this spatial decoupling has only minor effects on the weighting of the different targets.

5.2.3 Incorporation of Gaze Behavior Assumption

The layout of the MHMM tracking model allows the driver's gaze to jump between potential gaze targets and assumes that during a visual fixation the target object and the gaze move consistently. However, given that fixations exhibit a characteristic minimum duration and that saccades exhibit characteristic maximum speeds, the incorporation of this gaze behavior knowledge can be used to further tune the tracking model. In [68], the fixation duration is explained by a cost model for saccades implicating that short fixations are associated with higher effort and are thus penalized, and

[44] $\boldsymbol{\Gamma}_{ij}$ is chosen as fixed parameter.

a bounded actor model formulating the time required for visual processing. Both assumptions impose a certain fixation duration and can be translated into a modeled probability which should be higher after a change of the gaze target. The maximum saccade speed on the other hand can be used to decrease the modeled probability for physiologically impossible jumps.

Normally, the transition weights are constant model parameters (cf. equation (5.5b)). Now, the transition is conditioned on the previous gaze target and the measured objects

$$p(\boldsymbol{s}^t|\{\boldsymbol{s},\boldsymbol{m}\}^{t-1}) = \prod_{j=1}^{n+1}\prod_{i=1}^{n+1}\left(\alpha(s_j^t, s_i^{t-1})\beta(s_j^t, \{s_i, \boldsymbol{m}_i\}^{t-1})\right)^{s_i^{t-1}s_j^t}, \quad (5.27)$$

modeled by two potentials describing a temporal part $\alpha(\boldsymbol{s}^t, \boldsymbol{s}^{t-1})$ and a spatial part $\beta(\boldsymbol{s}^t, \{\boldsymbol{s},\boldsymbol{m}\}^{t-1})$. The temporal transition models the changing probability for fixations and saccades, while the spatial transition considers the geometric relations of the scene. For a shorter notation, again, the abbreviation $p_{ij} = \alpha_{ij}\beta_{ij}$ is used.

Temporal Switching Weight $\boldsymbol{\alpha_{ij}}$ The temporal switching weight realizes the idea of a minimum fixation duration and short saccades. Given the most probable gaze target, the fixation time τ is increased by the sample time if previous and current gaze target are the same and reset to zero if they are different[45]. The modeled probability of a fixation is then decreased with increasing fixation time. Once the estimated target with highest confidence changes, fixation time and switching weights are reset, thus favoring short saccades. A simple sigmoid shaped function is used to compute the transition weight for staying at the current gaze target. The remaining weight is then shared between all other potential targets:

$$\alpha_{ii}(\tau) = \kappa_0 - \frac{\kappa_1}{1+\exp(-\kappa_2(\tau-\tau_0))} \quad (5.28)$$

$$\alpha_{ij}(\tau) = \frac{1-\alpha_{ii}(\tau)}{n+1}, \quad (5.29)$$

where $\tau_0, \kappa_{0,1,2} \in \mathbb{R}$ are predefined parameters. As point of inflection of the sigmoid function, the parameter τ_0 reflects a typical minimum fixation duration.

[45] This rule is a very simple approximation of a higher-order Markov-chain as the switching probability is solely based on the most probable gaze target instead of determining separate probabilities for each possible transition.

Spatial Switching Weight $\boldsymbol{\beta}_{ij}$ The spatial transition weights describe physiological constraints on maximum saccade speeds and general gaze direction. Using the measured object positions, the geometric relations between objects can be exploited in the modeling. Here, again a sigmoid shaped function is applied depending on the angular difference δ_{ij} between gaze target i and the possibly following target j

$$\beta_{ij} = \left(\kappa_3 - \frac{\kappa_4}{1 + exp(-\kappa_5(\delta_{ij} - \delta_0))}\right) c_j, \tag{5.30}$$

$$c_j = \left(1 - \frac{|\delta_j|}{\pi/2}\right), \tag{5.31}$$

where the model parameters $\delta_0, \kappa_{3,4,5} \in \mathbb{R}$ need to be defined beforehand as well. The point of inflection is given by the parameter δ_0 which reflects the angular distance that gaze can cover within one sample cycle. The variable c_j depends on the angle δ_j which denotes the angle between looking at object j instead of looking straight to the front. It has the meaning that objects towards the front have, *in general*, a higher probability of being fixated than objects towards the sides just because glances to the front are more frequent. Since gaze yaw values of more than $\frac{\pi}{2}$ are practically not observable due to the hardware configuration, they also do not need to be considered for δ_j. At the end, the row sums of the resulting matrix $\mathbf{\Pi} = (p_{ij}) = (\alpha_{ij}\beta_{ij})$ are normalized to one so that $\sum_{j=1}^{n+1} p(s_j^t | \{s_i, \boldsymbol{m}\}^{t-1}) = 1$. In the experiments, it is shown that the effort of incorporating these gaze behavior assumptions leads to an increased contrast between targets and thus to an increased confidence when fixations are detected.

5.3 Experimental Results

Insights into the tracking behavior of the proposed method are provided on the basis of several short example sequences. As input for the compared algorithms, the dynamic object list, the free space spline control points and 3D gaze direction measurements with quality value are used[46]. The dynamic object list is a fused and partially classified object list obtained from the radar object list and the stereo camera object list. Single objects are tracked with Kalman filters and static objects are filtered out from the raw lists. Nevertheless, artifacts such as false positive or false negative detections

[46] A more detailed description of the used hardware setup is given in the following chapter which deals with the recording of a suitable evaluation dataset.

or slow tracking cannot be ruled out. Especially the rather low speeds in urban residential areas pose difficulties when distinguishing dynamic and static objects. The free space spline control points are obtained from the method presented in [153, 154] and are based on a free space gridmap representation. The gaze direction is measured by a multi-camera remote head-eye-tracking system which is fully calibrated, extrinsically referenced in the car and time-synchronized with the other data streams.

The following models for gaze target estimation are compared to each other. All approaches operate in the 2D vehicle coordinate system. The first two models are the basic geometric approaches that have already been introduced in the previous chapter. The intersection model (IS) determines the current gaze target as the object that intersects the gaze ray at the closest distance to the driver. In contrast, the threshold model (TH) does not determine one specific gaze target but counts all objects as seen which fall at least partially in the gaze region. This gaze region is defined by an opening angle around the gaze ray chosen here to be 6°. Counting all objects as seen can also be interpreted as distributing the probability weights equally among the "seen" objects. In that way, the model can be compared more consistently to the others. To approach the model proposed in this chapter, first, the measurement model alone is used given in equation (5.10). It results from the evaluation of the likelihood functions and choosing the maximum weight π_i^t and its corresponding mean $\boldsymbol{\mu}_i^t$ as gaze target and gaze location. As the maximum likelihood is used to determine the gaze target, the model is abbreviated with ML. In the following plots comparing the time course of the different model outputs, always the probabilities of each potential gaze target are plotted, not just the final, most probable target objects. The second model is a static multi-hypothesis tracking (MH) approach. The term *static* stands here as characteristic of the filter. It is assumed that one of the hypotheses, i. e. objects or static spline, is the gaze target. Switches in between different targets are not considered, i. e. $p_{ij} = 0, \text{if } i \neq j$. Finally, the MHMM tracking with $p_{ij} = \frac{1}{n+1}$ is used, expanded by the specific time-variant parameter modeling based on human gaze behavior presented in the previous Section (MHMMP). The final model is given by the MHMMP with reduced sampling effort (MHMMPS).

5.3.1 Runtime

As human gaze is highly dynamic, common head-eye-tracking systems generally operate at sampling frequencies of 50 Hz or higher. The applied SmartEye head-eye-tracking system samples gaze and head motion at 60 Hz.

Thus, for a real-time application of the tracking algorithm, the needed runtime for the filter itself must be below 16.7 ms. The proposed MHMM tracking algorithm is implemented in C++ and runs on the experimental vehicle as well as on a standard laptop equipped with a 2.50 GHz Intel Core i7-4710MQ CPU. The run times are evaluated on a ca. 100 s run, i. e. about 6700 time steps, with a varying number of dynamic objects in the object list. Using only highly reliable and relevant dynamic objects yields around $n \approx 2$ whereas using the raw radar list yields a somewhat worst case approximation with $n \approx 37$ objects in the list. However, a large portion of these object clusters represent reflections from the static scene such as buildings, parked cars, poles, etc. The average runtimes for different configurations are shown in Table 5.1. K describes the number of samples per object used in the sampling procedure. If confined to the relevant dynamic objects in the scene, the average runtimes stay well below the sensor update time of 16.7 ms if the reduced sampling approach or the measurement model is applied making the proposed method run sufficiently fast for real time use on the test vehicle. In case of detected violations of the required maximum runtime, it is possible to dynamically adjust the number of used samples to further reduce the computation time.

5.3.2 Tracking in Static Scene

The difficulty of evaluating the gaze target estimation of different models in real traffic is, as already mentioned, the missing ground truth data of where and at what a driver is looking. For this reason, the first experiment is constituted of an artificial static scene with three stationary objects, two vehicles and one person, at distances of about 10-20 m to the ego-vehicle. The test person is advised to stare at one object after another from left to right repeatedly for three times with decreasing glance duration. From these instructions, a ground truth can be constructed. Fig. 5.10 shows the scene as viewed by the vehicle's camera as well as a bird's-eye view with

Table 5.1: Runtime comparison, average runtimes in ms.

	$n \approx 2$ $K = 50$	$n \approx 2$ $K = 10$	$n \approx 37$ $K = 10$
MHMMP $O(n^2)$	2.8	1.5	120.5
MHMMPS $O(n)$	1.5	0.7	12.4
ML (always $O(n)$)	0.6	0.2	2.2

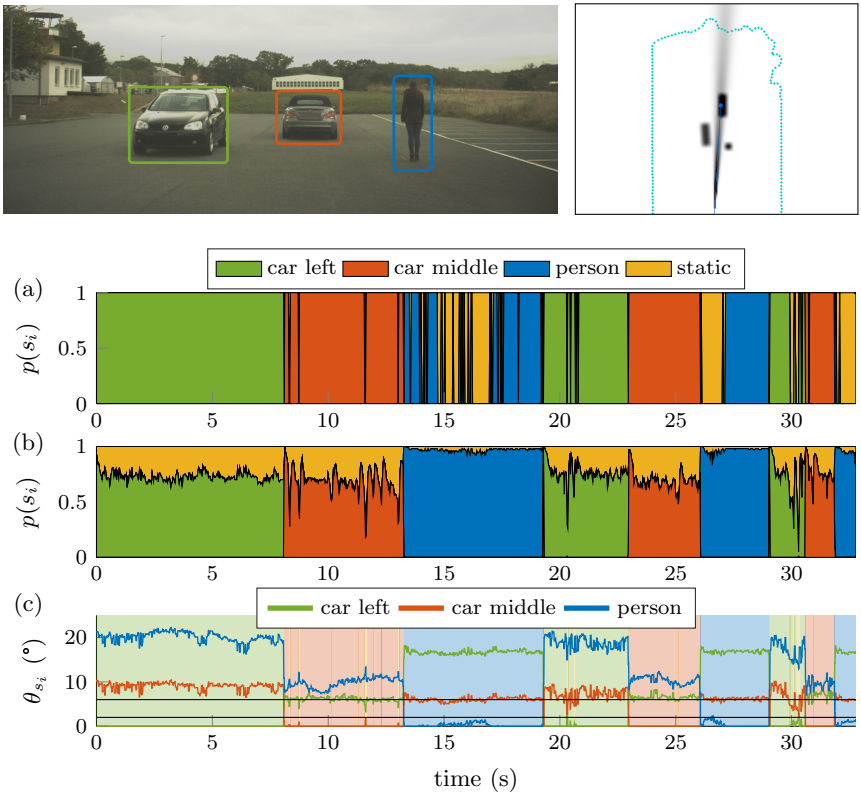


Figure 5.10: Static evaluation scene. (a) Intersection model (IS); (b) Tracking with gaze behavior modeling and reduced sampling effort (MHMMPS); (c) Angular deviation between the gaze heading and the direction of each object in the scene.

the objects and the free space spline points. The proposed model (MHMM) with gaze behavior parameter modeling (P) and reduced sampling (S) is compared to the already known simple geometric intersection model (IS). The respective gaze target estimations are plotted in Fig. 5.10a and Fig. 5.10b as area-diagrams over time. The x-axis describes the time course of the situation and the y-axis denotes the probability $p(s_i)$ of the gaze targets, where one s_i denotes one potential target in the scene. At each time slice, the gaze target is the object whose color has the largest height in

the plot, i. e. the highest portion of modeled attention. The confidence for visual fixations can thus be directly obtained from a colored area during a specific timespan. The bottom plot shows the angular deviation of the gaze ray to each potential gaze target of the yaw angle. If the gaze ray intersects an object, the deviation is zero. The background colors are the respective gaze targets from the MHMMPS model and just like in Fig. 4.9, the two horizontal black lines denote 2° and 6° deviations. Each object is also marked in the scene view with the respective color from the plots below. The static environment, which is not explicitly marked in the camera image, is always colored yellow in the plots showing the probabilities of the gaze targets.

It can again be observed that the gaze measurement accuracy does not suffice to employ a simple IS model, especially to detect visual fixations on smaller objects. Even the smallest deviation leads to a misclassification of the gaze target. Of course, the deviations might stem from other reasons for that imprecise configuration of gaze and surround representation. In contrast, the tracking approach reacts to abrupt changes in the gaze direction measurement under consideration of the surrounding scene. As can be seen in Fig. 5.10c, misclassifications or target jumps occur when the gaze changes abruptly. This corresponds to the intended behavior. During a visual fixation, i. e. periods of slow or no gaze direction change, the gaze target stays the same, even if the gaze ray does not intersect the target at all times. When the gaze target changes, the filter is able to follow this dynamic behavior. However, misclassifications are not completely excluded. Here, they might stem from stabilizing microsaccades as the fixation duration is rather long. Yet, the advantage of the new model is the provided confidence in terms of the modeled probability $p(s_i)$. Another aspect worth to discuss is the preference of objects with a smaller cross section in the bird's eye view. This is visible by the higher probability that the person in the scene obtains when it is looked at compared to the vehicles. On the one hand, this modeled preference might be too strict in dynamic scenes but on the other hand, the goal is the detection of fixations. Especially for small objects, it is more likely that the gaze ray does not directly hit the target. The precision of the gaze target detection for the given simple scenario are 0.83 for the IS model and 0.98 for the filter approach. More specifically, for the periods when the test person glances at the person in the scene, the respective precision are 0.53 and 1.

The reason for the presented scenario is the constructed ground truth and the advantages of the presented tracking approach become already visible. However, the static scene is not suitable to obtain also insights

regarding the model assumptions in dynamic driving scenes. Even though a ground truth is missing for real world scenarios, the subsequent section takes a closer look into the tracking model in dynamic real world situations and gives a qualitative comparison.

5.3.3 Tracking in Real World Driving

Probabilities

Scenario and Plot Description Using a hand-picked scenario, the benefit of the proposed MHMM tracking over simple geometric approaches discussed in the previous chapters are presented. Furthermore, the strengths of a suitable tracking are outlined through the comparison with MH tracking and the measurement model only (ML). Finally, also the benefits of the spatio-temporal modeling of the weight transition parameters p_{ij} are highlighted (abbreviated with MHMMP) as well as additionally the effect of the reduced sampling approach (model abbreviated with MHMMPS). The scenario is shown in Fig. 5.11 where the driver is facing a biker on his ego-lane who intends to turn left thus having a high probability of crossing the lane. The driver is thus also expected to look at the cyclist to visually secure the situation as at the same time a car is approaching from the opposite direction. Furthermore, the object list contains a false detection in front of the building. It has already been used in Fig. 4.8 in Section 4.2.2 where the drawbacks of intersection and threshold models have been discussed. Below, in Fig. 5.12, the area plots of the target probabilities $p(s_i)$ for different models are plotted for a short time-span of about 2 s. These area plots are especially suited since they provide an intuitive insight into the belief of the gaze target of a model. The size of the connected colored areas directly hints at the driver's perception of the specific object even without being able to quantify the driver's perception. The higher the confidence and the longer the high confidence is maintained, the higher is the probability of the driver having perceived the respective object. The first two plots are identical to Fig. 4.8a and Fig. 4.8b whereas the following show different manifestations of the approach proposed in this work. For a direct comparison of these different manifestations, in Fig. 5.13 only the tracking probability of the biker is picked out from each model for better visibility. From the plot in Fig. 5.12g, it can be seen that the driver's gaze clearly switches between two distinct targets, namely the biker and the static scene in the front[47]. Within this sequence, the oncoming car is never

[47] That the static scene is fixated towards the front can be observed in Fig. 5.15.

explicitly fixated.

Intersection and Threshold Models As discussed in Section 4.2.2, the main drawback of the intersection model is the high sensitivity to measurement or calibration errors. As soon as the gaze ray does not intersect the gaze target, a mismatch occurs preventing the model from detecting distinct fixations. In contrast, the threshold model does not specify a gaze target if several objects fall into the modeled gaze cone. It is this trade-off that the proposed approach tries to overcome. The first idea is to incorporate probabilistic assumptions instead of sharp boundaries for both, gaze and objects.

Measurement Model Using the introduced measurement models without any tracking (cf. Fig. 5.7d), the gaze opening cone is softened towards the side and has its highest likelihood on the gaze ray. As visible in Fig. 5.12c, this gives already some insights into the scene. The timespans when the gaze falls on the biker become more clearly defined. Yet, compared to the filtering approaches, the confidences for the gaze targets are still too low to decide on the current gaze target. This is especially because at one single time step, all potential targets within the line of sight experience similar evidence. Using only the current measurements, no model knowledge is incorporated, i. e. assumptions on how gaze and objects move jointly. Furthermore, compared to the filtering approaches, the ML estimate is much more susceptible to noisy data which is a very inherent characteristic. Noisy measurements and available model knowledge are a basic motivation for applying Bayesian filtering.

Figure 5.11: Image of the example scenario. The colors of the targets correspond to those in the diagrams below.

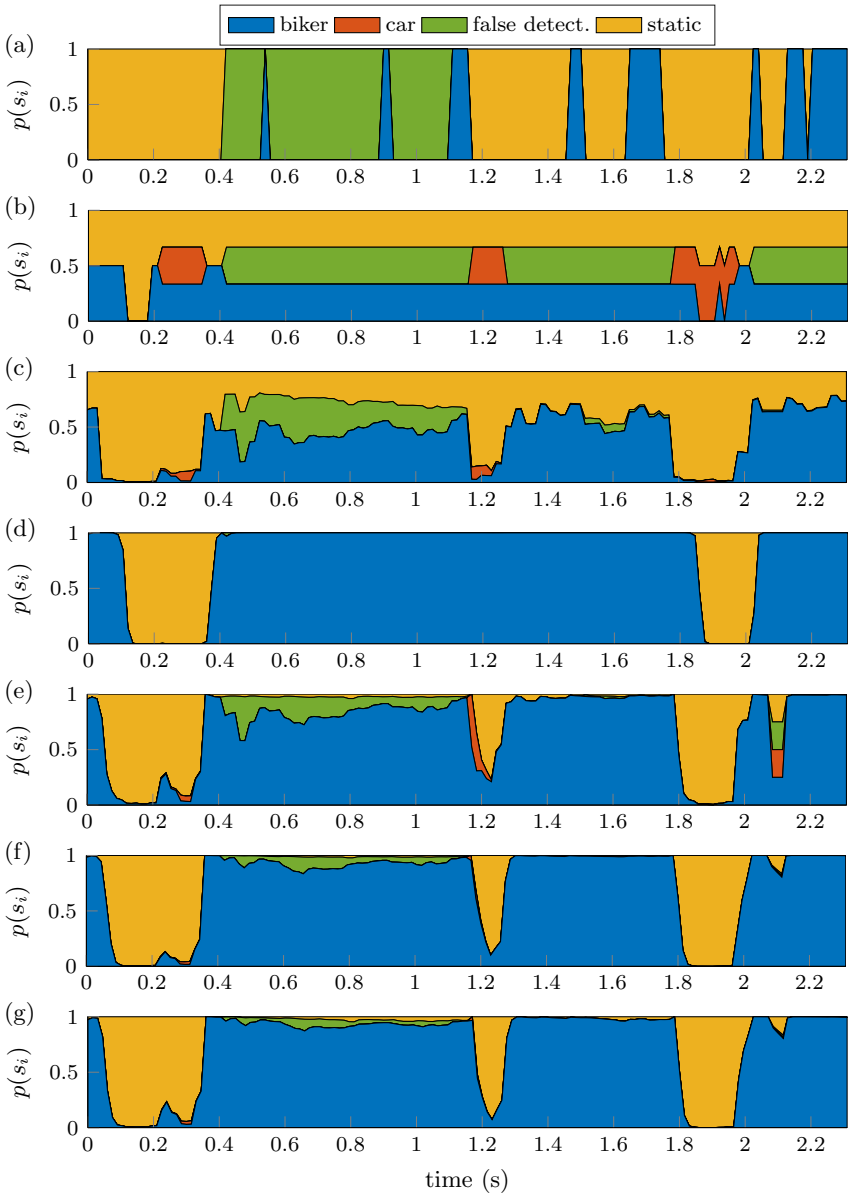


Figure 5.12: Estimated gaze target for different models. (a) Intersection model (IS); (b) Threshold model (TH); (c) Measurement model only (ML); (d) Static Multi-Hypothesis Tracking (MH); (e) MHMM Tracking with $\mathbf{\Pi} = \mathbf{1}$; (f) MHMM Tracking with gaze behavior modeling (MHMMP); (g) MHMM Tracking with gaze behavior modeling and reduced sampling effort (MHMMPS).

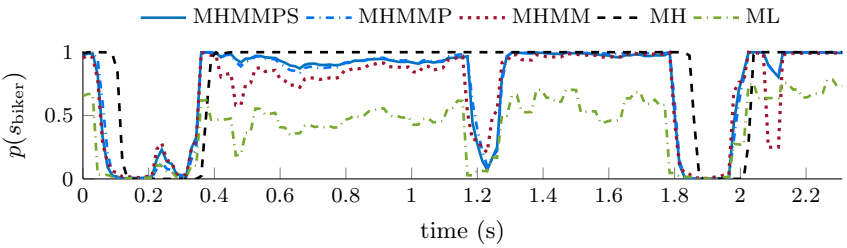


Figure 5.13: Target probability for the cyclist only. The ML model is prone to noise and its capability to decide for one target is weak. The MH estimate has low pass character and is unable to react to short glances. The choice of more elaborate transition parameters (MHMMP/MHMMPS) shows to increase the contrast between gaze targets compared to the basic MHMM. The reduced sampling has only little effect on the modeled probabilities.

Static MH Tracking In direct opposition to the measurement model stands the static multi-hypothesis tracking (MH). Again, in this case, the transition parameters p_{ij} are zero if $i \neq j$ which results in one separate filter for each hypothesis. It can be clearly seen in Fig. 5.13 that this approach exhibits low-pass character as jumps from one target to another are delayed by about 30-60 ms which corresponds to 2-4 sample steps. Due to the lacking trade-off between fixation and saccade motion, the probability of the target with the "best" model will converge to unity and will suppress the probabilities for alternatives. Therefore, a suitable regularization term in each time step is needed [14]. Even then, when the target is changing, the probability of the new target rises only slowly and only if it experiences enough evidence. Here, it can be observed that the short glance towards the front at t =1.2 s is missed. The dynamic potential of the MH tracking is thus mainly characterized by the regularization term[48].

MHMM Tracking The proposed MHMM ($p_{ij} = 1 \; \forall \; i, j$) tracking algorithm overcomes not only the drawbacks of intersection and threshold models but also the just mentioned problems of the ML and MH models: by tracking the possibility to stay at the current target, the estimate is smoother than the measurement model and by enabling saccades to different targets, the model is able to follow the flexible and fast dynamics of gaze motion.

[48] For this example, the regularization term is set to 10^{-5}.

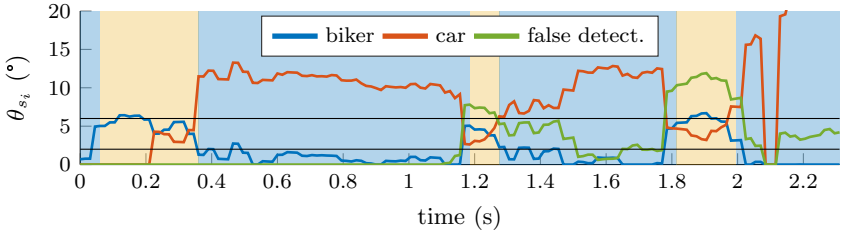


Figure 5.14: Angular deviation between the gaze heading and the direction of each object in the scene, identical to the deviations in Fig. 4.9. The colors in the background are the respective gaze targets from Fig. 5.12f.

Finally, the elaborate and time dependent choice of the transition parameters (MHMMP model) enhances the estimation quality compared to the naive parameter choice by increasing the contrast between the probabilities for different targets. This can be especially observed right after a gaze target change. In that case, the temporal modeling decreases the probability of jumping again right away and supports a minimum fixation duration. In direct comparison in Fig. 5.13, it is furthermore interesting to see that the reduced sampling method with the newly initialized variances after a saccade leads to a comparable result as the modeled probability of the biker being the gaze target is almost identical over the complete sequence.

In all courses of the MHMM tracking variants, it can again be observed that the smaller biker gets favored over the larger fault detection. Yet, in this case, this preference is enhanced since the gaze model also favors objects which are closer to the driver. This modeling aspect is reasonable since the proposed gaze model does not account for occlusions. It is quite intuitive to assume that the object in front is looked at by the driver if two or more detected objects are all in the line of sight. A very interesting consequence is that not necessarily the object closest to the gaze ray is classified as gaze target which is shown in Fig. 5.14. While the false object detection is closer to the gaze ray between 0.4 s and 1.2 s, it is the biker who is classified as the gaze target. Just like in the static scene in the previous section, Fig. 5.14 shows that the gaze target switches when the gaze has abrupt changes.

Another observable characteristic when comparing the ML model with the MHMM approaches is that jumps towards the free space spline are delayed by about 1-2 gaze samples whereas gaze jumps from the spline onto objects are not delayed, cf. Fig. 5.13 and Fig. 5.14. This is another

effect of the modeling. By assuming a larger transition variance for the static environment, this larger expansion implies also a larger uncertainty.

Spatio-Temporal Description

Since the information on the driver's gaze and head motion flows together with all other collected information such as vehicle dynamics and position in one sensor network, the current gaze target can not only be presented over time but can also be put in relation to the spatio-temporal evolution of the driving scene. Such a configuration of the just presented scenario with all object's trajectories as measured by the ego-vehicle together with the estimated driver's gaze target is shown in Fig. 5.15. On the left, the trajectories are shown with equal axis scaling. For better visibility of the gaze target and the jumps in between, the driver's visual behavior is shown on the right with a more favorable scaling. The dashed lines represent the times when the driver's gaze switches from one target to another. The currently "active" target is plotted with a thicker linewidth. In the zoomed view at the top right, it can be seen how the filter slightly changes the position of the area of attention compared to the object position, i. e. the mean of the main component of the filter is not on the mean of the object's position. Although the driver never explicitly fixates the oncoming car during the 2 s of evaluation, it is very important to not put the tracking result on a level with what the driver is actually aware of since peripheral vision is not modeled in this thesis. As long as the oncoming traffic stays on its proper lane and the space for the driver suffices to pass biker and car (so the risk of a collision with the car is very low), peripheral perception of the oncoming traffic might be sufficient. In the presented scenario, the driver concentrates on the object with higher collision risk, namely the biker.

Furthermore, fixations captured by the system can be classified according to the rules from [98]. In this way, the glances to the biker can be interpreted as "task-relevant fixations" while many salient areas (oncoming and parking car, bikes at the bike rack) do not experience direct attention. Plus, the visual sampling is intermittent. The driver does not stare at the biker for over two seconds. The behavior is better explained as "sampling" of the world. According to [98], guiding fixations are interleaved with look-ahead fixations.

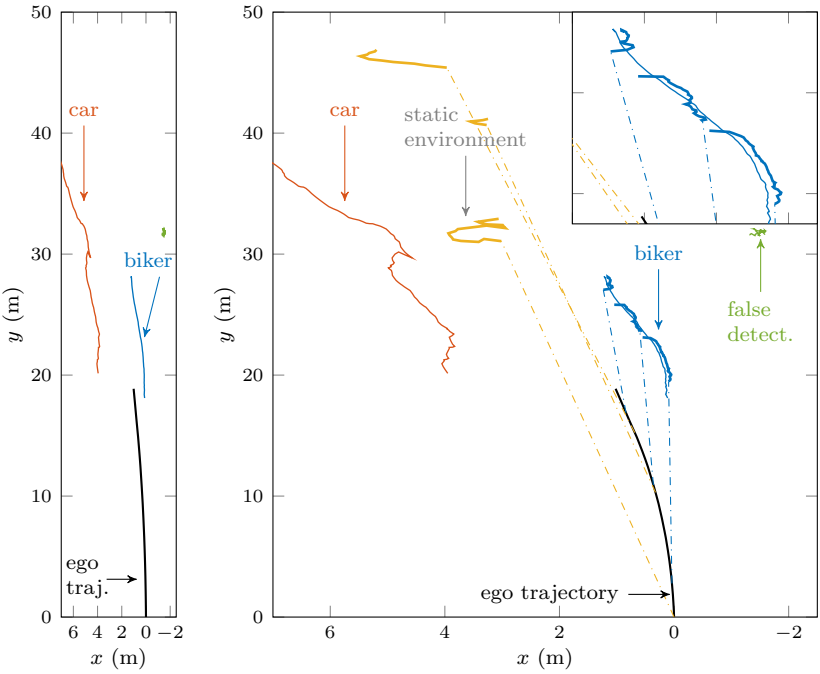


Figure 5.15: The driver's visual sampling of the scene. Left: the scene with equal axis scaling and trajectories only. Right: uneven axis scaling for better visibility. Dashed lines between ego pose (black) and targets represent the time of a gaze shift towards a new target. Thin colored lines show the object's trajectories while the trajectory of the current gaze target has a thicker linewidth. The colors correspond to Fig. 5.11 and Fig. 5.12. It can be seen that the driver fixates the biker and looks towards the front whereas the oncoming traffic is never explicitly fixated.

Additional Examples

To underline the capabilities of visual fixation detection of the proposed tracking approach, additional examples of real world driving scenes are provided in Fig. 5.16 and Fig. 5.17. In both scenarios, the discussed general characteristics of the different gaze target estimation approaches can be observed.

In the first scenario, it is interesting to see that even though all three road users are most of the time within the parafoveal field of view of the driver

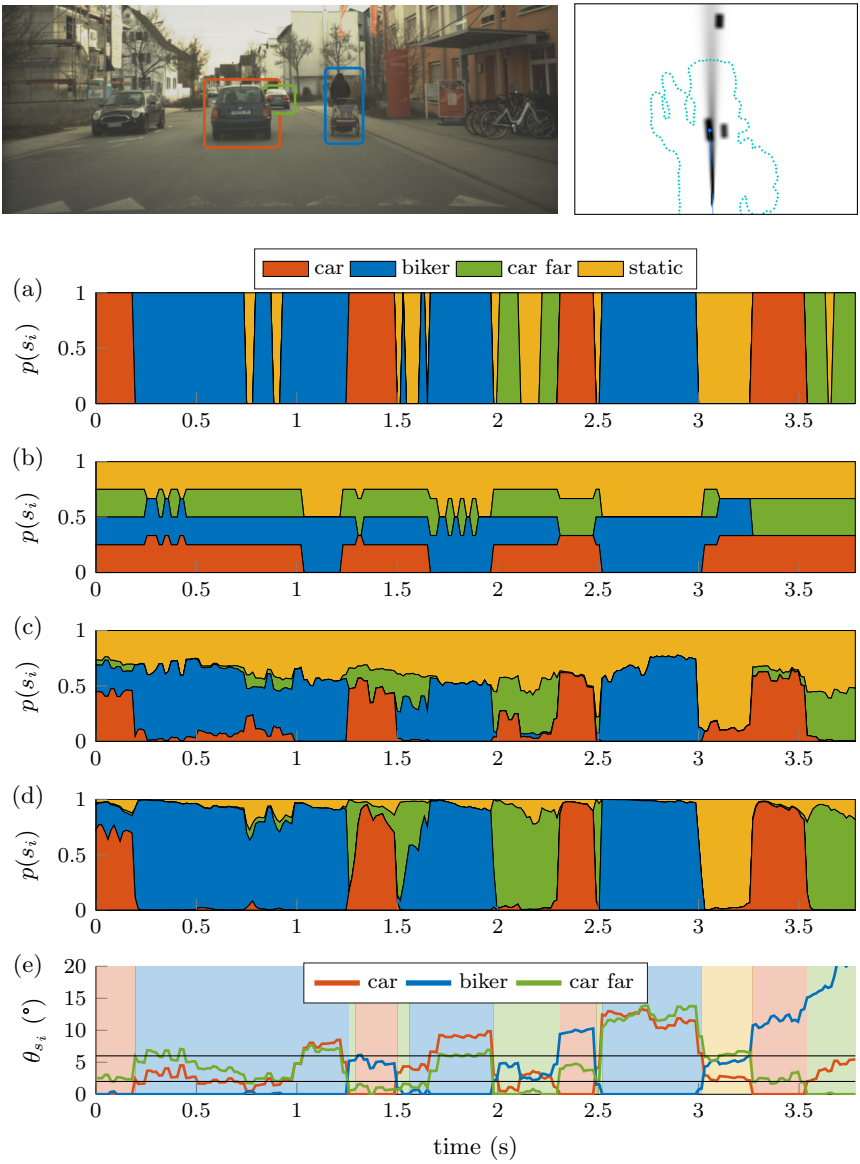


Figure 5.16: Real world scenario 2. The plots show the estimated gaze target for different models. (a) Intersection model; (b) Threshold model; (c) Measurement model only; (d) MHMM-PS Tracking Model; (e) Deviation of gaze to targets.

(cf. Fig. 5.16b and Fig. 5.16e), the gaze tracking approach separates the sequence into clearly distinguishable timespans of different targets. Again, the size of the connected colored areas directly hints at the perception of the specific object. E. g., a clear fixation onto the car in front is only detected at $t = 2\,\mathrm{s}$ (even though this road user is not relevant for the driving scene).

The second scenario shows three vulnerable road users (VRU) in the scene and once more, the weaknesses of common approaches become obvious. With the proposed approach, one distinctive fixation on the child on the bike is detected. In a potential assistant system, the information of when this fixation occurred can be put in relation to the motion of the child, e. g. if the child accidentally leaves the sidewalk. Interestingly again, oncoming traffic can be monitored from the peripheral view as no fixation on the left biker is detected and even does not fall in the 12°-cone.

5.4 Discussion

MHMM and Gaze Motion Assumptions

The incorporated assumptions on gaze behavior lead to a plausible behavior of the filter. Above all, by enabling the filter state to jump from one location to another, a plausible course of modeled probabilities is obtained. But also the time dependent choice of the transition parameters increases the contrast between different objects. Nevertheless, it is difficult to demonstrate the effect of every single parameter in the modeling of gaze motion characteristics. Together with the motion hypotheses of fixations and saccades, the decreased modeled probability of jumping right again leads to an increased confidence in the gaze target after a gaze jump. It is the intention of the tracking approach to increase the confidence in the target, however, not at the cost of false estimates. The obtained confidences of the intersection model and the static multi-hypothesis tracking (MH) in Fig. 5.12a and Fig. 5.12d are higher than in any other model, yet, it is obvious that at least at some time steps the estimate of the target is incorrect. So, of course the modeling assumptions are used to increase the probabilities for road users fulfilling those assumptions. And of course this can also imply higher false positive estimation rates than a simple intersection model. However, it must be taken care that strong but possibly inaccurate model assumptions do not dominate the tracking result yielding incorrect estimates. Some configurations will hardly ever become observable even with highest measurement precision. These include above all fixations

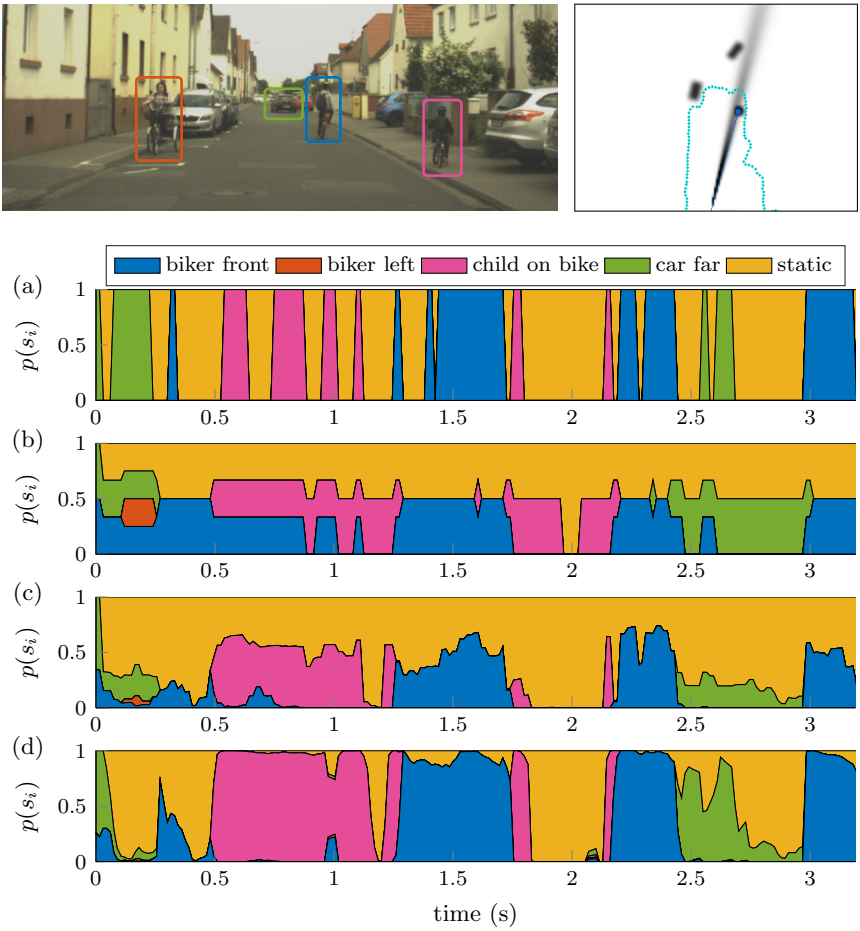


Figure 5.17: Real world scenario 3. Zoomed view of the scene; the car far away is not visible in the bird's eye image. The plots show the estimated gaze target for different models. (a) Intersection model; (b) Threshold model; (c) Measurement model only; (d) MHMM-PS Tracking Model.

on objects far away and the decision between objects which are very close to each other compared to the gaze direction measurement precision. E. g. in Fig. 4.12, it is possible that the driver actually looks at the biker far in front. But if the tracking approach is designed such that it also outputs the biker as gaze target, the same gaze target tracking approach will most probably fail in many other constellations. In the same way, it is important to not choose the model parameters such that the highest possible accuracy is achieved in the static scenario. At the end, however, it stays unclear whether the proposed tracking estimate is truly correct. For this reason, it is necessary to create reference data in order to quantitatively compare different approaches by means of characteristic statistical values. One possible approach is outlined in the following chapter.

Applicability

Firstly, the usage of the proposed method might not always be advantageous. As just mentioned, decisions on fixations on objects far away and between objects which are very close to each other compared to the gaze direction measurement precision will remain difficult simply due to the geometric relations. While simple intersection and threshold models are not directly affected, the incorporation of model knowledge might lead to the inference of the wrong target. Yet, this does not mean that simple models are better. Their outcome just does not depend on model assumptions. As gaze targets are fixated just in time for the relevant subtask [98], fixation detections over long distances are not of foremost importance for the application of the proposed model. The same problem of detecting the object of fixation holds if the scene contains many road users such as a highway scenario with crowded traffic. Since each object close to the line of sight gets some "portion" of the total probability weight, the weight of the actual target is potentially lower. For this reason, it is necessary to use an object list which contains only relevant potential gaze targets. The raw radar object list is a good counterexample since it contains many small clusters originating from all kinds of reflections. Many of these clusters, however, do not belong to relevant targets as they result from poles, fences, etc. Additionally, as just explained, these small objects are even favored due to small assumed variances.

Secondly, also the data precision needs to fulfill certain prerequisites in order for the tracking approach to work properly. This does not only include the gaze measurements but also the environment perception. The consistency of the measurements is of utmost importance as this consistency

of the driver's gaze and the spatial configuration of other road users is the basic assumption of the tracking approach. Imprecise or invalid gaze measurements, false positive or false negative object detections as well as imprecise object location estimates all lead to violations of the postulated consistency. The consequences include lower modeled confidences if the gaze target changes more often or if gaze direction and object location do not coincide. Even more disadvantageous is a higher number of target misclassifications or a high confidence for a wrong estimate, i. e. missed objects are counted as seen and seen objects counted as missed. When selecting the example scenarios from the recorded data, attention was paid to select meaningful situations and care was taken to ensure that the postulated consistency is satisfied in the data.

Fixation Detection and Perception

It was already mentioned that a perception score for each target object could be derived from the size of connected colored areas in the plots of the experiments. In other words, the integral of the probability of fixation over time for each object could be used. In order to only consider true fixations, thresholds of a minimum fixation duration, e. g. 150 ms, and minimum target probability, e. g. 0.5, could be established and only connected areas fulfilling these requirements could be counted. This is only an exemplary proposition but it is clear that the information of "which object is the momentary gaze target" needs to be abstracted to a "perception score" in order to be able to base decisions on the driver's gaze in specifically designed ADAS. The proposed model gives a tool at hand to determine the gaze target but even more than the true current gaze target is unknown, the actual perception of the very same is not observable and can thus only be investigated by evaluating the interaction of the assistance system with the driver.

Remaining Difficulties

Within the proposed tracking approach, a few modeling difficulties remain up to this point. The first drawback of the current model is the discrimination against objects of larger size that was already observable in the examples but which is even more severe for larger objects like trucks. Even though this discrimination is not intended, it is caused by the modeling of object likelihoods as Gaussian distributions. Larger measurement variances for larger objects also imply a larger uncertainty. This increased uncertainty

(a) (b)

Figure 5.18: Reprojection of the point of regard in the ADAS camera. (a) Correct point of regard estimate and thus reasonable reprojection. (b) Contradicting estimate in the image. Due to the lost height information, the estimated gaze target and the reprojection do not necessarily coincide. In both cases (a) and (b), the estimated target is the vehicle in front in the ego-lane.

is then reflected in the estimation by smaller likelihood functions Λ_{ij}. An increased weighting of the object likelihood is not helpful either, as the orientation of the large object within the gaze ray could then lead to an over-weighting of the target. A slightly higher preference of smaller objects is intended, though, as for small objects it is more likely that the gaze ray does not directly hit the target and therefore, stronger model assumptions are needed.

The second drawback of the proposed model is the lost height information by ignoring the gaze pitch angle. Even though gaze samples with larger pitch angles are filtered out, e. g. due to glances to targets inside the vehicle, glances within the "eyes-on-road" region with significant pitch angle are frequent. These glances might then be misinterpreted such as in Fig. 5.18 where a target object is assumed even though the driver might be looking onto the ground. Since only traffic participants are used as potential gaze targets, other regions of interest such as traffic lights, signs or road markings are discarded. Even though these potential targets could in principle be added, the increased number of objects would, as mentioned before, lead to lower probabilities.

Transfer to Alternative Coordinate Systems

As mentioned at the beginning, the general MHMM filter structure can, in principle, be modified to different coordinate systems. E. g. the tracking can be performed in 3D space or in image coordinates. For this, all it

remains to do is to formulate transition and measurement models in the respective coordinate system.

For the 3D vehicle coordinate system, the description can be derived straightforwardly from the 2D case. Reasons why here only the 2D approach was chosen are manifold. Firstly, the principled feasibility has been investigated together with secondly also runtime issues. Furthermore, driving mostly occurs in the 2D vehicle coordinates. That information is stacked vertically happens but is rather seldom, such like traffic lights which are visible for the driver by looking over the car in front of them. Nevertheless, the lost height information was already discussed in the previous paragraph.

For the modeling in image space, the formulation of motion and measurement models might not be straightforward if the reprojection of the gaze ray is used as given by equation (4.1). However, if formulations are found, the interpretability opens up new possibilities due to the rich scene information (cf. Fig. 4.11). In the subsequent chapter, the image space is used in parts for the evaluation of the proposed model.

5.5 Summary and Conclusion

In order to estimate and track the driver's focus of attention over time in a dynamic automotive scene, a Multi-Hypothesis Multi-Model probabilistic tracking framework has been developed in which consistency between machine and human perception during gaze fixations is postulated, i. e. gaze will move consistently with the attended object. Within this framework, spatio-temporal models of human-like gaze behavior for fixations and saccades have been incorporated in the motion transition. This elaborate design makes the target estimation robust and yet flexible to react to saccades. Additionally, each potential target's probability is directly obtained from the probabilistic description which includes the trade-off between several hypotheses. By incorporating dynamic and static potential gaze targets from an object list and a free space spline, the algorithm is in principle independent from the applied sensor setup. The benefit of the proposed model, i. e. the detection of fixations on objects with a confidence score, has been presented on three exemplary scenarios recorded in real world urban traffic. The filter's tracking performance, the effects of different modeling aspects and the driver's visual sampling have been discussed as well as the necessary data quality. However, different methods can only be compared qualitatively against each other as ground truth reference

data of the driver's gaze target and area of visual attention are totally missing. The recording of an evaluation dataset is presumably inevitable for a thorough and meaningful comparison of different methods. In the subsequent chapter, it is aimed at such a reference dataset.

6 Reference Dataset for Object-of-Fixation Detection

As argued in the previous chapter, only qualitative observations and statements about the driver's visual target can be made without proper reference data. In this chapter, a new approach is presented to create such a desired reference dataset which can be used to compare various methods on a quantitative basis. For the creation of this dataset, the remote head eye tracking device in the test vehicle is combined with a wearable gaze tracking device. The first part of this chapter addresses the problem of missing ground truth information for remote eye tracking systems in more detail. The second part is then devoted to the description of the data creation procedure and the thereby encountered difficulties. Finally, the third part presents exemplary evaluation possibilities followed by an evaluation and discussion of different models. Using the reference data, specific characteristics of the data fusion models are pointed out which would be impossible without the reference. The real world dataset was made publicly available and to the author's knowledge it is the first dataset to enable the comparison of different object of fixation algorithms. The main aspects of this chapter have been published by the author in [156].

6.1 Introduction and Motivation

Following the argumentation of the discussion in the previous chapter, it is necessary to develop a strategy to quantitatively evaluate different object of fixation detection algorithms. After all that has been discussed in this thesis up to this point, the series of visual fixations is not equal to what the driver has perceived. An evaluation via questionnaires such as for Level 3 responsibility handover[49] studies and situation awareness assessment is thus unreasonable. There, in driving simulators, the scene is shut down at a random point and the driver is asked for what they

[49] cf. Chapter 1: Level 3 of autonomous driving defines conditional automation. The driver has to expect a responsibility handover call to take back control over the vehicle.

remember [150]. As pointed out in Chapter 2, human perception is highly dependent on the task at hand. Furthermore, even though this aspect has not largely been discussed in this work, humans are able to perceive information from their peripheral visual field and at the same time, they are able to cognitively ignore information within their visual field as shown by experiments for change blindness [165] or intentional blindness [164]. Thus, this procedure would not cover the desired evaluation for two different reasons. Just because something was perceived does not mean it was also visually fixated. It could have been perceived in the peripheral view. And just because something was not reported in the questionnaire, e. g. because the driver could not remember everything, does not mean that they have not been aware of it. They might as well have fixated it. Thus, questionnaires do not give information on what the driver fixated or where they have been looking. However, the need for evaluation goes beyond the assessment of the eye tracking data quality alone. As presented in Chapter 4, different fusion approaches come with different distance functions to be minimized. Even if the eye tracker provides highly accurate gaze direction measurements, different fusion algorithms will also result in different estimates of the point of regard for the same gaze direction vector. The goal is to actually investigate how and why their outcomes actually differ and what error sources are responsible for sometimes bad gaze target estimates. Furthermore, the goal is to assess fixations on objects as well as how well the extracted point of regard is compared to where the driver truly looked. This is visualized in Fig. 6.1. However, up to now, there exists no possibility to compare different fusion approaches or verify their performances on a quantitative basis simply because there is no reference available. Model comparisons have only been made qualitatively or not at all and model evaluations have mostly been done through reasoning and argumentation. The previous chapter focused on object of fixation detection in real world driving scenarios. This chapter now aims at providing an idea of how such model outcome can not just qualitatively be described but quantitatively evaluated.

6.2 Problem Statement

6.2.1 Ground Truth for Visual Attention

The information of where and at what the driver is looking in the surrounding environment can be extracted as fixed [38] or dynamic [159] region in

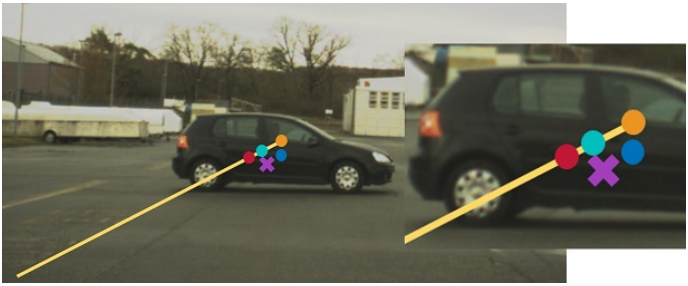


Figure 6.1: Scene image with the reprojection of the gaze ray (yellow) as measured by the remote eye tracking system, the measured points of regard of four different models (blue: model from Chapter 5, red: stereo intersection from [91], orange: infinite point on gaze ray, cyan: reprojection of the intersection of gaze ray with the radar object in 2D vehicle coordinates) as well as the reference point as measured by a wearable gaze tracking device and labeled manually (purple cross). In general, such ground truth label is unavailable.

vehicle space, object instance [11, 116, 158, 182] or pixel based region in the image space [5, 91, 96, 137, 200], cf. Chapter 4. From this collection, remote gaze estimation methods are used in [11, 38, 91, 137, 158, 159, 200] while in [5, 96, 116, 182], wearable *eye-tracking glasses* (ETG) are employed. However, when estimating the driver's point of regard, i. e. the point the driver looks at, all approaches share the problem that there simply exists no ground truth on where the driver is truly looking. This is due to the uncertainties of the gaze direction estimation, the sensor calibration precision and the fact that sharp vision is rather a cone with small opening angle than a single ray. Plus, there is no label telling the precise gaze location. For wearable eye trackers, these issues are often neglected due to the higher accuracy[50]. Since all of the mentioned works with *remote* head-eye-tracking system are exposed more or less to those named issues, there also exists no ground truth on how well a certain inside-outside sensor fusion model performs in a given hardware setup. This was identified as one major problem in [96, 158] and it is the reason why rich tolerance thresholds are often included in the gaze target or focus of attention estimation such as in [11, 96, 137, 200].

[50]For comparison: the accuracy for the wearable system used in this work is given as 0.5° in the data sheet of the device whereas the accuracy of the remote system is at about 2° for the test person (cf. Fig. 6.8).

6.2.2 General Problem of Ground Truth

Creating a suitable ground truth is hard to obtain in merely all domains. It often involves the need for measurements of higher precision which in turn leads to a second measurement system running in parallel. A simple analogy can be found in visual localization or visual odometry (VO) of which the KITTI dataset and benchmark [52] is a well-known example in the automotive research community. In order to evaluate VO results, it does not suffice to record images, run one's own VO and claim superior performance. Only with the high precision reference data based on separate measurement principles, qualitative comparisons of different models are possible. The success of the KITTI benchmark underlines the problem of the effort often associated with recording reference data as many people rely on existing benchmarks instead of recording an own ground truth dataset. In other cases, ground truth is obtained with expert knowledge or a large amount of manual annotations. Both involves high effort as well. The cityscapes dataset [30] is one example for tedious manual labeling. In the domain of driver monitoring, fatigue is a good example of a concept that is hard to grasp where often human expert psychologists are consulted or different methodologies like electroencephalogram (EEG) analysis or percentage of eye-closure (PERCLOS) are evaluated against indirect measures like driving performance.

In gaze region estimation, ground truth often comes from experts. One approach from Martin and Tawari [116] includes expert knowledge when determining the object of fixation. In contrast, Alletto et al. [5] found a way to annotate their DR(eye)VE dataset automatically by means of image feature matching and homographic transformations, however, the area of attention is only given as pixel region and not related to road users. Both approaches use wearable eye-tracking glasses and therefore, the dataset from [5] cannot be used as reference dataset in this work. These examples yet also show that the term ground truth does not mean absolute truth. Even carefully recorded or annotated datasets are exposed to uncertainties. In order to minimize such uncertainties for the special case of investigations and evaluations of driver visual behavior with respect to in-vehicle information and control systems, the international norm ISO 15007 [71] has been formulated for this specific field. In contrast to the goal of this chapter where the precise gaze location and target is of interest, the formulated guidelines mainly relate to predefined areas of interest and dwell times.

[51] Example image taken from https://www.cityscapes-dataset.com/examples/.

(a) KITTI VO [52], image taken from [25].

(b) Cityscapes [30][51]

(c) Dr(eye)ve [5]

Figure 6.2: Images of three different datasets in the automotive field. Creating reference data always necessitates comparatively high effort in recording and/or labeling. Images taken from the named sources.

6.2.3 Error Sources of Remote Object of Fixation Detection

Fusing the data of remote eye tracking with the information of environment perception sensors is exposed to several challenges. These include particularly an unknown time synchronization error which is very hard to identify without reference or specific stimulus. Furthermore, it can have large effects on the outcome of a fusion algorithm. In the worst case, a vulnerable road user is classified as seen even though they were missed. The same principal problem holds for an unknown total calibration error[52], both extrinsic and intrinsic of all cameras employed and of possibly other measurement systems, e. g. radar or lidar. Since this error leads to a certain bias of the estimate, it can be found more easily than a time synchronization error. A third difficulty is given by unknown shape and magnitude of the measurement noise of the gaze direction measurements. This does not only imply that variance and bias are unknown in magnitude and direction but also

[52]Precise calibration is inevitable for associating gaze with specific objects.

that the measurement noise could be direction dependent and might not follow a parametric distribution. Most of all, the distribution of the gaze direction error can be different for different test people. Further potential error sources such as time dependent extrinsic system calibration changes due to vehicle roll and pitch motion in real traffic, engine vibrations, and changing intrinsic camera calibrations due to the cameras' temperatures are mostly neglected[53].

One more error source which shall be named here but is not part of this work is that different people are differently qualified for high quality eye tracking. As certain anatomic differences will probably always pose difficulties [67], any system intended to reach series maturity is required to be tested with a large amount of people. As this work outlines an idea of how the evaluation of matching algorithms can be done, the problem of scalability is out of the scope of this work and only shortly addressed in the conclusion at the end of the chapter.

6.2.4 Proposed Method for Reference Data Recording

The performance identification of an eye tracker itself is a topic on its own [67] and it is often done in ideal environments. Yet, even if measurement confidences are determined, these can vary for different setups and might not hold for a setup in a car. Therefore, it is necessary to assess and evaluate a system in its final configuration. For this, calibrated 3D points on the windshield are used and it is determined how well gaze measurements of a test person fit (see Fig. 6.8c). This procedure can be applied in order to determine for which subjects the eye tracking device works suitably but it does not provide any measure on different fusion algorithms. Also, it would be better to test for certain points in the environment with characteristic depths of real traffic. However, their precise location is hard to obtain. Nevertheless, the evaluation of the eye tracker's performance can not assess the quantitative benefit of different fusion algorithms or tracking model assumptions. Thus, rather a way to compare different algorithms needs to be found. For this purpose, there are presumably only three options. The first idea is based on experimental setups with instructions to the test people on what to visually fixate. This suffers from the difficulty that gaze behavior is partially driven by bottom-up factors meaning that other stimuli can unconsciously affect the gaze motion. Therefore, it will not be

[53] In automotive applications, these can be assumed to have indeed a smaller effect than time synchronization or a calibration bias. For applications on a motorcycle however, these can become crucial.

guaranteed that test people actually follow these instructions and thus, the obtained reference data is not necessarily reliable. Such an approach was chosen in Chapter 5 in the static scene. In fact, the measured gaze ray does not intersect the respective target at all times and it is unknown whether the reason are imprecise measurements or small changes in gaze direction, e. g. due to microsaccades. Furthermore, this approach is impractical and dangerous for real traffic situations which is why "natural" gaze behavior can not be obtained with this method. The second idea is to create reference data in a simulator setup. This has the advantage of generating controllable scenarios from dynamic traffic situations without any risk. In a simulator though, model assumptions and algorithm performances can hardly be checked since the gaze target is determined by the distance between road user and gaze point on the screen. The potential to estimate depth, however, is one of the main aspects of real world inside-outside matching approaches. In order to provide such possibilities, the simulator environment description would need to be fused as 3D description with the driver's gaze. Nevertheless, even such a setup would not provide a reference for the driver's gaze position. The third option is to use another measurement device with a higher accuracy in parallel which is described in the following.

Even though high performance remote head eye trackers can be used as reference for lower quality systems[54], whenever precise information about the driver's gaze is desired, wearable eye tracking devices are used. A few examples include [5, 84, 98, 116]. In these cases, the measured data is taken as reference without considerable post-processing. This leads to the assumption and proposition that a wearable device with subsequent manual labeling could be used to assess the model quality of multi-camera remote eye tracking object-of-fixation detection algorithms by running those two systems at the same time. The idea behind this approach is that the wearable eye tracker has a higher accuracy and precision than a current state of the art inside-outside matching system of a remote eye tracking device in combination with other vehicle sensors. It is important to note that this proposed evaluation method is unable to assess the quality (precision and noise) of the remote eye tracker in terms of angular deviation between true and measured gaze direction. In contrast, the focus lies on the numerical comparison of different gaze target detection and gaze region estimation models. The next section presents the data collection

[54] E. g. a multi-camera and multi-lighting setup can be used to measure the performance of a single-camera or RGB-camera system.

and processing in more detail.

6.3 Reference Data Generation

6.3.1 Test Setup

The setup under test consists of the PRORETA 4 test vehicle equipped with state of the art series sensors for scene perception, i. e. a long range radar and a stereo camera. Additionally, a four-camera *SmartEye 6.2* head-eye-tracking system is employed, that captures the driver's gaze motion at 60 Hz. Depending on the test person, an accuracy of a few degrees ($\sim$ 1-4°) is reachable (see Fig. 6.8b). The sensor system is fully calibrated which means that the eye tracker's coordinate system origin is placed in the same position as the vehicle coordinate system at the middle of the car's front axle[55]. All cameras including the exterior camera are intrinsically and extrinsically calibrated. Furthermore, the system is time synchronized. Thus, all sensor measurements including but not restricted to gaze measurements, object list, and exterior image flow together in one network system. Fig. 6.3 shows the test vehicle and its setup. With this setup, it is possible to extract the gaze target at object level as shown in [158–160]. However, performance measurements and quantitative results regarding the precise gaze location are missing in these works where only qualitative argumentation was provided.

6.3.2 Verification Setup

In order to create reference data, it must be guaranteed that the gaze quality of the remote eye tracker is not affected too much by the additional measurement device. The remote eye tracker uses images of the eye, pupil and IR corneal reflections which is why the wearable device is allowed to occlude parts of the face but not the eyes. On the other hand, a system is needed whose precision and accuracy are expected to be higher than the one of the remote head-eye-tracking system. This can be achieved by selecting a suitable wearable device. It was found that the *SMI ETG 2* shown in Fig. 6.4, also used by [5, 96], fulfills these requirements[56]. The device has a frontal camera capturing the driver's field of view at 30 Hz and two

[55] cf. Appendix A for further information on the system calibration.

[56] Many thanks to Marius Hofmann from FZD, TU Darmstadt and Herbert Janssen from the Honda Research Institute Europe GmbH for lending their SMI ETG 2 for this work.

(a) Test vehicle quipped with radar and stereo camera.

(b) Multi-camera eye tracking system from SmartEye (2 out of 4 cameras visible in the image) installed in the vehicle's interior.

Figure 6.3: Test vehicle sensor setup.

inner cameras capturing the driver's eyes and gaze relative to the glasses at 120 Hz. The used output are the gaze point coordinates within the exterior images. These gaze point coordinates are then transferred in the image of the ADAS camera (see Section 6.3.5). During the search for an appropriate wearable eye tracking device, also the first generation Dikablis eye-tracking glasses have been tested which, however, proved to be impractical since the eye-camera occluded the driver's left eye at most times. As the eye-tracking glasses are not connected to the measurement framework of the test setup, time synchronization is of high importance. It is done via a flashing light signal induced by a flashlight visible in the exterior cameras of the vehicle as well as the eye-tracking glasses. From the time stamps, the offset parameter is determined and drift is verified by a second light signal at the end of each sequence. Even though the test setup is principally the same, i. e.

Figure 6.4: Wearable tracking device from SMI with removed IR filter lenses.

cameras targeting the driver's eyes and one facing outward, a wearable device in general has lower errors. This is because exterior camera and eye camera are much closer together and closer to the eye which results in smaller errors due to image resolution or calibration.

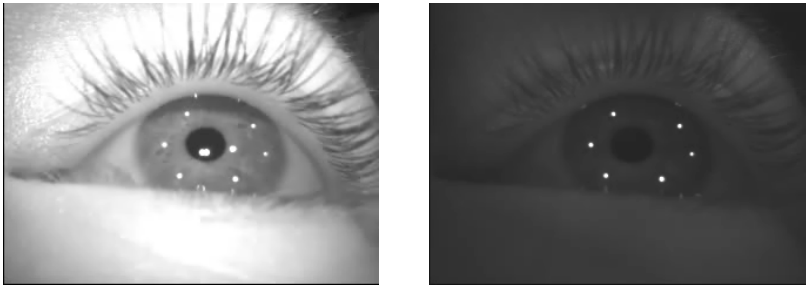
6.3.3 Individual Calibration

Both gaze tracking systems must be individually calibrated to each user (cf. Section 2.1.2). The individual calibration of the remote device is performed by letting the test person glance at known 3D points in the car and on the windshield. The SmartEye software determines the deviation from the known and measured directions and incorporates this knowledge directly in the gaze tracking. Different, but also known 3D points are used to verify the eye tracking performance.

In case of the wearable device, highest possible accuracy is of foremost importance since the data obtained is later used as ground truth. The calibration procedure of the wearable eye tracker consists of looking at three points on one even surface at a distance common to the field of use, i. e. for urban driving, buildings at distances above 10 m are used. Here as well, the calibration result is directly incorporated in the tracking output of the eye-tracking glasses.

6.3.4 Joint Usage

Since both systems rely on active IR illumination, their joint usage is not straightforward. Each device projects IR light onto the driver's eyes and computes the gaze direction from the corneal reflections on the eye which are detected by the respective tracking device. As can be seen in Fig. 6.5,



(a) (b)

Figure 6.5: Images of the wearable eye tracking device when used jointly with the remote eye tracker. (a) The two additional corneal reflections due to the flashes of the remote eye tracker are visible within the six reflections from the wearable device. Also, the image is much brighter compared to the right image. (b) Since the wearable device is working at 120 Hz while the remote is working at approximately 60 Hz, brighter images are followed by darker images without reflections from the remote device. The tracking algorithm of the wearable needs to get along with these challenges.

the light from the eye-tracking glasses is sent out constantly while the light from the remote head-eye-tracking system is pulsed. This leads to changing illumination and up to two additional corneal reflections within the image of the wearable. In the experiments, it seemed that these factors did not deteriorate the tracking performance as opposed to the following issue: the eye-tracking glasses must of course be used without the IR filter lenses which come with the device and whose purpose is to filter out disturbing IR light. The IR illumination of the SmartEye system would be filtered in the same way. It was quickly noticeable that data recording is difficult on sunny days or at weather with much stray light due to the increased IR noise emission.

On the other hand, the SmartEye system showed no problems with the light from the SMI since always two images are taken, one with and one without flash, so that interrupting IR light can be subtracted. This effect is shown in Fig. 6.6. However, even though the SMI's frame is large, it is also quite thick and can occlude the eyes which needs to be considered. In conclusion, the later evaluation can of course only be performed at time instances where both systems produce reliable results. These are recognizable by high values of the SmartEye confidence signal and the existence of tracking values for the SMI.

Figure 6.6: Image of the remote eye tracker when used with the wearable device. Note that there are no corneal reflections from the wearable device visible since the remote eye tracker compensates IR noise by taking two images with and without IR flashlight respectively. The large frame of the eye-tracking glasses is advantageous compared to other wearables since it occludes the eyes only at larger head rotations.

6.3.5 Annotation

In the labeling process, for each gaze sample from the SmartEye system, the sample closest in time from the wearable eye tracker is transferred to one gaze pixel position in the firmly mounted ADAS camera. In order to be as precise as possible, the labeling is done manually, not by a homographic transform as in [5] or [116]. Furthermore, objects of fixation are labeled based on the object list. For evaluations in the image space, the target category is as well labeled based on the semantic segmentation of the image[57,58] even if the object of fixation does not exist as reliable object in the object list of the vehicle. However, problems remain in areas with low texture or in special cases of parallax when the gaze target is occluded or unclear.

[57] The tensorflow implementation of PSPNet [201] available on github https://github.com/hellochick/PSPNet-tensorflow with the model trained on the cityscapes dataset is used without further training on the images from the test vehicle camera.

[58] Calculations of the semantic image segmentation for this work were conducted on Nvidia GPUs on the Lichtenberg high performance computer of the TU Darmstadt.

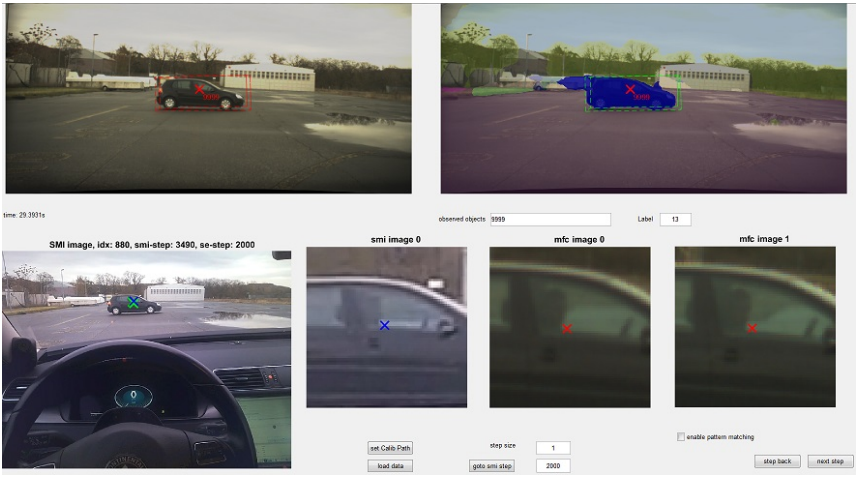


Figure 6.7: Screen shot of the labeling GUI: image of the wearable eye tracker with point of regard on the bottom left and images from the ADAS camera in the top row. The user gets a suggestion for the point of regard in the ADAS camera and can correct the pixel position in the magnified excerpt. The object of fixation label is automatically created if the point of regard lies within the bounding box reprojection of road users in the object list. The object type is inferred similarly from the semantically segmented image.

In Fig. 6.7, a screenshot of the labeling graphic user interface (GUI) is presented. The human annotator simply has to transfer the pixel position. Reprojections from traffic participants in the object list are highlighted and the semantically segmented image areas are processed in the background so that object of fixation labeling is simplified. In case that the measurement of the wearable device reports a small deviation to a clearly fixated target, the fixated object is still labeled as such. It is up to the human annotator to decide on fixations. Similarly, when the object is clearly hit but the position estimate of the radar object is not precise, i. e. the label does not fall into the bounding box of the object, the object is still labeled as gaze target.

6.3.6 Dataset

For each recording, two separate data streams of one test person wearing the eye tracker while also being remotely tracked have been recorded. The

data set consists of several short scenes on a test track with stationary ego-vehicle and only one single gaze target plus one recording of real world driving of about 5 min[59]. The real world recording has been divided into meaningful sequences, each a few seconds long. The obtained reference data is visible in the top right image of Fig. 6.7 consisting of the point of regard's pixel position in the vehicle's camera and the labels of the current gaze target as given by the vehicle's object list and the semantic image information.

Data Quality

In order to verify that for the test person the quality of the remote eye tracker is not severely affected, a quality assessment[60] is performed for known 3D points in the car with and without the eye-tracking glasses. The results are shown in Fig. 6.8. It can be seen that the eye tracking quality is within the same range for both setups. From this, it is concluded that in principle, the wearable device does not impair the tracking quality of the SmartEye system for the test person. At the same time, the wearable device has a tracking coverage of above 98 % in all of the presented recordings[61], so the tracking performance of the wearable system is not substantially affected either.

6.4 Experimental Results

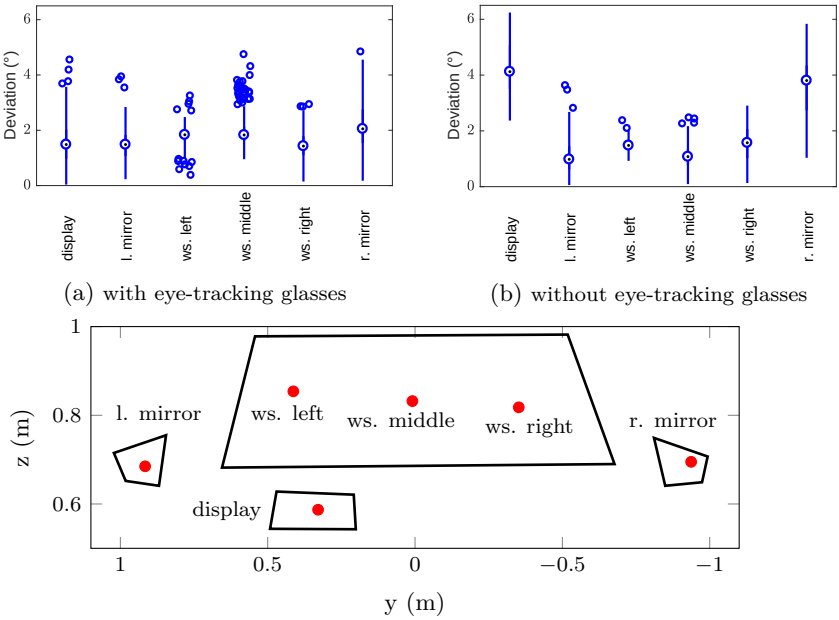
6.4.1 Models to Compare

As motivated, the new reference data serves as ground truth for different algorithms that deal with object of fixation detection or area of attention estimation. The following models are compared to each other. The first model is the already known simple intersection model (abbreviated here with 'I') where the 2D point of regard in vehicle coordinates is given by the

[59] The recorded dataset has been made publicly available in separate sequences at https://www.proreta.tu-darmstadt.de/p4data. It includes the gaze direction and gaze confidence measures as well as the ADAS camera image sequences and object lists from the test setup. From the verification setup, video sequences and gaze point coordinates are included.

[60] Thanks to Manfred Wilck for providing the quality assessment in the PRORETA 4 test vehicle.

[61] This holds for all sequences which are used in this thesis. Parts of the test track recordings have lower coverage of the wearable device due to more IR stray light on the day of recording.



(a) with eye-tracking glasses

(b) without eye-tracking glasses

(c) Arrangement of the quality points in the car

Figure 6.8: Quality point measurement of the remote tracking system with and without the wearable device. It can be seen, that the quality of the remote eye tracker did not suffer due to the wearable. However, the glass frame can of course hamper the tracking.

intersection of the gaze ray with the radar objects or the free space spline. It can again be seen as baseline for the Multi-Hypothesis Multi-Model filter from Chapter 5 (here, only the MHMMPS version with reduced sampling and gaze behavior modeling is used, for more simplicity abbreviated in this chapter with 'M'). The threshold model does not determine one specific gaze target point which is why it is not directly reasonable to compare it to the set of above mentioned models. However, in Section 6.4.6, the threshold model is picked up again in the comparison of gaze targets. In the set of the models to compare, only the two models I and M compute an estimate of the gaze target from the radar object list. The other models are purely image-based and without any object tracking within the image sequences. The first of these two models is just the reprojection of the point at infinity

on the gaze ray. Since it is a heuristic approach, the model is abbreviated with 'H' and it serves as baseline for the stereo intersection model ('S') presented in [91] and described and depicted in Section 4.2.2 and Fig. 4.11. Using this model, the depth estimate and the pixel position are obtained from the reprojection of the gaze ray and stereo images. Since the label is given in image coordinates, the reprojections of the point estimates from I and M are compared to the image based approaches H and S.

6.4.2 Selected Scenarios

For this evaluation, two artificial scenarios and one short snippet of the recording from real world traffic are picked out in order to highlight some characteristics of different gaze target estimation approaches and show possible comparisons[62]. In the first experiment, the gaze target crosses from the left to the right. Similarly, in the second experiment, the gaze target first recedes about 100 m, turns around and then approaches again, simulating longitudinal traffic moving in both directions. In both trials, the test person is asked to always fixate the gaze target, however, without specified fixation location on the target. The third scenario is in real traffic with a bike rider appearing from the right at the upcoming T-junction. Scene shots from all three scenarios are shown in Fig. 6.9.

6.4.3 Evaluation Criteria: Statistical Measures

When evaluating the object of fixation performance, appropriate measures need to be specified. Most importantly, objects should not be counted as seen, when they are not. In a two class classification scheme, the *false positive rate* (FPR) or *fall out*

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{6.1}$$

where FP denotes the false positive and TN the true negative classifications, is an adequate measure when the cost of FP is high. Especially when the respective object is never fixated, the FPR is a reasonable choice. Yet, in this case, even small values of the FPR are to avoid because most of the

[62] The quantitative results presented in this chapter deviate from the results presented in the original publication [156]. Almost all statistical values for all compared models are higher in this evaluation. This is due to an improved time synchronization and renewed labeling leading to an improved data consistency.

(a) crossing experiment

(b) longitudinal experiment

(c) Real world driving scene

Figure 6.9: Scene shots of the three evaluated scenarios.

time that an object is observable in the sensors' field of view, a driver is not directly looking at it yielding a large number of TN.

Secondly, objects should not be counted as missed, if they have been clearly fixated. However, it is not most important since current ADAS without driver monitoring always assume this case of an inattentive driver. The *true positive rate* (TPR) or *recall*, given by

$$\mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{6.2}$$

where TP denotes the true positive classifications and FN the false negatives, is appropriate when high cost is associated with false negatives. Recall measures how often the true gaze target is actually detected during a fixation.

In contrast to the FPR, recall can only be obtained if the respective object is looked at at least once so that TP + FN $\neq 0$. Since, as mentioned, the number of true negative classifications can be very high even if the object is looked at, *precision*, given as

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{6.3}$$

can be used as measure of the number of FP related to TP if TP + FP $\neq 0$. Precision is a measure of how many of the predicted positives are actual positives and it decreases for each sample where the model predicts the object as target even though it is not. The reason to use precision is also to compute the F1 score

$$\mathrm{F1} = 2\frac{\mathrm{PrRe}}{\mathrm{Pr} + \mathrm{Re}} \tag{6.4}$$

as a measure to balance precision and recall and thus also the cost of FP and FN in one single number. The often applied accuracy, given by

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{6.5}$$

might not be the best measure to balance the cost of FP and FN due to the already mentioned potentially high number of TN leading to high accuracy even though a model might exhibit low precision and recall scores.

In the following investigations, the statistical values of recall, precision and F1 are computed for the object of fixation classification of the models in vehicle space and also for the class labels in image space of the respective gaze targets. Due to the experiment design, TP > 0 holds for the artificial scenarios.

Additionally to the statistical classification scores, the computed point of regard is compared to the respective label in image space. Since the label is given as pixel position in the exterior image, the distances between label and model outputs are reported in the form of means and variances.

6.4.4 Applicability of Reference Data:

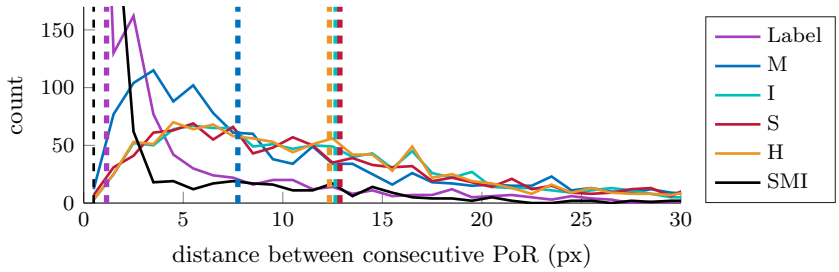


Figure 6.10: Distance between two gaze points for the SMI recording and the label compared to the models for the first experiment (100-bin histogram) with a crossing vehicle. The median distance for each model is plotted as dashed vertical line in the respective color.

Median Distance of Models vs. Label

One might argue that using the manually labeled pixel position as ground truth observation is not straightforward due to a potentially induced labeling error. Besides the fact that wearable eye trackers like the one used in this work are current state of the art for studies in psychology and ergonomics, another argument is shown in Fig. 6.10. During visual fixations, the angular gaze motion and thus also the pixel distance of the point of regard between two consecutive time steps is small. Additionally, fixations last longer than saccades so that the median pixel distance is a measure of the jump size within a fixation. If the median distance of the label is much smaller than those from the models, then it can be assumed that the larger jump size results from measurement errors. Thus, the label will have higher accuracy than the models and can therefore serve as ground truth. This relationship can in fact be seen in Fig. 6.10. The median distance of the point of regard in the SMI recording is below and the median distance of the label is close to 1 px. This reflects a typical characteristic for visual fixations whereas all results from the remote eye tracker exhibit comparatively large distances. From this, it is assumed that the label can be used as reference.

6.4.5 Results in Artificial Scenarios

Even with labeled reference data, the rating of different models is cumbersome and the results not always evident. For this reason, this subsection

not only assesses the models by their statistical score but also tries to unveil potential causes for various behaviors.

Crossing Experiment

In this experiment, one gaze target at a distance of about 20 m to the driver is crossing from left to right. The driver is asked to fixate the car at all times, however, without specifying the exact location on the car. Table. 6.1 shows the statistical values for each gaze target estimation approach. The first column reports the measures of the fusion in image space, e. g. the estimation result at each time step is counted as TP if the semantic class label as well as the pixel label at the computed point of regard in image space are 'car'. The second column presents the statistical scores for the two models I and M in vehicle space. And the last column shows the mean distance of the estimated point of regard to the label position and the standard deviations in both directions of the eigenvectors of the covariance ellipses, i. e. the ellipses' major axes, in pixel. The third column becomes more comprehensible with Fig. 6.11. There, the deviations of the estimated point of regard to the respective labels are plotted as single data points. Therein, the label position is always centered at the origin and the models' deviations are drawn relative to the label position together with their respective mean deviation and their respective 50 % covariance ellipses[63] plotted with dashed lines. The plot is enhanced by two red circles indicating approximate angular deviations of 1° and 2°[64].

Table 6.1: Precision, Recall and F1-Score of the class-label 'car' in image space and of the object of fixation detection for the models in vehicle space. The third column shows the mean distance of the estimated point of regard to the label position and standard deviations in direction of the eigenvectors of the covariance ellipses in pixel.

	class			obj. of fix.			pixel dist.		
	Pr	Re	F1	Pr	Re	F1	∅	σ_1	σ_2
H	0.98	0.89	0.93	-	-	-	36.4	31.7	24.5
S	0.98	0.91	0.95	-	-	-	34.2	32.0	25.8
I	0.98	0.94	0.96	0.98	0.98	0.98	35.1	33.8	24.0
M	0.98	0.94	0.96	0.98	0.95	0.97	36.0	33.8	24.1

[63] 50 % are chosen simply due to a better visibility compared to larger percentiles.

[64] The pixel deviation itself can only serve as comparison between different models. Only by providing a more general reference, the results can also be compared to setups

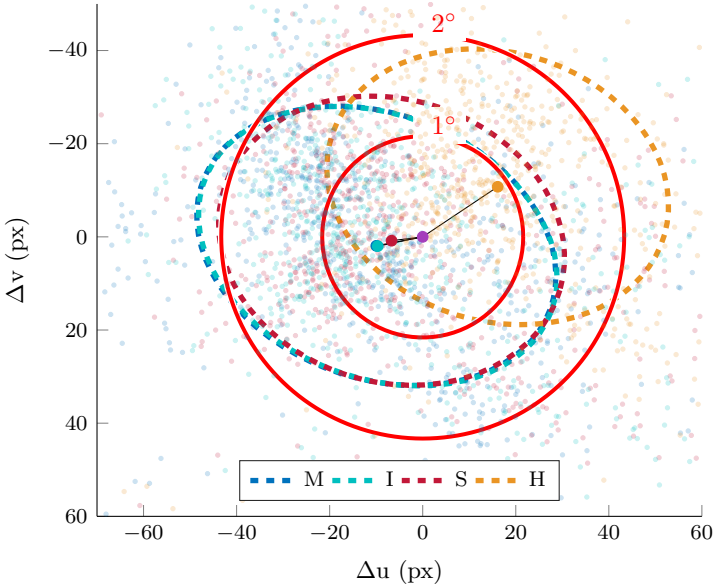


Figure 6.11: Deviation of the estimated point of regard for different models. The figure shows the mean position in relation to the label as opaque circle and single data points relative to their corresponding label as transparent circles. The ellipses are the 50 % covariance ellipses. The red circles indicate approximate angular deviations of 1°and 2°. The mean position of the tracking model is plotted below the dot for the intersection model.

High statistical scores for all models show that this experiment is comparatively simple. All gaze target estimation methods are able to reliably determine the correct fixated object. Only the heuristic model has a slightly lower recall score, indicating that it missed the image area with the car for a few samples in the recording. This is in accordance with the deviations shown in Fig. 6.11. In this simple configuration with the object being not too far away and moving slowly, the covariance ellipses of all models have similar size. The three models which estimate depth also exhibit a similar direction of the deviation while the error of the heuristic model is towards

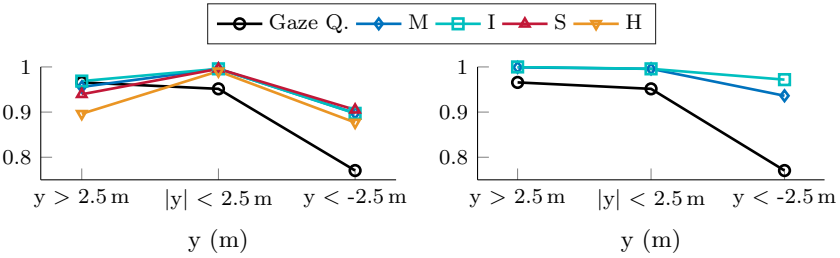
with different cameras. The focal length of the used camera is 1240 px which means that 22 px correspond to roughly 1° deviation and 43 px to 2°. However, the deviation between label and model estimate can as well stem from a wrong depth estimate. Thus, the two red circles are only indications.

the upper right. Due to the characteristics of projective geometry and the camera position related to the driver, the heuristic model output lies always higher and more to the right than the stereo output.

The high statistical scores for all models show that the correct semantic class was predicted by all models for most of the gaze samples. This leads to the assumption that the remaining errors or the point of regard position in the image stem from errors, respectively deviations, of gaze direction measurements. The magnitude of the observed deviations is similar to the deviations in the quality point check above in Fig. 6.8a. Error sources of calibration or time synchronization can be ruled out with high probability. If the error came solely from the extrinsic calibration, the ellipses would be narrower. Furthermore, the time synchronization was done carefully which is also visible in the following Fig. 6.13c.

Interestingly, the precision is exactly 0.98 for all models and for the semantic class prediction as well as the object of fixation classification. For a few samples in the scene, the label is not on the car leading to a small and for all models quite similar number of false positive classifications. It can be suspected whether the measured point of regard of the SMI gaze tracker was slightly imprecise. This shows that the scores of course also depend on the label and the reference data quality as mentioned in Section 6.2.2.

Why is the simple model better than the complex model? Even though the differences are only minor, it is investigated why the simple intersection model performs slightly better regarding the recall and F1 score compared to the complex tracking module which is designed such that the expectation of the gaze region follows the object's motion. Trying to answer this question, first the F1 scores are computed based on the lateral gaze target position presented in Fig. 6.12. It can be seen that the tracking performance decreases towards the right of the scene, i. e. for negative values of y. Since both, intersection and tracking model should not be affected by these geometric relations, neither in image nor in vehicle space, the average gaze quality signal is investigated. The gaze quality signal drops as well when the car drives towards the right side. While the intersection model is not directly affected by a lower quality signal, this can be a hint for jumps in the gaze direction signal. Therefore, the already known graph types of the gaze target probabilities and the angular deviation of the gaze measurement to the target object are plotted in Fig. 6.13a and 6.13b. In Fig. 6.13b, the background color indicates the



(a) Semantic class prediction (b) Object prediction

Figure 6.12: F1-Scores for different lateral distances of the gaze target in the crossing scenario. The object prediction is only computed for the models in vehicle space. The additional black data points denote the average gaze quality as obtained from the SmartEye system for the respective segments.

gaze target of the tracking model. First of all, it can be observed that a few gaze jumps away from the object have been measured leading to predictions of the static environment. But also at time instances, where the angular deviation of the gaze ray to the object is zero, the tracking model predicts the wrong target. One such misclassification occurs around $t = 6\,\mathrm{s}$. In Fig. 6.13c, the horizontal pixel coordinates for the label, the intersection model and the tracking model are plotted. At $t = 6\,\mathrm{s}$, there is a jump observable in the label and the intersection model. At the same time, the angular deviation remains zero indicating that the gaze jumped from one point on the object to another. Thus, the gaze target stays the same but the gaze direction jumps. This jump leads to a short change of the gaze target of the tracking model and thus to a FN prediction. This happens because in this specific setup, the static free space behind the car obtains measurement evidence as well, which is why the gaze behavior assumptions are ineffective. It is important to note that the present scenario does simply not reflect natural human gaze behavior. After a short period, the probabilities of car and free space reach a state of balance as can be seen in the zoomed view in Fig. 6.13a. Furthermore, the mentioned lower gaze quality also leads to false negative classifications in the last thirds of the recording. Visible by the pixel position jumps of the intersection model, the jumps in the yaw direction measurements violate the tracking assumptions made in Chapter 5 affecting the tracking model's results. The intended behavior of the tracking model is visible in the middle of the recording. Here, the gaze position of the tracking approach is smoothed compared to

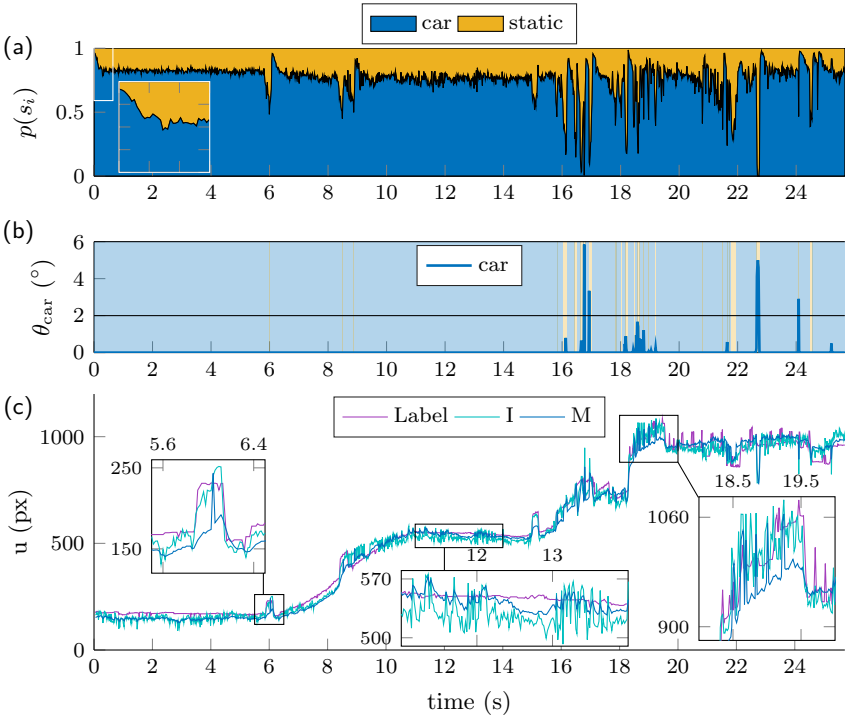


Figure 6.13: Time course of the crossing scenario with (a) the modeled probabilities from the tracking model, (b) the deviation of the measured gaze to the gaze target, and (c) the reprojected horizontal pixel position of the estimated point of regard for the intersection and tracking model compared to the labeled position.

the basic geometric model and does predict the correct target.

Point of Regard Depth Estimation with Stereo Model The 3D gaze points of the models I and S are further compared since the intersection model exhibits also a slightly higher recall score in image space than the stereo model. At first, this seems counterintuitive. However, this is due to several problems of the stereo model. In [91], it is claimed that the stereo method can extract the 3D point of regard. Even though this is formally correct, the precise extraction of the 3D point of regard is highly dependent on very accurate disparity computations and gaze measurements which

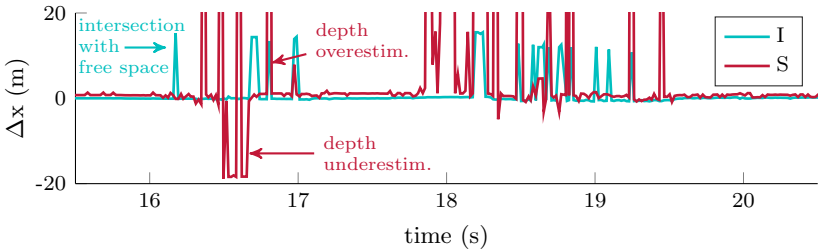


Figure 6.14: Error in the depth estimate of the point of regard for the stereo and intersection model. Each time the deviation is larger than about 1 m, the point of regard does not coincide with the target object in vehicle space. Underestimating the depth leads to a larger deviation in image space than overestimating depth. The axis range is chosen for better visibility. In fact, the overestimations of the stereo model can be much larger than 20 m.

might both not always be given. In Fig. 6.14 a few seconds of the depth estimation error of the 3D point of regard are shown. It can be seen that the stereo model overestimates the depth several times. These overestimations occur when the gaze ray intersects something in the background which accidentally exhibits a lower depth error. They can also occur when the disparity values have smooth transitions at the edges of an object in the image or when the depth estimate at the true intersection position is not precise. E. g., here, the target is a black vehicle with only little texture. The same holds also for road surfaces around, so the stereo matching might not always be sufficiently accurate. Whether the estimated point of regard actually hits the correct target in these cases is, just like in the case of the heuristic model, rather random and depends on the size and shape of the object in the image or at which point on the object the driver looks at. The number of overestimations can be reduced if the absolute distance function (equation (4.2)) is applied instead of the relative distance function (equation (4.3)). However, this increases the problem of depth underestimation which is as well visible in Fig. 6.14. Due to the counterintuitive distance function which is used in the stereo model and explained in Section 4.2.2, the minimum distance between gaze ray and scene can occur at much lower distance, i. e. in front of the actual target. In contrast, it can be seen that the 3D point of regard from a simple model in vehicle space is extracted more robustly provided that the target's distance is measured. On the other hand, when the driver's

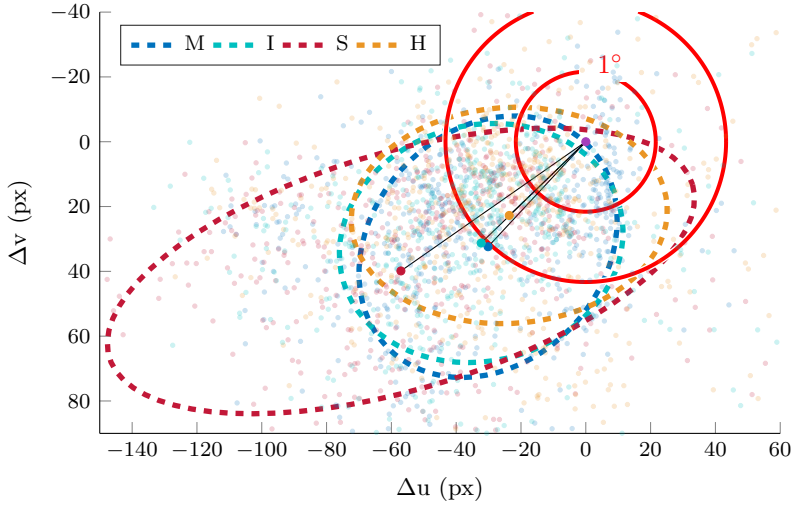


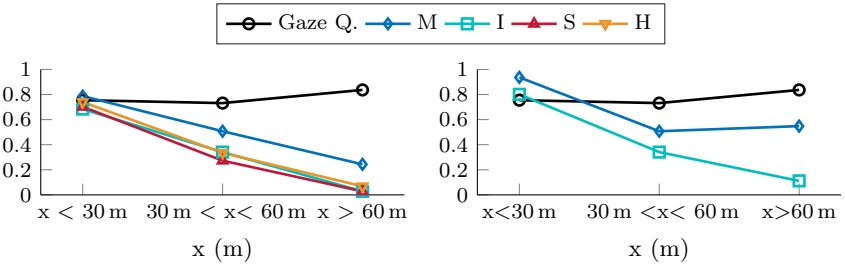
Figure 6.15: Deviation of the estimated point of regard. The figure shows the mean position in relation to the label, single data points relative to their corresponding label and the 50 % covariance ellipses.

gaze does not fall on a known object, the depth can only be estimated when using further measurements like a limiting free space or assumptions such as a ground plane. Here, whenever the depth error is larger for the intersection model, the gaze ray intersects with the free space spline.

Longitudinal Experiment

In the second experiment, the gaze target first recedes about 100 m, turns around and then approaches again. Just like before, the test person is asked to always fixate the gaze target. The statistical values for each gaze target estimation approach are shown in Table. 6.2 and the deviations of the estimated point of regard to the respective labels are presented in Fig. 6.15.

Two differences to the previous scenario directly attract attention. First, the classification scores are lower, especially the recall values indicating a higher number of false negative classifications, i. e. the models missed the correct target. Second, the stereo model exhibits a much larger covariance and a large bias to the left in the point of regard estimation.



(a) Semantic class prediction (b) Object prediction

Figure 6.16: F1-Scores for different longitudinal distances of the gaze target in the longitudinal scenario. The object prediction is only computed for the models in vehicle space. The additional black data points denote the average gaze quality as obtained from the SmartEye system for the respective segments.

Gaze Target in the Far The lower recall results are a direct consequence of basic geometry. If the object moves further away, the probability that the gaze ray intersects the object becomes lower if the gaze direction measurement accuracy stays the same.In Fig. 6.16, the F1 scores are shown for three different distance ranges[65]. For a cross check, the mean gaze quality score is added, yet, the gaze measurements are comparably reliable in the three sections. As expected, the F1 scores decrease with increasing distance, mainly because of decreasing recall values. For distances above 60 m, the chance of hitting the object with the gaze ray is neglectable visible

Table 6.2: Precision, Recall and F1-Score of the class-label 'car' in image space and of the object of fixation detection for the models in vehicle space. The third column shows the mean distance to the label position and standard deviations in direction of the eigenvectors of the covariance ellipses in pixel.

	class			obj. of fix.			pixel dist.		
	Pr	Re	F1	Pr	Re	F1	∅	σ_1	σ_2
H	0.89	0.32	0.48	-	-	-	52.4	41.5	28.3
S	0.88	0.30	0.44	-	-	-	81.2	79.7	31.1
I	0.88	0.30	0.45	0.95	0.31	0.47	57.7	37.5	30.9
M	0.83	0.44	0.57	0.94	0.63	0.76	55.5	37.3	30.4

[65] An error in the script responsible for the plots in the longitudinal scenario has been corrected in the preparation of this evaluation. The plot in the original publication [156] accidentally showed higher F1 scores at far distance than they really are.

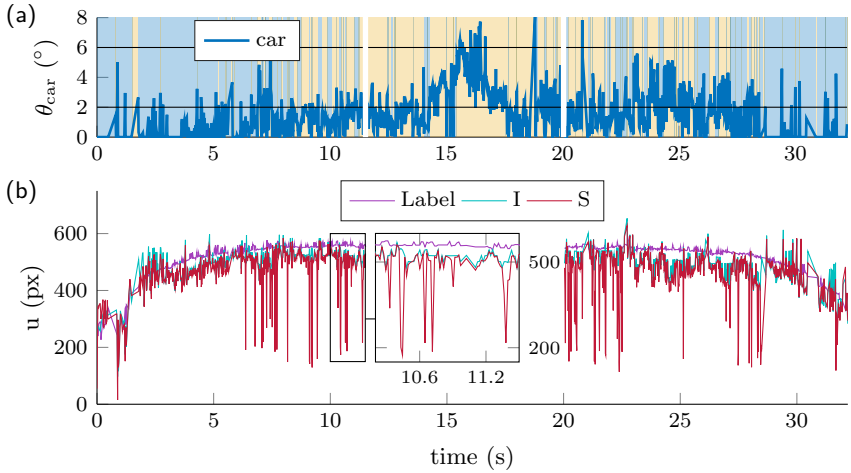


Figure 6.17: Time course of the longitudinal scenario with (a) the deviation of the measured gaze to the gaze target and the determined gaze target from the tracking model, and (b) the reprojected horizontal pixel position of the estimated point of regard for the intersection and stereo model compared to the labeled position. During the time between 11.5 s and 20 s, the sequence was not labeled.

by the intersection model in Fig. 6.16b. Also the classification in image space decreases significantly with the object's distance (cf. Fig. 6.16a). Especially here, the advantages of the tracking algorithm become visible as the number of FN is reduced in relation to the TP at all distances compared to the other models[66] meaning that more gaze samples on the object have been correctly classified. Fig. 6.17a, where the angular deviation is plotted with the tracking target from M as background color, shows that the angular deviation is seldom equal to zero. Nevertheless, the tracking algorithm succeeds to make correct classifications. This results in higher recall values and thus also higher F1 scores, yet at the cost of slightly more FP. The still high precision values are due to the fact that gaze target is always fixated, so there are more positives (looking at target) than negatives (not looking at target). According to the instructions, the latter should actually be zero meaning that false positive predictions should

[66] For this evaluation, the same parameters as in all other scenarios have been applied for the tracking model. It would be possible to design the model such the number of TP is even higher in this scenario, yielding however a too strict tracking behavior in natural driving scenes.

be impossible but this is not the case. First it is not sure whether the test person really succeeds in keeping their gaze stable on the target and secondly, the measured gaze position of the SMI tracker does not constantly fall on the target, especially if it is at far distance. When at its farthest at about 100 m, the car (approximately 2 m wide) is only about 20-25 px large, so that gaze-object matching becomes challenging. 22 px, respectively 1.8 m width at 100 m distance correspond to roughly 1° difference in the gaze direction, meaning that the required accuracy to correctly identify a match between gaze and object is similar to the SMI's accuracy reported to be 0.5° [161]. Fig. 6.17a also shows that a simple threshold model with 12° opening angle, i. e. a maximum deviation of 6° would manage to find almost all TP classifications. Such a threshold model seems advantageous here, but nevertheless, as argued in the previous chapter, this might lead to FP predictions if more potential targets were present in the scene.

Bias of the Stereo Model The second observation of the large bias and variance of the stereo estimate[67] has already been discussed in the scope of the crossing experiment. Yet, here it becomes more evident. When the reprojected gaze ray enters the image from the left side, each of the image points where the reprojected gaze ray falls onto is considered as potential point of regard. Especially when the supposed true target is at far distance, points very close can result in comparably lower distance error values. The problem of underestimating the depth is visualized in Fig. 6.17b and it underlines the fact that underestimating the depth can lead to errors much more severe than overestimating it. This is a specific problem of the camera's position in the test vehicle which is the regular position of a series camera behind the windshield for ADAS functionalities.

6.4.6 Results in Real World Scenario

With the two presented artificial experiments, it is almost impossible to capture false positives since the target is always fixated by the test person. Also, since the presented scenarios do not reflect natural human gaze behavior, a further example from real world driving is considered. Just like in the presentation of real world situations in the previous chapter, the sequence is much shorter than the artificial scenarios. This is mostly due to the fact that many traffic participants are not visible in the sensors'

[67] In the original publication [156], these large deviations of the stereo model have mistakenly been removed as outliers, even though they reflect the model's behavior.

Table 6.3: Precision, Recall and F1-Score of the class-label 'biker' or 'bike' in image space and of the object of fixation detection for the models in vehicle space. The third column shows the mean distance to the label position and standard deviations in direction of the eigenvectors of the covariance ellipses in pixel.

	class			obj. of fix.			pixel dist.		
	Pr	Re	F1	Pr	Re	F1	∅	σ_1	σ_2
H	0.97	0.62	0.74	-	-	-	43.6	40.8	22.5
S	0.96	0.57	0.72	-	-	-	38.8	36.3	25.9
I	0.96	0.66	0.78	0.63	0.06	0.10	37.9	40.1	21.5
M	0.93	0.40	0.56	0.82	0.67	0.74	61.8	58.9	23.5

field of view for very long except for a preceding road user. The situation is shown in Fig. 6.9c and again in Fig. 6.19. A bike rider is appearing at the upcoming T-junction from the road from the right in which the driver wants to turn in, thus it is expected that the test person has noticed the bike rider. Just like for the previous scenarios, the statistical values for each gaze target estimation approach are listed in Table. 6.3. Here, only the biker as a gaze target is considered. The other road user in the scene is considered later in the discussion. Also, the known plot of the deviations of the estimated point of regard to the respective labels is given in Fig. 6.18.

Like in the previous example, the tracking model M achieves the best results for object of fixation detection in vehicle space. In image space however, the statistical results and deviations are the worst. At first glance, this seems to be a contradiction. The explanation is given by the poor geometric consistency of the measurements as visualized in Fig. 6.19. Because the biker's position in the object list lags behind, the geometric intersection approach fails in this case whereas the MHMM tracking approach manages to find the correct target since the measured gaze is still close enough to the direction of the biker. Conversely, this also means that the estimated point of regard in vehicle space is 'drawn' towards the biker's measured position and thus its reprojection does not fall on the biker in the image just like the reprojection of the object does not coincide with the biker in the image either. In the reprojection, the models without tracking assumptions therefore reach higher scores and have a smaller distance to the label.

In vehicle space, the intersection model fails to detect the fixation on the biker. Since the intersection model is simply a special case of the threshold model with opening angle parameter of 0°, in a final step, the use of the threshold model is taken up again in order to find a more suitable threshold parameter. Given the reference data, it is now possible

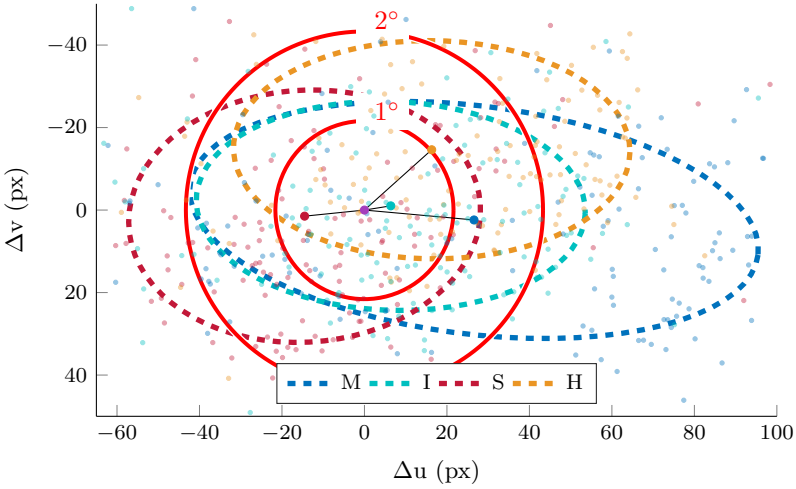


Figure 6.18: Deviation of the estimated point of regard. The figure shows the mean position in relation to the label, single data points relative to their corresponding label and the 50 % covariance ellipses.

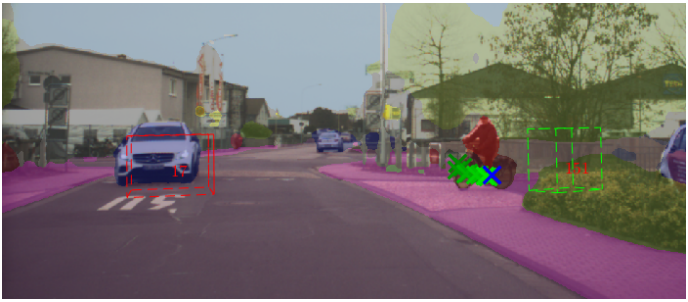
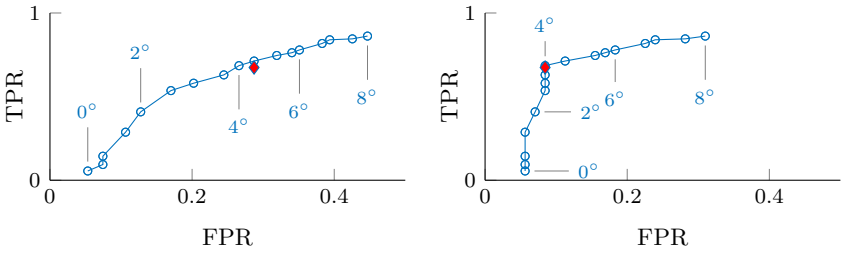


Figure 6.19: Tracking small traffic participants with low velocity can be erroneous. As stated before, the object is labeled as 'seen' even though the point of regard does not lie within the object's bounding box but on the object. The wrong object position shifts the tracking output slightly to the right.

to compare the results at least within one sequence. Fig. 6.20a shows the ROC curve (receiver operating characteristic) of the threshold model as binary classifier for various threshold values. The binary classification is given by deciding whether the biker is the current gaze target or not. The



(a) Evaluation of complete object life time. (b) Evaluation time without first false positive fixation (see Fig. 6.21).

Figure 6.20: TPR over FPR (receiver operating characteristic) of the threshold model for different threshold parameters (0.5°-steps from 0° – equivalent to model I – up to 8°) and for the binary classification of the biker being the gaze target. The red diamond marker marks the performance of the tracking algorithm as comparison.

ROC curve plots the TPR or recall over the FPR. Requirement for this is an existing TPR. As mentioned in Section 6.4.3, the TPR can only be computed if TP $\neq 0$, i. e. the respective object was looked at at least once. Not surprisingly, the TPR raises at the cost of an also increasing FPR. Yet, it is unknown which threshold value represents a sensible parameter choice. For better interpretability, the performance of the tracking model is added in the diagram which is close to the threshold of 4°. Fig. 6.21 shows the object of fixation result over time for the intersection model, the threshold model for 4° and 6°, as well as the tracking model and the label. The ROC curve in Fig. 6.20a was computed on the complete 4.5 s. As can be seen, by increasing the threshold value, the number of TP increases quickly but at the cost of a falsely detected fixation at the beginning of the scene. If this false detection is left aside for the ROC computation, i. e. starting the evaluation after the first 0.5 s, the ROC curve in Fig. 6.20b is obtained. Again, the tracking result is added and it is close to the 4° again. Even more, at this point, the curve exhibits a sharp bend. By increasing the threshold value, the FPR increases more than the TPR which is undesired. Conversely, by decreasing the threshold value, only the TPR decreases which is unfavorable as well. This not only suggests that a threshold value of 4° is a reasonable choice for this model. It also suggests that the parameters of the tracking model are well chosen. Now it could be argued why to actually use a complex tracking model if

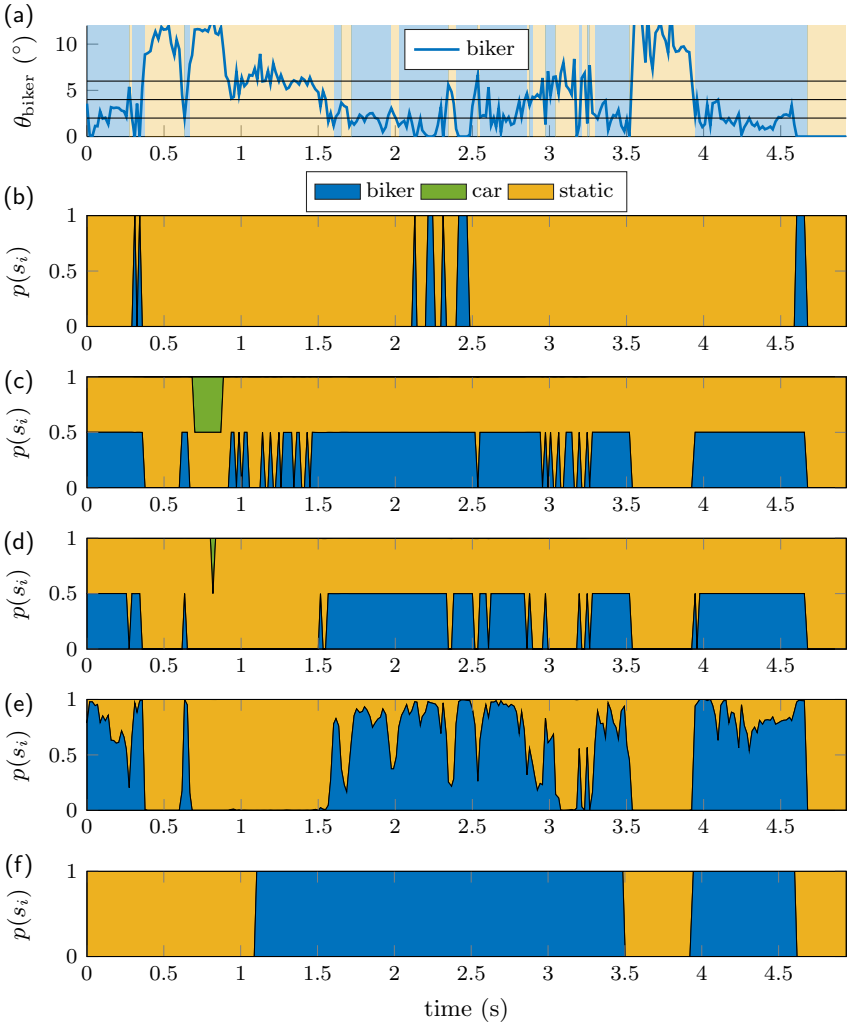


Figure 6.21: Time course of the real world scenario. (a) Angular deviation with tracking target result in the background. (b)-(d) Gaze target results of the threshold model for different opening angles (0° correspond to I, 4° and 6°). (e) Detection result for the tracking approach. (f) The label of the fixated object.

the threshold model with "correctly" chosen threshold parameter has a comparable performance. There are two main reasons which complement each other. First of all, the threshold model does not, as also argued in the previous chapter, decide on one object. In certain scenarios like the second real world example, potential gaze targets might be very close to each other regarding their direction from the driver. In such case, the FPR for at least one of the targets increases. The tracking model, on the other hand, provides a modeled probability for each target, i. e. a confidence measure. It furthermore favors connected time spans which are not captured in the statistical scores. Like this, unprobable gaze targets like the oncoming vehicle in this scene are successfully suppressed.

6.5 Discussion

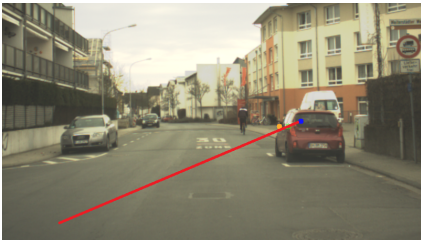
The previous section contained discussions on many different details of the different compared models. For this reason, this section summarizes the findings of the experimental results and discusses the approach to the problem of missing ground truth information.

6.5.1 Discussion of Models and Experimental Results

From the comparison of the four different models, it can be seen that each model exhibits certain characteristics which could be supposed beforehand but which have now been demonstrated using the reference data.

Reprojection and Depth Estimates

Interestingly, the estimate of the intersection model was in all scenarios among the best when it comes to estimating the point of regard in image space. From this, it is concluded that the simple intersection model is most robust in the estimation of the correct depth. Even though the approach of estimating the depth via stereo images is a basic geometric approach as well, the depth is sometimes heavily underestimated leading to bad point of regard estimates in the image. Even the heuristic approach, which does not compute depth at all provided more robust results in the image. When no object list or only a monocular camera is available, a definition of a typical depth might therefore already be sufficient for a reasonable point of regard estimate in the image. Fig. 6.22 furthermore presents one suggestion, how also the stereo approach can get more robust. By placing the camera on



(a) Actual camera position.
($x = -1.0\,\mathrm{m}, y = -0.1\,\mathrm{m}, z = 1.3\,\mathrm{m}$)

(b) Camera on driver side behind the cluster intruments comparable to a position of a head up display;
($x = -0.6\,\mathrm{m}, y = 0.3\,\mathrm{m}, z = 1.0\,\mathrm{m}$)

Figure 6.22: Virtual change of exterior camera position (untouched rotation). There exist better configurations for the fusion in image space, which are expected to be more robust.

the driver side behind the cluster instruments, the reprojected gaze ray is shorter and enters the image from the top. The reprojected ray is not only shorter, but especially the closer points on the ray are usually compared to points in the scene far away like the sky or buildings. In this configuration, the stereo approach is expected to be comparably robust as the intersection model.

Object of Fixation Detection

The detection of visual fixations on objects can pose a challenge when applying only basic point-based decision rules, e. g. using only the single point of regard pixel position or the intersection of the gaze ray with the scene. Already small measurement errors of the gaze direction can lead to deviations of several tens of pixels off regardless of the subsequent fusion model. Applying threshold models[68] leads to an improved TPR but at the potential cost of increased false positive detections. However, false positive detections should be avoided as much as possible. Based on this conflict, the MHMM tracking approach provides a promising trade-off by incorporating probabilistic assumptions, e. g. that objects closer to the gaze ray are more likely to be the visual gaze target. Given that extrinsic system calibration, environment and gaze measurements are precise and consistent,

[68] The extension of point based approaches to threshold models in image space is straightforward as a circle or ellipse around the computed point of regard can be considered.

good and reliable results can be achieved in the near field around the driver (up to about 30 m). Also, the FPR of visually not fixated road users is improved. However, if the mentioned prerequisites are not fulfilled, e. g. by an imprecise calibration, the incorporated model assumptions can lead to mistakenly high confidences for the wrong targets.

At the end, for an application, the specific score is not decisive but rather the question whether the driver has seen the respective road user. As long as the false positive detections remain low, it is more decisive to detect clear fixations regardless whether more or longer fixations have occurred. Therefore, when developing a tracking algorithm for gaze target tracking, it is important that the model does not read too much into the measurement data. The parameters of the presented MHMM are designed for this balancing act following the principle to keep the false positives low but if the measured gaze is close to a road user, then a visual fixation, i. e. a connected time span on that object, should be detected.

6.5.2 Discussion of Reference Data Recording Approach

In Section 6.2.4, the difficulty of the assessment of what a driver has looked at motivated this chapter. The created reference dataset provides the possibility to compare the result of different algorithms with a label directly instead of only describing qualitative aspects of a model. This comparison enabled not only the findings which have just been discussed, it also allowed to identify and narrow down the latency of the SmartEye system in the sensor network to 60-70 ms. Previously, the latency of the system could only roughly be estimated. The improved statistical results in the last sequence of all algorithms which have been compared underline the great importance of precise time synchronization for this purpose. Nevertheless, some general difficulties remain which are listed in the following.

First of all, the evaluation still concentrated on short, extracted sequences where gaze on single road users has been investigated. A truly reliable result of a method ranking requires a large testing catalogue which includes scenarios of different complexity. For this, two prerequisites need to be fulfilled. First, a reliable and precise environment perception is necessary. Often, potential true positive predictions can not be counted because objects are detected after the driver's first glances towards them or objects are not tracked robustly. Second, the results of different models can, of course, be only as good as the label. Also the measured point of regard

of the wearable SMI tracker exhibits a certain variance which affects the gaze target and the image position of the label. This was especially visible in the second investigated scenario with the gaze target at far distance. Therefore, a testing catalogue needs to be restricted to verifiable scenarios with valid labels. Yet, up to now, the labeling procedure is extremely tedious due to the high gaze sample frequency of 60 Hz. Suggestions to automate the labeling process using a homographic transform have been made in [5, 116]. This approach does not only pose additional difficulties, also the precision of the automatic label is unclear. The question of label accuracy is also linked to the joint usage of the systems. It was shown that the presented eye trackers do indeed work simultaneously. However, a joint usage of systems has to be investigated every time anew in case-by-case reviews depending on the hardware choice, the camera and IR lighting configuration and also depending on the test person. Thus, the scaling of the proposed approach poses a large challenge.

Independent of the prerequisites, the actual target function needs to be discussed and defined. Truly interesting would be to assess complete visual fixations on objects or even how good a method is to detect the driver's first fixation on a traffic participant. As discussed in the previous chapter, the results from all presented models, which have been assessed time step by time step in this chapter, would need to be postprocessed in actual fixation detections.

6.6 Summary and Conclusion

In the first part of this chapter, the problem of missing ground truth information for remote eye tracking in the driving context has been addressed. Even though eye tracking systems can be assessed in laboratory or stationary setups, it is difficult to evaluate an eye tracking and information fusion system under natural driving conditions. The second part of this chapter outlined the approach, procedure and the encountered difficulties to create reference data for the sensor system in the project's test vehicle. Despite some challenges, e. g. active IR illumination of each device, sunny weather, occlusion of the eyes or time synchronization, it was possible to find a suitable and functioning configuration of a combined sensor setup comprised of a wearable and a remote eye tracking device. Most importantly, in the selection of the wearable device, it was paid attention that each gaze tracking system does not impair the tracking performance of the respective other device. This reference data was used in the third part of

the chapter to compare different models and algorithms from the field of visual attention region estimation, point of regard estimation and object of fixation detection, among those also the model introduced in the preceding chapter. One achievement of the presented work is that different algorithms can now be quantitatively compared and potential systematic error sources can be detected. Nevertheless, the evaluation still concentrated on short, extracted sequences where gaze on single road users has been investigated. The recorded dataset with one test person has been made publicly available. It is up to further research, how the proposed procedure can be scaled and thus evaluation of different fusion algorithms be generalized more. Future work could thus focus on further experiments and a more standardized evaluation. Especially more recordings from different test people in real world scenarios are necessary, potentially leading to a testing catalogue since experiments can not be repeated. Concerning the data collection, a fully integrated sensor setup where all data is collected and time-stamped in one network and where the IR illumination is specifically designed to suit each device could avoid some of the encountered problems. Yet, the development of such a system would be cumbersome. Furthermore, a high labeling effort is necessary to evaluate larger time series of gaze behavior. One aspect of future work could therefore include the automation of the labeling process.

7 Driver Gaze Behavior in PRORETA 4

This chapter presents the prototypical adaptive City Assistant System, developed in the course of the interdisciplinary research project PRORETA 4, which bundles concepts and findings of the four research assistants involved in the project. The emphasis of this chapter is of course put on the incorporation of the driver's gaze, the overall topic of this thesis. Therefore, before the City Assistant System is fully introduced, the benefits and potentials but also the challenges of the incorporation of driver gaze information are outlined. This argumentation constitutes the bridge between the previous chapters of this work that focused on specific fusion methods and this one focusing on possible applications. Secondly, an overview of the system focusing on the incorporation of gaze and the Human-Machine Interface (HMI) is given. The third part of this chapter is dedicated to one particular challenge related to the use of the driver's gaze, namely the procedure of personal gaze calibration which is often unnatural and bothersome. Here, the solution implemented in the PRORETA 4 City Assistant System to avoid personal calibration is presented.

7.1 Awareness Estimation for ADAS

The motivation to include an estimate of the driver's situation awareness into new ADAS is well explained in [104] and is therefore given here with a similar argumentation. State of the art assistance systems that do not include driver monitoring systems must compulsorily assume that the driver is unaware of an upcoming risk. The question whether to raise a warning or even actively intervene is thus solely based on the configuration of traffic participants and the resulting risk[69]. At the same time, current systems

[69]Risk measure and modeling in traffic scenarios is a research field in its own. Nevertheless, it is important to note that the instantaneous risk of a situation is not affected by knowledge on the driver, e. g. the course of the driver's focus of attention. But the understanding of the driver and their actions can affect the predicted risk and

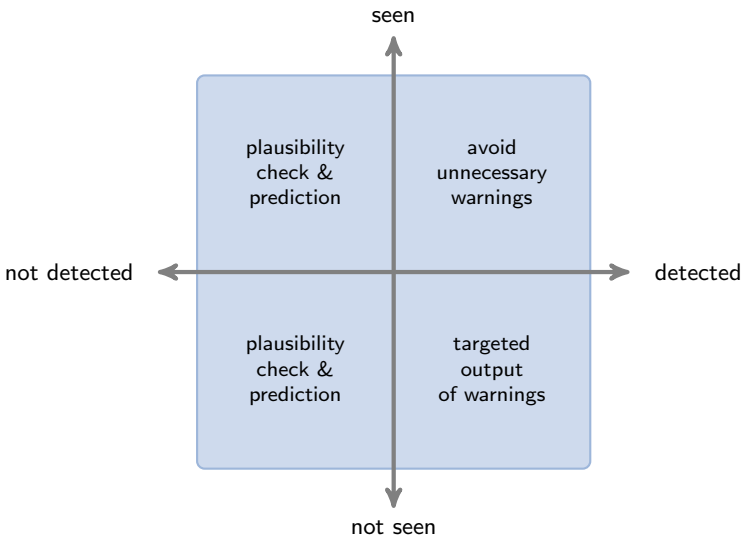


Figure 7.1: Exploiting the situation analysis depending on the situation awareness of system (horizontal axis) and driver (vertical axis) according to [104, 105] for the adaptation of warning/intervention strategies. On the right side, the gained axis of seen and missed road users can contribute to the mitigation of the classical warning dilemma. Operating on the left side, warning strategies can exploit the knowledge of where a driver should look compared to where the driver's visual attention actually lies.

assume a complete environment model and thus do not take the possibility for undetected road users into account. According to [104], current systems thus act solely in the lower right quadrant of Fig. 7.1. However, in many situations in real world traffic, the driver is aware of other road users (upper half plane) and sometimes even interacts with them while at the same time, the environment model of the vehicle might be incomplete (left half plane). As presented in Fig. 7.1, the author of [104] divides the space of possible actions in the right half plane into the application of targeted warnings and the prevention of unnecessary warnings. However, it is additionally argued here that the gained axis of seen and missed road users can further contribute to the mitigation of the classical warning dilemma. This dilemma

thus opens up the possibility to influence its progression at an earlier stage than it is possible without estimation of the driver's situation awareness.

also implicates that late interventions might be heavier than necessary and might not prevent an accident. Early interventions and recommendations might be more subtle but are exposed to higher uncertainty regarding the evaluation of a situation. For an appropriate strategy, an estimate of the driver's mental state can be a helpful source of information as it can not only help to prevent warnings but also to adapt the *Time To Intervention* (TTI) and the warning escalation process itself. New design strategies can act like always attentive co-pilots that specifically aid the driver without overloading them with distracting unnecessary advice [106], leading to a closer cooperation between the vehicle and the driver. Even in the case of the left half plane of Fig. 7.1, i. e. an incomplete environment representation, knowledge about the driver's mental state can be incorporated in the system behavior as exemplarily presented later in Section 7.2.2: the knowledge of where a driver *should* look due to the traffic rules in place can be checked against where the driver's visual attention actually lies. The detection, prediction and consideration of non-visible road users, e. g. due to occlusion, as presented in [104] is another aspect of this left half plane that is not considered in detail in this chapter.

In order to incorporate the driver's visual behavior as one source of information into ADAS, the driver's gaze and head pose data needs to be mapped with all other information necessary for a specific application, e. g. ego-speed, position and motion of other road users and even traffic rules. This new aspect of driver monitoring systems, namely being coupled to active assistance functions is until now object of research. Prototypical assistance systems of different maturity realized in vehicles are presented in [11, 104, 119, 137, 157, 183, 189]. Whether an explicit modeling of the objects perceived by the driver such as in Chapter 5 is necessary, is a question of the targeted application and its realization. It might often be sufficient to couple a rough attention estimate with an estimate of the driver's intention such as in [105, 106, 147] but somehow, the potential of the vertical axis in Fig. 7.1, i. e. of seen and unseen road users, should be included in the system design. In the next section, a straightforward example of how gaze can contribute to a classical warning strategy is presented in the left-yields-right intersection use-case of the PRORETA 4 City Assistant System. This system is designed to aid drivers in different urban intersection scenarios in a personalized and adaptive way. So in the following, first the overall motivation and system design is given, before the use of the driver's gaze is highlighted.

7.2 The PRORETA 4 City Assistant System

7.2.1 Introduction and Motivation

Statistics from 2017 in Germany show that most (over 68 %) road traffic accidents involving personal injury occur within urban areas. More than 60 % of the slightly injured, more than 50 % of the seriously injured and about 30% of the killed people have been involved in an accident in this area [170]. Therein, the major type of accidents are turn and crossing accidents, thus caused by the violation of the effective rules of right of way. This can be due to unseen traffic participants or misjudgment of the current situation and the driver's own capabilities. All these reasons are directly linked to decisions regarding maneuver execution.

On the market, many systems from different vehicle manufacturers are available that support the driver in the longitudinal vehicle guidance or warn or even intervene to prevent potential front collisions or collisions at junctions. However, such systems do not provide active support in maneuvering decisions, e. g. there is currently no system that tells the driver whether the current gap between two oncoming vehicles suffices to turn left or whether the driver should wait for a larger gap. This is partly due to the fact that for active maneuver recommendations the individual driving style and capabilities as well as the driver's current attention and emotional state play a major role. A general design of driver assistance systems as it is done for "common" collision warning systems is not possible here. In order to develop new systems that actively assist the driver in making maneuver decisions, there exists a need for functional customization in terms of driving style and driver's attention. Such individualization has already been identified as one future research focus for new ADAS [17, 195].

The TU Darmstadt met this demand together with Continental AG in the research project PRORETA 4. The research project, successor of PRORETA 1 – 3, is dedicated to the adaption of driver assistance systems to the situation and the individual driver. The project was scheduled from 2015 to 2018 and four research assistants from three different institutes of the TU Darmstadt worked together on this interdisciplinary project (see Fig. 7.2). Within this frame, several publications comprising new algorithms for driver intention detection and online driver adaptation [4, 33–36], visual localization and mapping [23, 108–110] and, the topic of this thesis, driver gaze target estimation [156, 158–160] have been published as well as articles on safety approval of machine learning algorithms in

Figure 7.2: PRORETA 4 core research and development team with prototype vehicle. From left to right: Hien Dang (Machine Learning and Driver Adaptation), Julian Schwehr (Driver Monitoring and Driver Gaze Target Estimation), Stefan Luthardt (Vehicle Localization and Environment Perception), and Maren Henzel (System Architecture and Safeguarding).

the automotive context [63]. For more detailed information, the reader is referred to the related dissertations [32, 62, 172]. Many of the core ideas can also be retrieved from [157] where the exemplary prototypical assistance system has been introduced. The main aspects from [157] are outlined in the following Section 7.2.2.

7.2.2 System Description

The PRORETA 4 City Assistant System issues an individual, situation-adapted maneuver recommendation or warning in demanding urban traffic situations by observing the driver, their current driving style and the environment. This maneuver recommendation was implemented for three use-cases (cf. Fig. 7.3) on the test vehicle:

1. Left-turn scenario: Give a recommendation to wait or perform a left-turn at intersections with oncoming traffic. The recommendation is personalized according to a short-time driver profile.

2. Roundabout scenario: Give a recommendation to wait or enter a roundabout. The task at hand is similar to the left-turn scenario.

3. Left-yields-right scenario: Give recommendations and warnings at

(a) Left-turn scenario: The driver has to decide which gap to take.

(b) Roundabout scenario: The driver has to decide when to enter.

(c) Left-yields-right scenario: The driver should behave safely.

Figure 7.3: Use Cases in which the PRORETA 4 City Assistant System supports the driver[70].

left-yields-right intersections when the driver's behavior lets to expect a violation of the rules of right of way or does not exhibit cautious elements.

The different main function blocks are shown in Fig. 7.4. They firstly include an environment perception and situation comprehension module and secondly a module to grasp the driver style and infer the individual adaption. Thirdly, the incorporation of the driver's visual behavior is addressed. These three main function blocks are controlled by a behavior coordination and planning module that sends out all necessary information to the HMI. In the following section, the incorporation of gaze is discussed in more detail as it outlines one potential usage of driver monitoring and specifically gaze estimation for future ADAS. Furthermore, an overview on the HMI issuing recommendations to the driver is presented in order to provide an understandable description of the system's behavior. Other

[70]Images taken from https://www.ikiwiki.de.

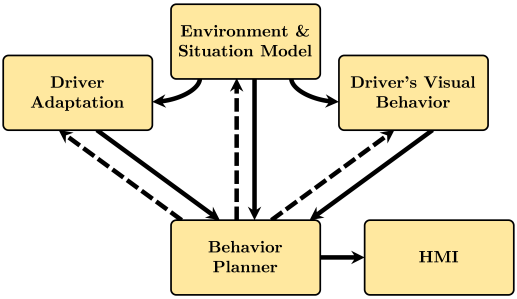


Figure 7.4: Overview of the function blocks that compose the PRORETA 4 City Assistant System. The solid arrows depict the flow of information to generate and deliver the adaptive recommendation to the driver. The dashed arrows indicate the situation dependent control of the function blocks. Image according to [157].

modules, however, which are not in the focus of this work, are not described here. The interested reader is referred to the related article [157].

Incorporation of the Driver's Gaze

The PRORETA 4 City Assistant System incorporates the driver's visual behavior in two ways. For the left-turn and roundabout recommendations, it is analyzed whether the driver is visually distracted and has turned to secondary tasks when waiting for a gap. However, this use-case is not specifically suited to demonstrate the advantage of driver monitoring and driver awareness estimation for ADAS. Here, the detection of distraction rather presents a comfort function that allows the driver to take attention away from the driving task while waiting for a gap in standstill. For more safety-relevance, situations like crossing left-yields-right intersections, passing zebra-crossings or overtaking vulnerable road users are much better suited. From this list, the left-yields-right intersection scenario fits best to the idea of maneuver based assistance in urban intersection scenarios which is why the City Assistant System has been extended by this use-case. The use-case is furthermore advantageous as the traffic rules are simple and the execution of the maneuver often follows a stereotypical pattern. Even more, the relevant traffic participants can be identified from the same situation model as in the left-turn or roundabout use-cases.

Distraction In the recommendation and warning strategy of the City Assistant System, driver distraction is only considered if the driver is waiting at an intersection to turn left or at the entry point of a roundabout and a line of cars is visible in the sensor range. Only then, when no arriving gap is expected to fit the driver, it is checked whether the driver shifts their visual attention away from the driving task to a secondary tasks, e. g. texting, or talking to other passengers. In that case, the system notifies the driver when a sufficiently large gap is approaching.

In the style of existing driver distraction detection algorithms [85], a basic distraction detection model was designed. First, application specific areas of "eyes-on-road" are empirically determined. This means, that the definition of "eyes-on-road" and thus "driver visually attentive" changes with the respective use-case, i. e. left turn or roundabout. The goal of the model should be that in order to classify the driver distracted, a certain time must evolve where the driver looks away, i. e. outside of that predefined area. This leads to a natural latency of the distraction detection. When the driver re-involves in the driving task, the detection latency should be as short as possible. Since these requirements stand to some extent in contradiction to the robustness of a model, the implemented algorithm works as follows to realize the trade-off. Over an evaluation time span of the last 1.5 s, each gaze sample inside the eyes-on-road region increases a counter by 2 and each sample outside decreases it by 1. If the counter over the last 1.5 s is negative, the driver is classified as distracted. The latency and the switching behavior can be controlled by the counter increments and the evaluation length. The realization via such an asynchronous counter has led to a more meaningful system behavior than a synchronous counter or just regarding the mean values of gaze pitch and heading.

Left Yields Right Intersection Giving priority to the right is in many countries with right-hand driving the fundamental rule of right of way. Often, these left-yields-right intersections come without the regulation to come to a full stop, rather it is enough to slow down to be able to give way to the other road user approaching the junction from the right. Furthermore, these junctions are widely used in residential areas due to the lower speed and generally calm traffic. However, two main risk sources can be observed. Firstly, local residents are often used to their common routes and therefore know where the expectation of approaching road users is low potentially leading to increased speed and sloppy attention. This habitual effect was also observed in the recorded data set where each driver

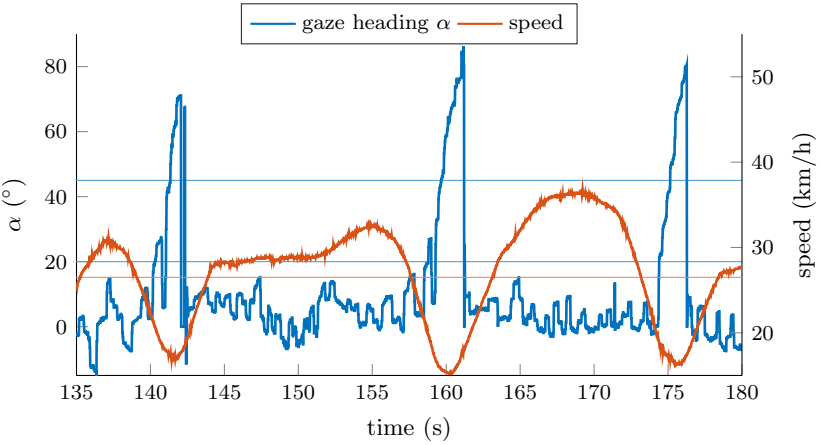


Figure 7.5: Example of typical behavior of gaze heading and speed over time in a residential urban area with left-yields-right intersections. The deceleration in front of the intersections as well as the shift of the driver's visual attention towards the junctions are distinctively visible. The three horizontal lines are the thresholds that are checked in the assistance function: 20° and 45° for gaze and 26.5 km/h for the speed.

drove the same route 30 times. Secondly, junctions are not necessarily well visible from afar so that unfamiliar drivers might miss these intersections.

The intuition that eye movements and driving maneuvers are related has often been addressed and exploited for tasks like maneuver prediction such as in [36, 99, 117, 120]. As investigated in [120], drivers exhibit characteristic behaviors when performing certain maneuvers such as lane change or crossing a left-yields-right intersection. When approaching a left-yields-right intersection, drivers usually slow down and visually secure the intersection sufficiently far away from the junction. When entering the intersection, drivers usually look into the merging road [120]. Three exemplary and consecutive approaches to a left-yields-right intersection are shown in Fig. 7.5. The deceleration as well as the visual gaze shifts are clearly visible. In the assistance system, the driver's approaching behavior is observed and compared to this typical expected behavior. Thus, it is not simply checked whether the ego-speed decreases. Rather, while approaching the intersection, it is also verified whether the driver's gaze exceeds different yaw angle thresholds (here, 20° and 45° are checked, marked also in Fig. 7.5) at different distances to the intersection, i. e. it is checked whether the driver

behaves correctly according to the situation. Besides the characteristic viewing behavior, the most critical regular event – it shall not be thought of any exceptional outlaw event – would occur if a vehicle approaches from the right which has thus right of way and at least one of the drivers in the situation does not see the other one and misses to brake. So additionally, if another road user with right of way approaches from the right, it is evaluated whether this road user poses a potential traffic hazard to the ego-vehicle and whether the driver visually perceives this object (cf. Fig. 7.6). A common collision avoidance system would raise a warning if the risk exceeds a certain threshold[71]. This warning or intervention normally occurs as late as possible due to the warning dilemma. Here in contrast, if no fixation on the other road user is measured, the warning can be given as early as possible since it is assumed that the driver might have overlooked the other road user. Contrarily, if the driver has seen the other road user, an emergency brake maneuver[72] "in the last moment" could be sufficient. During the project's final event, this rare event of a potential collision did not occur. Instead, the less critical event of an approaching vehicle with no collision risk was quite common. Here as well, it is expected from the driver to visually secure the intersection. In case that they do not secure and "miss" the opposed vehicle, the system gives an auditory feedback to the driver to pay more attention to the traffic from the right next time.

In order to bring all the necessary information together, traffic objects are filtered from the set of objects so that only objects relevant to the situation for which the warning strategy is formulated remain. This filtering was done on a set of formulated rules. However, in more general scenarios, the importance of other road users can as well be learned [130].

The specific approach in the left-yields-right use-case covers all four quadrants in Fig. 7.1 in its warning strategy. The right half plane is tackled by the different system outputs depending on the risk and whether the driver has seen the other road user. The left half plane is considered through a strategy that does not contradict the present situation even if another road user remains undetected. In case of unavailable gaze direction or head pose data, the system has a fallback strategy to operate on the horizontal axis of Fig. 7.1. Such a rule-based approach can in principle of course be extended to arbitrary use-cases and warning strategies. However, for more sophisticated assistance systems, a more general approach, e. g. a

[71] Here, risk is defined in terms of a simple 2D *Time To Collision* (TTC) with circular safety buffer according to [69].

[72] Active intervention was not part of PRORETA 4, however, for a complete user experience, an emergency brake maneuver would pose a sensible extension of the system.

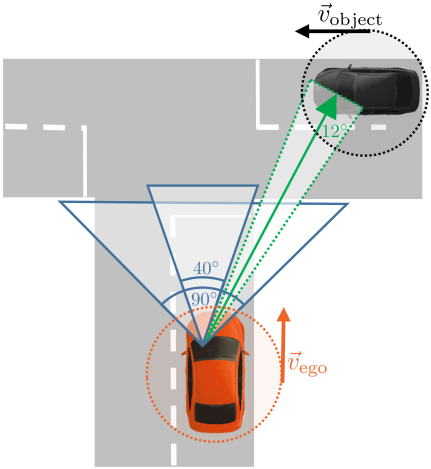


Figure 7.6: Incorporating driver visual cues for the adaption of warning strategies: When the driver approaches a left-yields-right intersection, their approaching behavior consisting of a decreasing ego-speed and a visual securing strategy is observed. For the driver's gaze, different thresholds (here, 20° and 45° to the right) at different distances to the intersection are checked for. Additionally, if another road user with right of way approaches from the right, an early warning can be raised if, e. g. no visual fixations on the road user are detected and the observed TTC of the circular buffer areas is below a threshold. A visual fixation is detected if the measured gaze ray intersects the bounding box of the road user with a tolerance of 6°. In the future, the simple threshold model can of course be replaced by the tracking model presented in Chapter 5.

probabilistic one as in [104], would be desirable.

For the detection of the driver's gaze target in real driving setups, most models use broad tolerance thresholds for the gaze measurements for increased robustness [11, 96, 137, 200] as remote eye tracking systems in automotive applications often do not reach the necessary precision to rely solely on the measurements themselves [156], cf. Chapter 4 and 6. In the City Assistant System as well, a basic threshold model was applied. A tolerance of 6° was added to either side the gaze vector when computing the intersection with the objects' bounding boxes (see Fig. 7.6). Similar to the model in [11], a minimum fixation time was considered. As described in Section 4.2.2, such a threshold model is more robust compared to a simple intersection model. For the project's scope, it also posed a reasonable

trade-off between implementation effort and detection quality. In the future however, this simple object-of-fixation detection model can of course be replaced by the tracking model presented in Chapter 5.

One practical issue within the City Assistant System is the individual gaze calibration (cf. Section 2.1.2). Normally, the eye tracking system needs to be calibrated to each user which is bothersome and sometimes lengthy. One approach to obtain gaze measurements that at least satisfy the requirements of the presented use-cases is presented in the following Section 7.3.

Another difficulty, that is out of the scope of gaze target computation, is the rapid, early and robust detection of relevant road users. In order to reliably act in the right half of Fig. 7.1 and warn the driver, an object should be detected before the driver has seen it. Otherwise, a road user might be detected by the driver but falsely classified as missed. In the case of a suddenly appearing object, this time frame can be as short as the duration of a typical fixation, i. e. a few hundred milliseconds. This however was not part of the research topics in PRORETA 4.

HMI and Recommendations

The goal of the PRORETA 4 City Assistant System is to provide the driver with individual and situation dependent recommendations and warnings tailored to the driver's needs and capabilities. This is why also the HMI of the system, which has been developed at Continental[73] in close collaboration with the PRORETA 4 research team, follows a multi-modal, holistic approach.

Left-Turn and Roundabout Scenario For the recommendation of the gap to take, different concepts including static and dynamic visualizations had been discussed. The final choice fell on a dynamic animation of the approaching gaps in the oncoming traffic colorized by the intuitive and common colors red and green signalizing whether a gap is too small or large enough to take (see Fig. 7.7). This approach actively supports the driver to anticipate the arrival of the gap to take. It supports the intuitive and natural expectation that the first gap becomes gradually smaller before a new gap opens up. Similar dynamic display concepts were investigated in [92] for merging scenarios, supporting the design choice. A static interface

[73] Many thanks to everyone involved, most notably Christoph Wannemacher and Herbert Deckenbach from Continental for the design and development, and Lars Mohrmann from Interative Pioneers for the implementation.

(a) Recommendation in a left-turn scenario.

(b) Recommendation in a roundabout scenario

Figure 7.7: Instrument cluster visualizations of recommendations for maneuver execution of the City Assistant System (left) with an image of the respective situation (right): The red arrow recommends to let the oncoming/crossing traffic pass. The dynamic visualization of gaps helps the driver anticipate the arrival of the "green" and thus sufficiently large gap.

in contrast, might lead to increased reaction times which would need to be considered in the system behavior by making use of a larger safety buffer. This in turn, might rather lead to misleading recommendations and as a consequence to a rejection instead of the acceptance of the recommendation system. In situations with dense traffic, the recommendation system relieves the driver from the stress of finding an appropriate gap suiting their needs and not to risk a too small gap. Especially inexperienced drivers could benefit from such a system.

Additionally to the visual cues depicted in Fig. 7.7, which are displayed in the instrument cluster, auditory signals support the driver. In cases when the driver is waiting for a suitable gap to arrive and is further shifting their visual attention to secondary tasks, the system prompts the driver to focus on the situation approximately 2 s before an appropriate gap arrives. Since roundabout scenarios exhibit high dynamics and often only short time windows of "green" gaps, here, the auditory outputs signaling the driver to enter the roundabout are of increased importance. Often, glances to the instrument cluster take too much time and a gap is gone before it can be taken. In these situations, starting to drive at the sound signal

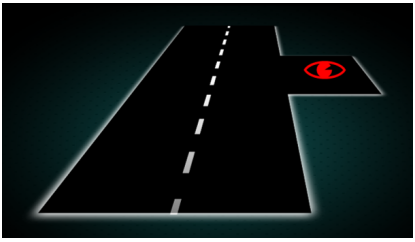
saves the necessary tenths of a second.

Left-Yields-Right Intersection Scenario Building on the insights presented in Section 7.2.2, a cascade of prompts and warnings has been constructed. If the driver does not slow down or secure visually or even fails to do both, recommendations according to the indications shown in Fig. 7.8 are displayed on the instrument cluster reinforced by an auditory signal with increasing frequency the closer the junction gets. When entering the intersection, it is again checked whether the driver truly secured the intersection and a feedback in form of a speech output is generated. All modalities except for the cluster outputs can be activated or deactivated according to the driver's needs. Thanks to the multi-modality, the driver can also turn off admonishing feedback and choose a lauding feedback instead for situations where they have seen the vehicle or have visually secured the intersection well. A simplified flow chart of the warning strategy is shown Appendix C.

7.2.3 Summary

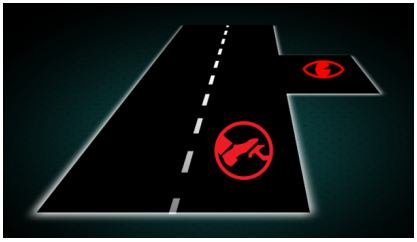
Within this Section 7.2, the incorporation of the driver's gaze into a prototypical next-generation assistance concept with a comprehensive understanding of scene, situation and driver has been outlined. The concept of the PRORETA 4 City Assistant System is based on the observation that especially in complex urban scenarios, there exists no *one-fits-for-all*-configuration of an assistance system. Additionally to the driver's momentary driving style, the driver's current visual attentive state is incorporated into the situation understanding, providing insights whether the driver visually secures intersections. The pursued warning strategy at left-yields-right intersections covers all four quadrants of Fig. 7.1 and adapts the HMI output according to the situation. Thus, the PRORETA 4 City Assistant System supports the driver in right of way decisions of several kinds in a personalized and adaptive way. The full prototypical system presented in this section was implemented on the test vehicle provided by Continental and has been successfully demonstrated to the public on the project's two day final event in October 2018. The participants were encouraged to experience the City Assistant System on a test route in real urban traffic, in a static setup and in video demonstrations[74].

[74] A video showing the City Assistant System in all use cases from a driver's perspective can be found at www.proreta.de



(a) Look to the right.

(b) Slow down.



(c) Slow down and look to the right.

Figure 7.8: Instrument cluster visualizations of warnings in left-yields-right intersections: Depending on how the driver approaches a left-yields-right intersection, they receive different HMI recommendations. Depending on the distance to the intersection, the visual cues are reinforced with an acoustical signal of increasing frequency.

7.3 Implicit Gaze Calibration

7.3.1 Motivation

As every human being exhibits a naturally occurring anatomic deviation between the visual and optical axis of the eye as outlined in Chapter 2, suppliers of commercially available head-eye-tracking systems often provide their tracking software with an individual gaze calibration function. As explained in Section 2.1.1, the optical axis describes the best fit of a line through the centers of curvature of the refracting surfaces while the visual axis connects the fovea and the point of fixation and corresponds to the line of sight (cf. Fig. 2.1). The personal calibration procedure aims at correcting the measured gaze ray of each eye based on known 3D fixation points and providing one single gaze vector as joint gaze direction of both eyes compensating for the anatomic effects as much as possible.

As early as in the concept phase of the left-yields-right use-case it

became clear, that a personal gaze calibration of each guest driver will be infeasible at the project's final event, due to exceeding effort of the procedure. Moreover, a personal calibration is also disadvantageous in a series application as a driver might not want to take the time to perform it. For some applications like simple "eyes on road" estimation, a rather rough gaze estimate might be sufficient. For the question, however, whether the driver visually fixated another road user, a high precision of the gaze direction estimate is crucial (cf. with insights from Chapters 4, 5 and 6). So the motivation of personalizing different parts of an assistance system to the individual driver is met again in this specific sub-problem. According to [56, 57], methods without any personal calibration seem not to be capable to create high precision estimates given the state of the art eye and gaze models. This conflict is addressed by [193], where an implicit calibration procedure is proposed as the personal calibration is argued to be unnatural and bothersome. The term implicit indicates that the calibration procedure is hidden to the user. It does not mean the calibration is obsolete, on the contrary, it is hidden by making use of what the person is probably looking at. In this approach, a fixation point can be known if e. g. a point on a screen serves as stimulus. Given multiple cameras and multiple light sources, the personal calibration can be reduced to a simple one-point calibration procedure [55].

However, such a procedure was infeasible in the PRORETA 4 test vehicle since only the head-eye-tracking system information was provided and no raw image data due to limited bandwidth. Thus, the question at hand was how to obtain a gaze direction signal from the uncalibrated gaze signal which is as precise as possible, i. e. as close to the signal that would be obtained with personal calibration. The approach that was finally implemented and some exemplary results are presented in the following subsections.

7.3.2 Approach

An approach of an implicit correction of the gaze data was pursued of which the user does not take any notice. The general idea is based on the hypothesis that a *general gaze pattern* can be observed for different drivers provided they drive structurally similar routes and have performed a personal gaze calibration. More precisely, the accumulated histograms of gaze pitch and gaze yaw are expected to have similar shape and location, i. e. a principal gaze direction and a similar width of the main lobe, when driving e. g. in urban residential areas or on highways. In the course

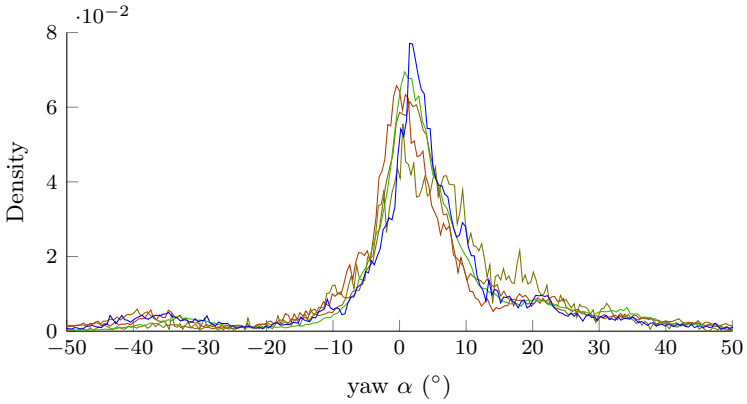


Figure 7.9: Comparison of discrete sample distributions of gaze yaw of four subjects from the recorded data set. The peaks lie within a span of 2°. The main lobe representing the "eyes-on-road" region is within a narrow range for all drivers.

of the project, two main data recordings have been performed with 32 respectively 14 drivers of different age, gender and driving experience. In the first recording, each of the 32 drivers drove 30 times the same route in an urban area of approximately 4-5 min resulting in about 2-3 hours of driving per person. In the second recording, the drivers had to drive structurally comparable routes. All drivers had performed a personal gaze calibration. Analyses of the recorded gaze data support the outlined idea of similar general gaze patterns. In Fig. 7.9, five examples of gaze yaw distributions for different people are given. As stated, all subjects drove the same route. It can be observed that all sample distributions indeed show similar shape and location. From this data, a person with qualitatively good tracking results in terms of precision and coverage and with a rather average distribution, i. e. not with the peak being far to the left or right and a medium main lobe width, was selected to provide a reference histogram as approximation of the expected gaze direction distribution for other drivers (the blue distribution in Fig. 7.9). The qualification of this person was verified with recordings from the same person from the second data recording and several hand-picked scenes.

The correcting procedure for uncalibrated drivers is conceivably simple and without raising a claim to perfectly correct the deviation between visual and optical axis. However, since the deviation of the axes are anatomically

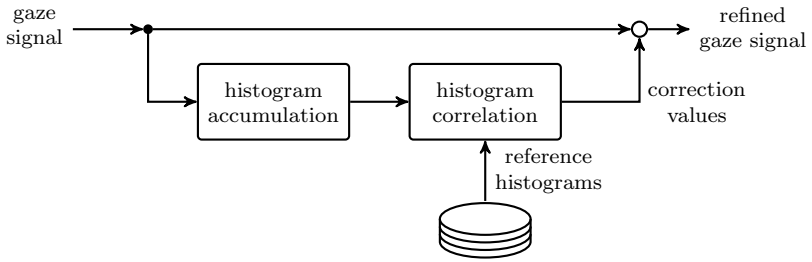


Figure 7.10: Procedure of the implicit gaze calibration: The driver's gaze signal is accumulated and compared to reference data. The obtained additive offsets are then added to the signal.

constant and gaze does not have a roll angle component, it is assumed for simplicity that the gaze direction error can be compensated through a constant bias in each direction. In order to do so, the signals of gaze pitch and yaw are accumulated in histograms. As not all gaze values are relevant, e.g. when driving backwards or when standing still, gaze is only collected while driving forward. Additive offsets to the reference data histograms are then computed as the shift maximizing the normalized cross-correlation (NCC), also known as Pearson correlation coefficient, and added to the gaze signal as shown in Fig. 7.10. Using the NCC corrects above all the location of the main peak. The main effect of this approach is thus to force the "eyes-on-road"-region to occur in roughly the same direction for all drivers. The additive model is not capable to capture other error sources such as possible scaling factors or nonlinearities and thus cannot lead to improvements in the edge regions, where other error sources might be predominant. A more detailed discussion on this aspect is given in the following section where the proposed approach is investigated. Similar approaches to correct the uncalibrated gaze direction are explained in [40, 138]. In the former, a user's measured gaze data statistics are similarly compared with reference statistics, however for wearable eye-tracking glasses. And in the latter different peaks in the gaze direction histograms are explicitly extracted and used to obtain calibration parameters.

7.3.3 Results and Discussion

In order to investigate the effects of the implicit gaze calibration, it is necessary to record raw video sequences of the head-eye-tracking system

and process them offline once with a personal calibration of the driver and once without any calibration. The expectation is that the deviation between these two signals decreases if the implicit calibration is applied to the uncalibrated.

In Fig. 7.11, the effects of the implicit calibration procedure are shown for one exemplary driver. First of all, using the final obtained values of the computed shifts on the whole sequence leads to pitch and yaw histograms close to the reference histograms. The Pearson correlation coefficients are 0.97 for the yaw angle histogram and 0.99 for the pitch angle histogram after 5 min of driving. Without implicit calibration, the coefficients are 0.89 and 0.44 respectively. This improvement is not surprising, the results correspond to the maximization of the target function, namely the correlation coefficient. At the same time, the resulting histograms are also closer to the histograms of the gaze signal with personal calibration (cf. Fig. 7.11a). However and also not surprisingly, it can also be observed that the gaze signals are of course only shifted towards the reference signal and not necessarily towards the signal of the calibrated gaze. This can be seen in the fact that in this example, the pitch histogram is overcompensated while the yaw histogram is not. That the calibration offset indeed leads to lower deviations can be seen in the Figs. 7.11b and 7.11c. The absolute value of the deviation to the calibrated gaze signal decreases when the implicit calibration is applied. However, as can be seen in Fig. 7.11b, the simple assumption of a purely additive offset is not fulfilled. If that would be the case, the distributions of the deviations would exhibit one very narrow peak or even a singleton. This however, is not the case as deviations of several degrees can be observed. As outlined above, the presented additive model compensates mainly the large "eyes-on-road"-peak in the histograms and is unable to capture other error sources. E. g. if the driver bends their head to the side, i. e. exhibits a roll movement with their head, the gaze pitch and yaw angles are still output and corrected with respect to the vehicle coordinate system. However, a constant anatomic bias is with respect to the head fixed coordinate system. Thus, the assumption of an always constant bias is violated by a head roll rotation. Moreover, the original as well as the remaining errors in the yaw direction are larger than in pitch direction. This is probably not only because the dynamic range is larger in yaw direction but also because the additive assumption is violated by another circumstance: while the anatomic deviation in the pitch angle between visual and optical axis is in the same direction for both eyes and thus in accordance with the assumption, it is in the diametrically opposed direction for the heading angle (cf. Fig. 2.1 in Section 2.1.1). Therefore, the

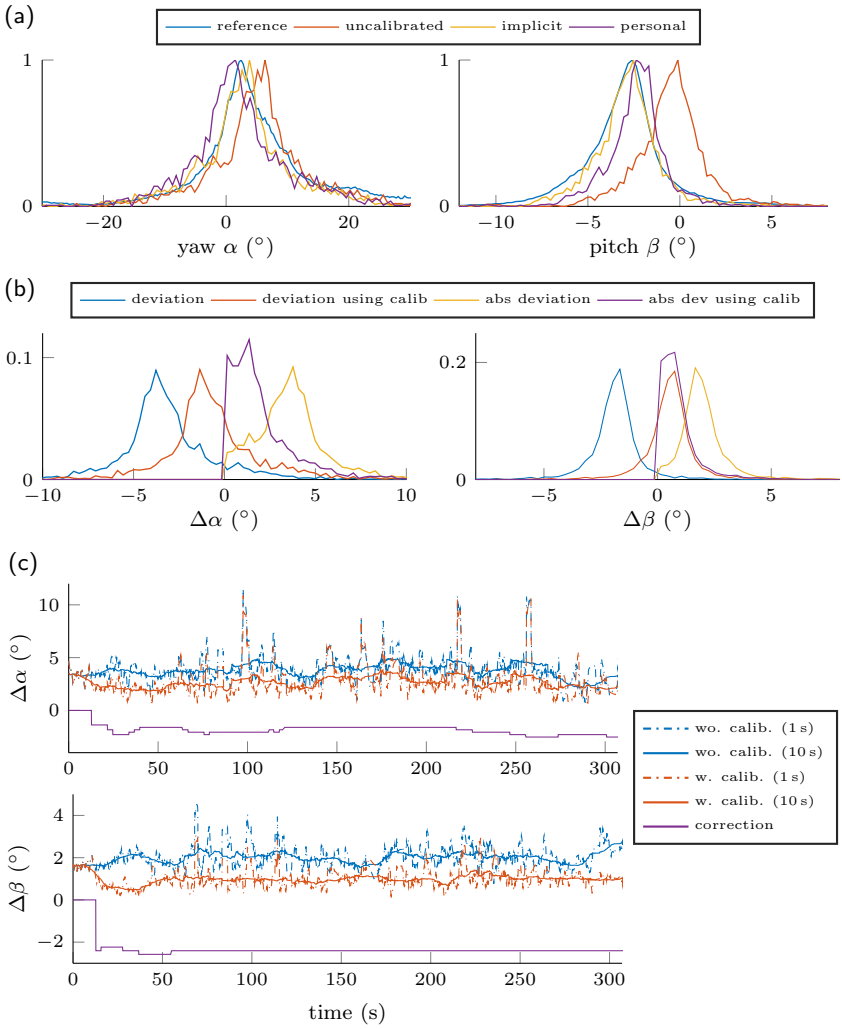


Figure 7.11: Implicit calibration result for one exemplary driver. (a) Normalized histograms of reference signal as well as uncalibrated, implicitly calibrated and personally calibrated signal. (b) Distribution of the deviation and absolute value of the deviation of the uncalibrated and implicitly calibrated signal to the signal with personal calibration. (c) Moving average (1 s and 10 s) of the deviation together with course of the computed shift. For the histograms in (a) and (b), the last obtained correction value is used.

joint gaze vector resulting from left and right gaze direction measurement does not necessarily exhibit a constant angular shift. It is thus interesting to observe that the implicit calibration nevertheless improves the gaze direction measurement.

Tests in preparation of the project's final event have shown that the resulting corrections are within a few degrees, which is a straightforward result: As long as invalid gaze samples, e. g. due to standstill or reversing, are filtered out robustly in the histogram accumulation step, the maximization of the correlation between the histograms should converge. Even though the presented approach worked well at the final event[75], it is clear that the approach does not generalize without further ado. E. g. it is not clear how the reference data is actually affected by the environment and the driven route and how much driving data is necessary so that the reference histograms converge as well as the test histograms. This being said, the proposed approach can be applied to skip a bothersome personal calibration for demonstration purposes of applications such as the PRORETA 4 City Assistant System but for a series application a more sophisticated implementation of an implicit calibration is necessary.

7.4 Conclusion

In conclusion, the PRORETA 4 City Assistant System fuses driver and environment information for an individualized and situation specific driver assistance, targeting a promising direction in the development of future ADAS. The understanding of the driver and their actions opens up the space for new assistance concepts that anticipate driver behavior, give early recommendations and advice on maneuver execution and thus have the potential to become attentive co-pilots. Herein, the PRORETA 4 system represents one possible realization taking the driver's capabilities and preferences as well as their viewing behavior into account. With the implicit calibration procedure, the gaze signal quality can be slightly improved by means of online adaptation. This can be helpful if a precise gaze signal is necessary for the targeted application. However, from the current point of view, it seems more realistic to work with the more robust head pose or larger gaze sectors in a series function. The following and final chapter summarizes the main results of this thesis and points out potential future research directions.

[75] Only one person out of about 15 reported that the gaze related functionality did not work, however, this was due to an overall lack of the head tracking for this person.

8 Conclusion

8.1 Summary

Since human error is still the most frequent accident cause, a joint understanding of the scene *and* the driver is one approach to proper assistance of the driver in case they are still in charge of the vehicle. Thus, the key motivation of this thesis was to contribute to the question how information of the driver's visual behavior can enhance the functionality of future ADAS. Driven by this motivation, the contribution of this thesis followed the four key questions i) how information of gaze and environment can efficiently be combined, ii) at what the driver is looking, iii) how the answer to the previous question can be assessed, and iv) how the newly obtained information can be applied in future assistance functionalities.

The first two content chapters created the theoretical frame of this work by presenting the foundations of psychophysical gaze behavior as well as of probabilistic tracking. In Chapter 2, the anatomy of the human eye, principal gaze motion behavior and visual attention characteristics were introduced, each followed by a review of existing computational models. These reviews underline that all aspects of human gaze have been and still are (probably more than ever before) an active research field. However, preliminary findings have not yet made their way into practical vehicle applications. In contrast, probabilistic tracking approaches are used in many practical applications of various domains, only one of which is e. g. object tracking for ADAS. Chapter 3 provides an introduction to probabilistic Bayesian filtering with a specific focus on the leading towards the method applied in Chapter 5. The goal was to provide the necessary theoretical foundations to readers from various research backgrounds which is why the derivations started from the fundamental optimal Bayesian filter and the basic Kalman filter for linear dynamic systems finally leading to Multiple Model filtering approaches and their computational variants.

The first research question, namely how the driver's gaze and information on the environment can be fused, was tackled in Chapter 4. An intensive literature review was provided, discussing various aspects of the problem

at hand such as the appropriate fusion coordinate system or the trade-off between robustness and sensitivity of a threshold model with different gaze opening angles. First experimental comparisons of simple geometric models led to the consideration of how to deal with uncertainty in the estimation of the driver's gaze target. Together with the vast knowledge from human gaze behavior research, this provided the motivation for a probabilistic modeling approach pursued in Chapter 5. Here, a tracking approach in vehicle coordinates has been presented, which directly allows for visual fixations and saccades through the choice of the filter's computational structure. Using the probabilistic tracking approach, hard decision criteria such as threshold parameters are replaced by a joint consideration of how well the measured gaze direction and environment fit together. The biggest advantage of the proposed method arises from the confidences for each object of being the current gaze target which are directly obtained from the filter result. Nevertheless, it should be noted that the filter is not independent of errors in previous process steps and is therefore not immune to false detections. Especially the object list and the gaze quality are decisive for a proper function. If consistency of gaze and environment is provided, the confidences indicate the reliability of the gaze target estimation – a measure unavailable when using simple geometric models.

At this point, the third key question became relevant. How should be decided how confident a fusion model actually is? In order to gain a more detailed insight, Chapter 6 presented a procedure how the proposed approach can be quantitatively compared with other methods which are all based on the head-eye-tracking system in the test vehicle. Evaluation scenarios with an additional wearable eye tracking device have been recorded, and through manual labeling reference data has been created. Based on this reference data, it could be shown on example scenes that the fusion in vehicle coordinates is more robust concerning the gaze point's depth estimation. Furthermore, comparisons of the tracking approach with different gaze opening angle thresholds suggest that the filter creates a reasonable trade-off between robustness characteristics found in threshold models and, if desired, the determination on a specific target based on the model weights.

After Chapters 5 and 6 focused on the details of how to detect the current gaze target, Chapter 7 considered the driver's gaze more on a system application level and exemplarily showed how the driver's gaze can be integrated into future assistance systems. Within the frame of the research project PRORETA 4, a prototype assistance system for urban scenarios with different use cases was developed and integrated in the test

vehicle. The system operates in real traffic in selected scenarios. As a prototypical use-case, a warning strategy for left-yields-right scenarios was designed, which not only takes into account the driver's gaze behavior when approaching intersections but also the detection of gaze targets.

With the insights obtained from the considerations on gaze and environment data fusion, gaze target estimation on object level in the vehicle's vicinity becomes a feasible task. Together with an appropriate definition of risk and a well designed HMI, the formerly fictitious scenario from the introductory chapter is indeed within reach. Nevertheless, further research is necessary to realize this vision in series applications.

8.2 Future Research

Especially for series production, camera based driver monitoring is still in its first stages. The future task will be to incorporate knowledge about the driver such as an estimation of their situation awareness into the functionality of ADAS. This can start with simple applications such as starting to drive again after a full stop in the ACC functionality only if the driver is attentive. Other examples include more advanced adaptive warning strategies based on the driver's gaze direction such as in the PRORETA 4 City Assistant System, or an extension of the blind spot detection assistance. Moreover, the driver's gaze can also be used for new advances in the HMI conception. Together with augmented reality displays, completely new driver-vehicle interaction is imaginable. As presented in this thesis, there are various approaches in different domains of the scientific literature related to driver monitoring that have yet to assert themselves, both in terms of system integration feasibility and true driver benefit. Of course, this also applies to the results of this work.

As far as the modeling approach is concerned, there remains a need for research on what information is actually relevant for assistance systems, how to efficiently measure the necessary information and how to efficiently represent the driver's perception in a useful mental model. This work uses a more direct approach, where first the gaze targets have been determined from which indicators for the perception of objects can be derived. These perception scores can then be compared with situation-specific rules. Another approach would be, e. g., a purely learning-based approach as indicated in Chapter 2. From learned observations it can be deduced what the driver should be warned against if their gaze behavior deviates from its expectation. Even the mapping of gaze and environment can be based on

learning data, i. e. the association of gaze direction to objects in the scene. The difficulty here, however, would still be the required corresponding labels which would have to be obtained from reference systems. Whether wearable systems like the one used in this work are sufficient for this task remains to be fully clarified. The consideration of what information is necessary inevitably leads to the task of being able to evaluate different approaches against each other at the overall system level. For this, approaches from the field of ergonomics are certainly helpful.

It is emphasized here that the methods and approaches presented in this thesis can be applied to all areas in which a human being monitors a process. Totally different applications are conceivable such as the usage in quality assurance processes. Employees at a production line in charge of identifying deficient items might always miss to look at every single item. Approaches from this thesis can be used to identify these missed items and return them automatically in the process.

In final conclusion, it can be said that remote eye-tracking is about to step out of the laboratory settings, potentially becoming a mass product in the automotive field. Cameras inside vehicles have an enormous untapped potential for new functions in terms of safety and comfort. However, the increasing complexity of the systems and the concrete benefits pose large challenges and still leave much room for future research.

A System Calibration

In order to combine the driver's gaze with the information from the exterior sensors, all data needs to be available in one common coordinate system. The tracking approach in Chapter 5 operates in the 2D vehicle space. Object list and free space spline are already given in that coordinate system. Gaze, however, is collected in camera coordinates of one of the cameras, the main camera, of the head-eye-tracking system. In Section A.1, the calibration of that reference camera with respect to the vehicle coordinate system laying on the middle of the front axis on the ground is presented. In Chapter 6, the tracking approach is compared to the "stereo model". This model operates completely in image coordinates of the exterior camera, so in that case, gaze does not necessarily need to be given in vehicle coordinates but can rather be transformed directly into the camera coordinates of the exterior camera. A possible way to perform this calibration is given in Section A.2.

In general, the transformation between coordinate systems is given by a special Euclidean transformation with rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ and translation vector $\boldsymbol{t} \in \mathbb{R}^3$ describing the geometric relations [111]. The affine coordinate transformation $\boldsymbol{X}_2 = \boldsymbol{R}_{21}\boldsymbol{X}_1 + \boldsymbol{t}_{21}$ of a point $\boldsymbol{X}_1$ can be written as linear matrix multiplication by means of the homogeneous representation

$$\bar{\boldsymbol{X}}_2 = \underbrace{\begin{bmatrix} \boldsymbol{R}_{21} & \boldsymbol{t}_{21} \\ \boldsymbol{0} & 1 \end{bmatrix}}_{\boldsymbol{T}_{21}\in\mathbb{R}^{4\times4}} \bar{\boldsymbol{X}}_1. \tag{A.1}$$

Determining a camera's pose with respect to a calibration target as well as extrinsic calibration of cameras with an overlapping field of view are common procedures from the computer vision area. Furthermore, intrinsic calibration of all cameras is a basic prerequisite and can be performed with common tools like the Camera Calibration Toolbox for Matlab or the respective library from OpenCV. It is therefore presumed that the multiple cameras of the eye-tracking system are completely calibrated and that only the transformation of the reference camera to vehicle coordinates or exterior camera coordinates is to be determined.

A.1 Extrinsic Calibration of the Eye-Tracking System

As mentioned, all exterior sensors are calibrated and referenced to the middle of the front axle at the ground plane. The goal is to obtain the respective transformation for the reference camera of the head-eye-tracking system so that gaze is given in vehicle coordinates[76].

- First, a calibration checkerboard is fixed inside the car so that the corner points are visible in the image of the reference camera. From common computer vision methods, the camera pose in relation to the checkerboard is retrieved.
- Second, a Leica 3D Disto is placed outside of the vehicle so that the same points on the checkerboard can be targeted with the laser to obtain 3D coordinates in the coordinate system of the 3D measuring system. From the same position, the center of the visible front wheel is measured.
- Third, the laser is placed on the other side of the vehicle and again, the points on the checkerboard and the center of the other front wheel are measured. Since it is known that the measurements of the checkerboard points from both sides are actually measurements of the same points and that the two points from the wheel centers lie on a line parallel to the y-axis of the vehicle coordinate system, the pose, i. e. rotation and translation, of the interior camera related to the vehicle coordinate system is completely determined except for the pitch angle, i. e. the rotation around the front axle.
- This pitch angle is obtained by placing a checkerboard (again visible in the camera) orthogonally to the ground plane. Again, the camera pose is obtained from computer vision methods.
- Finally, the height offset between the front axle and the ground plane needs to be compensated.

Given the extrinsic calibration of the reference camera with respect to the vehicle coordinate system, the SmartEye software directly transforms the gaze direction and origin into vehicle coordinates so that it can be fused

[76] Thanks go to Manfred Wilck at Continental, who carried out the calibration several times during the project.

with the environment model. Furthermore, the quality points mentioned in Chapter 6 are also obtained through 3D laser measurements.

A.2 Calibration of Two Cameras without Common Field of View

In order to transform the gaze measured in the interior's camera coordinates into the coordinate system of the exterior ADAS camera, a calibration of the two cameras, also known as "eye-to-eye calibration" is necessary. These two cameras have a disjoint field of view as on camera points on the driver and the other one is directed towards the outer scene. Yet, both cameras are mounted firmly on a rigid body (the car) and hence it is known that the pose transformation between the cameras is fixed.

In [102], different calibration setup categories are outlined, one of which is the calibration by a "calibration rig" with target patterns firmly mounted to the rig and visible in each camera respectively as introduced in [107]. Just like the pose transformation between the cameras denoted with $\boldsymbol{Y} \in \mathbb{R}^{4\times4}$, also the transformation between the target patterns $\boldsymbol{X} \in \mathbb{R}^{4\times4}$ is fixed but unknown. For each image pair taken by the cameras at a certain relative position to the rig, a pair of relative pose estimations $\boldsymbol{A} \in \mathbb{R}^{4\times4}$ and $\boldsymbol{B} \in \mathbb{R}^{4\times4}$ is obtained. Fig. A.1 shows a schematic view of the setup. From the picture, also the equality of the series of pose transformations $\boldsymbol{AX} = \boldsymbol{YB}$ can be seen. Using several measurements, i. e. several image pairs, this equation can be solved iteratively for $\boldsymbol{X}$ and, more importantly, for $\boldsymbol{Y}$, the extrinsic calibration between the two cameras [101, 102, 107].

For this project, a special calibration rig was set up specifically designed for the test vehicle[77] [112]. A depiction of the rig and the usage in the car are given in Fig. A.2. The rig can be rotated around two spatial axes. One joint is built in the mounting and the other axis is given by the pivot mounted wheels on which the rig is based. Through the wheels, it is also possible to change the distances of the calibration boards to the cameras, altogether providing the possibility to realize a variety of poses. In [101], optimization strategies are explained to cope with the narrow interior and restricted fields of views of the cameras which prevent an optimal measurement distribution.

The described calibration procedure was used in the project to verify the

[77] Thanks to Marius Möller for the design of the rig and his support with the calibration of the camera system.

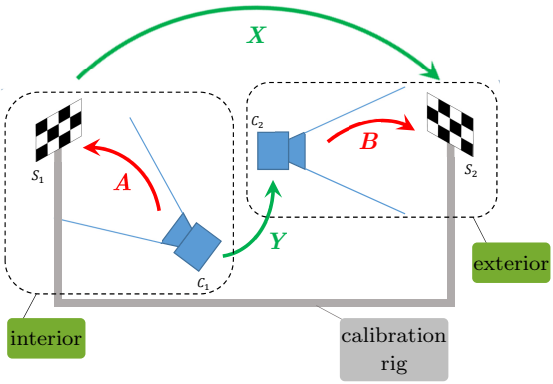
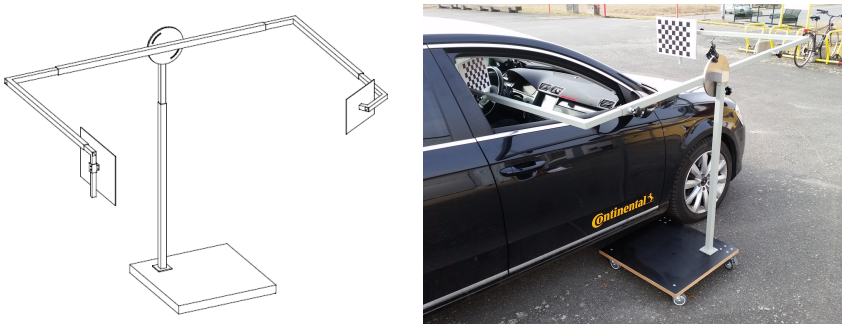


Figure A.1: Calibration setup: The fields of view of the cameras are disjoint. Two calibration targets (checkerboard patterns) are mounted on the calibration rig such that each target is visible in one of the cameras. Since cameras and calibration targets are mounted on rigid bodies (the car and the rig), the transformations $\boldsymbol{X}$ and $\boldsymbol{Y}$ are fixed. The transformations $\boldsymbol{A}$ and $\boldsymbol{B}$ are estimated for taken image pairs. It can be seen, that the transformation sequence from camera 1 to calibration target 2 given by $\boldsymbol{AX}$ must be equal to the sequence $\boldsymbol{YB}$. Image taken and adapted from [112].



(a) Calibration rig. Image taken from [112]. (b) Usage of the calibration rig with the test vehicle.

Figure A.2: The calibration rig is designed such that one calibration pattern is visible in each camera. Through the movable platform and the rotatable joint, the poses with respect to the cameras can be varied.

calibration described in the section before. Since the main development of this thesis concentrated on the tracking algorithm in 2D vehicle coordinates, a calibration of the head-eye-tracking system to the vehicle coordinate system was required.

B Filter Parameters

The parameters of the MHMM gaze target tracking algorithm presented in Chapter 5 are set as follows. The parameters are listed in the order of their occurance.

Symbol	Value	Page
$\mathbf{\Gamma}_{ij}$	$\begin{bmatrix} 0.054 & 0 \\ 0 & 0.054 \end{bmatrix}$	81
$\mathbf{\Gamma}_{i,n+1}$	$\begin{bmatrix} 0.87 & 0 \\ 0 & 0.87 \end{bmatrix}$	81
σ	0.035	84
d_{fs}	0.25	86
ρ	0.5 (0.2 for ML)	86
K	100	88
κ_0	0.9	93
κ_1	0.5	93
κ_2	20.0	93
τ_0	0.15	93
κ_3	0.9	94
κ_4	0.8	94
κ_5	20.0	94
δ_0	7.5	94

C City Assistant System – left-yields-right

Fig. C.1 shows a simplified depiction of the left-yields-right warning strategy in the PRORETA 4 City Assistant System. The warning strategy is altered depending on the driver's driving and viewing behavior in relation to the distance of the ego-vehicle to the intersection. A fallback strategy considers the driver always attentive if the gaze tracking performance is low. A warning to slow down is thrown at any time that the a collision risk is detected. If the driver is not attentive while approaching the intersection, a sound with increasing frequency is started. This sound is stopped as soon as the driver looks to the right. After completion of the situation, the driver obtains either a lauding, admonishing or no feedback depending on their system settings.

The parameters are set as follows:

Parameter	**Value**
speed threshold	26.5 km/h at 30 m
first gaze threshold	20° at 23 m
second gaze threshold	45° at 5 m
gaze opening angle	12°
head opening angle (if no gaze avail.)	24°
minimum fixation duration	250 ms
visually secured intersection threshold	500 ms
collision risk threshold	3 s TTC

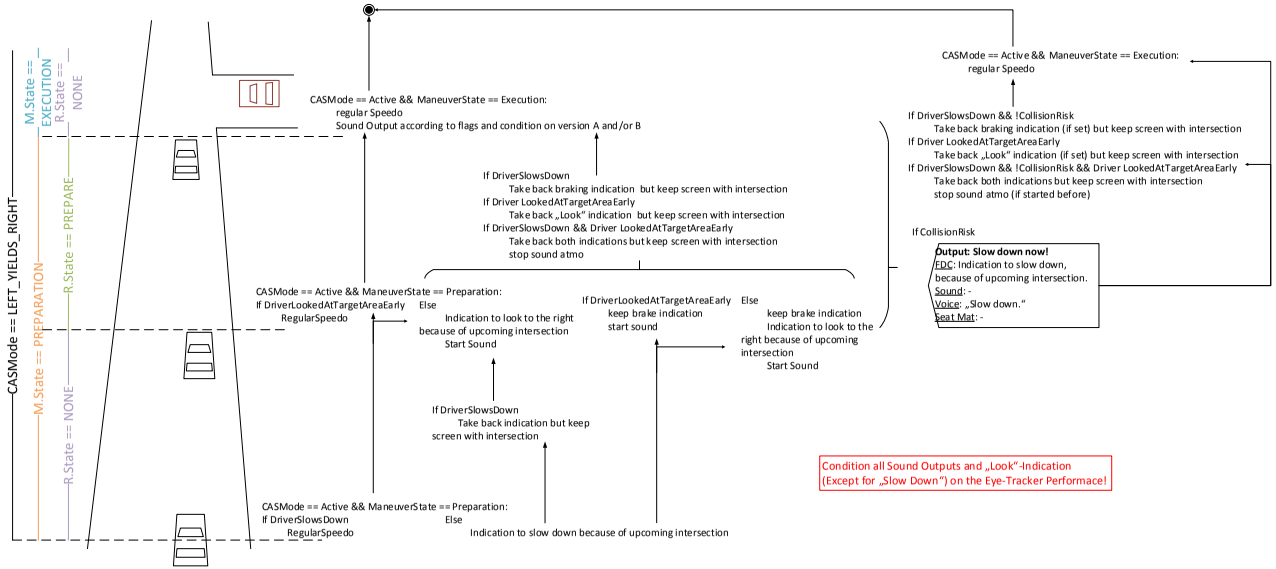


Figure C.1: Simplified depiction of the left-yields-right warning strategy in the PRORETA 4 City Assistant System.

D Publications and Supervisions

D.1 List of Publications by the Author

D.1.1 Journal Publications

- J. Schwehr; S. Luthardt; H. Dang; M. Henzel; H. Winner; J. Adamy; J. Fürnkranz; V. Willert; B. Lattke; C. Wannemacher; M. Höpfl. The PRORETA 4 City Assistant System: Adaptive Maneuver Assistance at Urban Intersections using Driver Behavior Modeling. In *at-Automatisierungstechnik*, 67(9):783 – 798, Sep. 2019.

D.1.2 Conference Publications

- J. Schwehr, M. Knaust, and V. Willert. How to Evaluate Object-of-Fixation Detection. In *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*, pages 570-577, Paris, France, June 2019

- J. Schwehr and V. Willert. Multi-Hypothesis Multi-Model Driver's Gaze Target Tracking. In *Proc. of the IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1427-1434, Maui, HI, USA, Nov. 2018

- J. Schwehr and V. Willert. Tracking des Aufmerksamkeitsziels des Fahrers mittels eines Multi-Hypothesen Multi-Modell Filters. In *12. Uni-DAS e.V. Workshop Fahrerassistenz und automatisiertes Fahren.*, pages 95-105, Walting, Deutschland, Sep. 2018

- M. Buczko, V. Willert, J. Schwehr and J. Adamy. Self-Validation for Automotive Visual Odometry. In *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*, pages 573-578, Changshu, China, June 2018

- J. Schwehr and V. Willert. Driver's Gaze Prediction in Dynamic Automotive Scenes. In *Proc. of the IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1-8, Yokohama, Japan, Nov. 2017
- A. Barth, J. Siegemund, and J. Schwehr. Fast and Precise Localization at Stop Intersections. In *IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pages 82-87, Gold Coast, Australia, June 2013

D.2 List of Supervisions by the Author

- F. Xiong. Entwicklung und Implementierung des JIPDA Trackers zur Reduzierung von FP/FN Objekten. Master's Thesis (administrative supervision), June 2019
- H. Jin. Single Image Depth Prediction Jointly with Semantic Segmentation. Master's Thesis (administrative supervision), Dec. 2018
- Y. Wang. Erlernen eines Fahrer-Blickverhaltensmodells mittels Methoden des Reinforcement Learning. Master's Thesis, Aug. 2018
- F. Euteneuer. Clustering of Driver's Gaze Behavior. Master's Thesis, July 2018
- M. Li. Object Specific Driver Perception using Semantic Segmentation Models. Master's Thesis, June 2018
- G. Ramajayam. Learning of Attention Maps using Fully Convolutional Networks. Master's Thesis, discontinued
- X. Wu. Semantic Representation of Vectors from Gaze Data using Word2vec. Master's Thesis, Nov. 2017
- F. Mogwitz. Entwurf einer Simulationsumgebung für die Evaluierung von Blicksensormodellen. Bachelor's Thesis, Nov. 2017
- C. Jin. LSTM Autoencoder for Clustering of Diver's Gaze Direction. Master's Thesis, Oct. 2017
- A. Büyükkakac. Mustererkennung in Fahrerblickbewegungen. Bachelor's Thesis, Apr. 2017

- R, Chen. Konzeption und Implementierung einer Closed-Loop Testumgebung zur Realisierung und Bewertung komplexer Fahrerassistenzfunktionen im Automobilbereich. Master's Thesis (administrative supervision), Apr. 2017
- M. Möller. Optimierungsverfahren zur Bestimmung geometrischer Abhängigkeiten zwischen zwei Fahrzeugkamerasystemen ohne gemeinsames Sichtfeld. Bachelor's Thesis, Mar. 2017
- K. Dracopulous. Bestimmung der Kopfpose durch visuelle Odometrie mittels TVL1 Fluss. Bachelor's Thesis, Dec. 2016
- F. Schmitt. Modellierung Objektorientierter Wahrnehmung. Bachelor's Thesis, Nov. 2016
- M. Fischer. Detektion und Verfolgung von Gesichtsmerkmalen zur Fahrerbeobachtung unter Einbeziehung der Kopfpose. Diploma's Thesis, Nov. 2016
- D. Kettner. Hand-/Arm-Klassifikation auf Basis von Tiefenbildmessungen. Master's Thesis, Nov. 2016
- W. Ding. Detection and Classification of Secondary Tasks During Driving. Master's Thesis, Sep. 2016
- D. Kraus. Kalibrierung einer Tiefenbildkamera und einer Monokamera. Bachelor's Thesis, July 2016
- N. Maier. Bestimmung der Blickrichtung über Convolutional Neural Networks. Bachelor's Thesis, Apr. 2016
- G. Lechner. 2D-Gesichtsdetektion und -tracking für die Fahrerbeobachtung. Bachelor's Thesis, Mar. 2016
- C. Linnhoff. Bestimmung der 3D Kopfpose in Videodaten mittels eines Bewegungsmodells aus Kugelgelenk und Kalmanfilter. Bachelor's Thesis, Mar. 2016
- J. Lee. Online-Kamerakalibrierung. Bachelor's Thesis, Feb. 2016

Bibliography

[1] B. Abendroth and R. Bruder. Die Leistungsfähigkeit des Menschen für die Fahrzeugführung. In *Handbuch Fahrerassistenzsysteme*, pages 3–15. Springer Vieweg, Wiesbaden, Germany, 2015.

[2] C. Ahlström and K. Kircher. Review of Real-time Visual Driver Distraction Detection Algorithms. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*, pages 2:1–2:4, Eindhoven, The Netherlands, Aug. 2010.

[3] A. Al-Rahayfeh and M. Faezipour. Eye Tracking and Head Movement Detection: A State-of-Art Survey. *IEEE Journal of Translational Engineering in Health and Medicine*, pages 2100212–2100212, Nov. 2013.

[4] A. Alekseenko, H. Dang, G. Bansal, J. Sanchez-Medina, C. Miyajima, T. Hirayama, K. Takeda, and I. Ide. ITS+DM Hackathon (ITSC 2017): Lane departure prediction with naturalistic driving data. *IEEE Intelligent Transportation Systems Magazine*, 2018.

[5] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara. DR(eye)VE: A Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 54–60. Las Vegas, NV, USA, June 2016.

[6] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof. A comparison of scanpath comparison methods. *Behavior Research Methods*, 47(4):1377–1392, Dec. 2015.

[7] N. Apostoloff. *Vision in and out of vehicles: an integrated approach to lane tracking.* Ms thesis, The Australian National University, Canberra, ACT, Australia, 2002.

[8] N. Apostoloff and A. Zelinsky. Vision In and Out of Vehicles: Integrated Driver and Road Scene Monitoring. volume 23, pages 513–538, Apr. 2004.

[9] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Aug. 2002.

[10] D. A. Atchison and G. Smith. *Optics of the human eye*. Butterworth-Heinemann, Oxford, UK, 2000.

[11] T. Bär, D. Linke, D. Nienhuser, and J. M. Zollner. Seen and missed traffic objects: A traffic object-specific awareness estimation. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 31–36, Gold Coast, Australia, June 2013.

[12] T. Bär, D. Nienhüser, R. Kohlhaas, and J. M. Zöllner. Probabilistic driving style determination by means of a situation based analysis of the vehicle data. In *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1698–1703, Washington, D.C., USA, Oct. 2011.

[13] T. Bär, J. F. Reuter, and J. M. Zöllner. Driver head pose and gaze estimation based on multi-template ICP 3-D point cloud alignment. In *15th International IEEE Conference on Intelligent Transportation Systems*, pages 1797–1802, Anchorage, AK, USA, Sep. 2012.

[14] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation: Theory, Algorithms and Software*. Wiley-Interscience, New York, NY, USA, 2001.

[15] D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, Cambridge, UK, 2012.

[16] D. Baron. A brief history of singular 'they'. `https://public.oed.com/blog/a-brief-history-of-singular-they/`, 2018. accessed 05.08.2019.

[17] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner. Three Decades of Driver Assistance Systems: Review and Future Perspectives. *IEEE Intelligent Transportation Systems Magazine*, 6(4):6–22, Oct. 2014.

[18] C. M. Bishop. *Pattern recognition and machine learning*. Springer, New York, NY, USA, 2006.

[19] S. S. Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):5–18, 2004.

[20] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl. Visualization of eye tracking data: A taxonomy and survey. In *Computer Graphics Forum*, volume 36, pages 260–284, Feb. 2017.

[21] G. Boccignone. Advanced statistical methods for eye movement analysis and modelling: a gentle introduction. *arXiv preprint arXiv:1506.07194v4*, Aug. 2017.

[22] A. Borji and L. Itti. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[23] S. Boschenriedter, P. Hossbach, C. Linnhoff, S. Luthardt, and S. Wu. Multi-session visual roadway mapping. In *IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 394–400, Maui, HI, USA, Nov. 2018.

[24] M. Buczko and V. Willert. How to distinguish inliers from outliers in visual odometry for high-speed automotive applications. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 478–483, Gothenburg, Sweden, June 2016.

[25] M. Buczko and V. Willert. Flow-decoupled normalized reprojection error for visual odometry. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1161–1167, Rio de Janeiro, Brasil, Nov. 2016.

[26] G. T. Buswell. How people look at pictures: a study of the psychology and perception in art. 1935.

[27] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci. Online Recognition of Driver-Activity Based on Visual Scanpath Classification. *IEEE Intelligent Transportation Systems Magazine*, 9(4):23–36, Oct. 2017.

[28] R. C. Coetzer and G. P. Hancke. Driver fatigue detection: A survey. In *AFRICON 2009*, pages 1–6, Nairobi, Kenya, Sep. 2009.

[29] M. Corbetta. Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences*, 95(3):831–838, Feb. 1998.

[30] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, Las Vegas, NV, USA, June 2016.

[31] M. Cornia, D. Abati, L. Baraldi, A. Palazzi, S. Calderara, and R. Cucchiara. Attentive Models in Vision: Computing Saliency Maps in the Deep Learning Era. In *AI*IA 2017 Advances in Artificial Intelligence*, pages 387–399. Springer International Publishing, Bari, Italy, Nov. 2017.

[32] H. Dang. *Adaptive Personalization in Driver Assistance Systems*. Dissertation, Technische Universität Darmstadt, Darmstadt, Germany, in preparation.

[33] H. Dang and J. Fürnkranz. Exploiting maneuver dependency for personalization of a driver model. In *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen"*, volume 2191 of *CEUR Workshop Proceedings*, pages 93–97. CEUR-WS.org, 2018.

[34] H. Dang and J. Fürnkranz. Driver information embedding with siamese LSTM networks. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 935–940, Paris, France, June 2019.

[35] H. Dang and J. Fürnkranz. Using past maneuver executions for personalization of a driver model. In *IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 742–748, Maui, HI, USA, Nov. 2018.

[36] H. Dang, J. Fürnkranz, M. Höpfl, and A. Biedermann. Time-to-lane-change prediction with deep learning. In *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Japan, Oct. 2017.

[37] Dominik Haumann. Herleitung des Kalman-Filters. internal document, 2012.

[38] A. Doshi and M. Trivedi. Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 887–892, Xi'an, China, June 2009.

[39] A. T. Duchowski. *Eye tracking methodology: Theory and practice*. Springer, London, UK, 2009.

[40] T. J. H. Edwards and N. J. Langdale-Smith. Automatic calibration of a gaze direction algorithm from user behavior, Mar. 17, 2015. US Patent 8982046.

[41] M. R. Endsley. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[42] J. Engström, G. Markkula, and N. Merat. Modelling the effect of cognitive load on driver reactions to a braking lead vehicle: A computational account of the cognitive control hypothesis. In *Proceedings of the Fifth International Conference on Driver Distraction and Inattention, Paris*, Paris, France, Mar. 2017.

[43] Euro NCAP. *Euro NCAP 2025 Roadmap: In Pursuit Of Vision Zero*. Leuven, Belgium, 2017.

[44] European Commission. *Advanced driver assistance systems*. Nov. 2016. European Commission, Directorate General for Transport.

[45] N. Fecher and J. Hoffmann. Fahrerwarnelemente. In *Handbuch Fahrerassistenzsysteme*, pages 675–685. Springer Vieweg, Wiesbaden, Germany, 2015.

[46] S. Feuerstack and B. Wortelen. A model-driven tool for getting insights into car drivers' monitoring behavior. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 861–868, Los Angeles, CA, USA, June 2017.

[47] L. Fletcher and A. Zelinsky. Driver Inattention Detection based on Eye Gaze—Road Event Correlation. *The International Journal of Robotics Research*, 28(6):774–801, May 2009.

[48] V. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello. Bayesian filtering for location estimation. *IEEE Pervasive Computing*, 2(3):24–33, 2003.

[49] L. Fridman, J. Lee, B. Reimer, and T. Victor. 'Owl' and 'Lizard': Patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision*, 10(4):308–314, June 2016.

[50] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundations. *ACM Transactions on Applied Perception*, 7(1):1–39, Jan. 2010.

[51] H. Gao, A. Yüce, and J. Thiran. Detecting emotional stress from facial expressions for driving safety. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5961–5965, Paris, France, Oct. 2014.

[52] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, Providence, RI, USA, June 2012.

[53] Y. Gu and M. Veloso. Multi-model Motion Tracking Under Multiple Team Member Actuators. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 449–456, Hakodate, Japan, May 2006.

[54] 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 International Conference on Human System Interactions (HSI)*, pages 264–269, Ulsan, South Korea, July 2017.

[55] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, June 2006.

[56] R. I. Hammoud, editor. *Passive Eye Monitoring: Algorithms, Applications and Experiments*. Signals and Commmunication Technologies. Springer-Verlag Berlin Heidelberg, Germany, 2008.

[57] D. W. Hansen and Q. Ji. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, Feb. 2010.

[58] M. M. Hayhoe and C. A. Rothkopf. Vision in the natural world. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(2):158–166, Sep. 2011.

[59] F. Heinrich. *Vorhersage der Fahrerbelastung während der Fahrt.* Diplomarbeit, Universität Stuttgart, Stuttgart, Germany, Sep. 2012.

[60] J. Helmert and J. Marx. Schneller als Gedanken: Blickbewegungen bei Gefahr! Wettbewerbsbeitrag zum Thema „Tempo–die beschleunigte Welt. Institut für Psychologie III an der Technischen Universität Dresden, 2003.

[61] C. Hennessey and P. Lawrence. Noncontact binocular eye-gaze tracking for point-of-gaze estimation in three dimensions. *IEEE transactions on bio-medical engineering*, 56(3):790–799, Mar. 2009.

[62] M. Henzel. *Analyse der Generalisierbarkeit von maschinell gelernten Algorithmen in Fahrerassistenzsystemen.* Dissertation, Technische Universität Darmstadt, Darmstadt, Germany, 2019. in preparation.

[63] M. Henzel, H. Winner, and B. Lattke. Herausforderungen in der Absicherung von Fahrerassistenzsystemen bei der Benutzung maschinell gelernter und lernender Algorithmen. In *11. Workshop Fahrerassistenzsysteme und automatisiertes Fahren*, pages 136–148, Walting, Germany, Mar. 2017.

[64] T. Hirayama, K. Mase, C. Miyajima, and K. Takeda. Classification of Driver's Neutral and Cognitive Distraction States Based on Peripheral Vehicle Behavior in Driver's Gaze Transition. *IEEE Transactions on Intelligent Vehicles*, 1(2):148–157, June 2016.

[65] M. Höffken, E. Tarayan, U. Kresel, and K. Dietmayer. Stereo vision-based driver head pose estimation. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 253–260, Dearborn, MI, USA, June 2014.

[66] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer. *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press , 2011.

[67] K. Holmqvist, M. Nyström, and F. Mulvey. Eye Tracker Data Quality: What It is and How to Measure It. In *ETRA '12 Proceedings of the Symposium on Eye Tracking Research and Applications* , pages 45–52, Santa Barbara, CA, USA, Mar. 2012.

[68] D. Hoppe and C. A. Rothkopf. Learning rational temporal eye movement strategies. *Proceedings of the National Academy of Sciences of the United States of America*, 113(29):8332–8337, July 2016.

[69] J. Hou, G. F. List, and X. Guo. New Algorithms for Computing the Time-to-collision in Freeway Traffic Simulation Models. *Computational Intelligence and Neuroscience*, 2014:57:57, Jan. 2014.

[70] S. Hu and G. Zheng. Driver drowsiness detection with eyelid related parameters by Support Vector Machine. *Expert Systems with Applications*, 36(4):7651–7658, May 2009.

[71] International Organization for Standardization. ISO/DIS 15007 Road vehicles – Measurement and analysis of driver visual behaviour with respect to transport information and control systems.

[72] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 20(11):1254–1259, Nov. 1998.

[73] M. Jabon, J. Bailenson, E. Pontikakis, L. Takayama, and C. Nass. Facial expression analysis for predicting unsafe driving behavior. *IEEE Pervasive Computing*, 10(4):84–95, Apr. 2011.

[74] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3182–319, Santiago, Chile, Dec. 2015.

[75] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena. Brain4Cars: Car That Knows Before You Do via Sensory-Fusion Deep Learning Architecture. *arXiv preprint arXiv:1601.00740*, Jan. 2016.

[76] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena. Recurrent Neural Networks for Driver Activity Anticipation via Sensory-Fusion Architecture. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3125, Stockholm, Sweden, May 2016.

[77] S. Jha and C. Busso. Probabilistic Estimation of the Gaze Region of the Driver using Dense Classification. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 697–702, Maui, HI, USA, Nov. 2018.

[78] Jinwoo Kim, Kitae Kim, Daesub Yoon, Yongbon Koo, and Wooyong Han. Fusion of driver-information based driver status recognition for co-pilot system. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1398–1403, Gothenburg, Sweden, June 2016.

[79] J. Jo. Vision-based method for detecting driver drowsiness and distraction in driver monitoring system. *Optical Engineering*, 50(12):1 – 25, Dec. 2011.

[80] L. Johnson, B. Sullivan, M. Hayhoe, and D. Ballard. Predicting human visuomotor behaviour in a driving task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1636), Feb. 2014.

[81] S. J. Julier and J. K. Uhlmann. New extension of the Kalman filter to nonlinear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI*, volume 3068, pages 182 – 193, July 1997.

[82] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, Mar. 1960.

[83] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt. Driver Behavior Analysis for Safe Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3017–3032, 2015.

[84] E. Kasneci, G. Kasneci, T. C. Kübler, and W. Rosenstiel. Online Recognition of Fixations, Saccades, and Smooth Pursuits for Automated Analysis of Traffic Hazard Perception. In *Artificial Neural Networks: Methods and Applications in Bio-/Neuroinformatics*, pages 411–434. Springer International Publishing, Cham, Switzerland, 2015.

[85] K. Kircher. Issues related to the driver distraction detection algorithm AttenD. In *First International Conference on Driver Distraction and Inattention*, volume Sep. 2009. Gothenburg, Sweden.

[86] K. Kircher and C. Ahlstrom. Minimum Required Attention: A Human-Centered Approach to Driver Inattention. *Human factors*, 59(3):471–484, Oct. 2017.

[87] R. Klette. Vision-based Driver Assistance Systems. `researchgate.net/publication/272199860_Vision-based_Driver_Assistance_Systems`, Feb. 2015. accessed 16.07.2018.

[88] R. Klinke, H.-C. Pape, and S. Silbernagl. *Physiologie*. Thieme, Stuttgart, Germany, 2005.

[89] A. Kolli, A. Fasih, F. A. Machot, and K. Kyamakya. Non-intrusive car driver's emotion recognition using thermal camera. In *Proceedings of the Joint INDS'11 ISTET'11*, pages 1–5, Klagenfurt, Austria, July 2011.

[90] E. Kowler. Eye movements: The past 25years. *Vision Research*, 51(13):1457–1483, July 2011.

[91] 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 (IV)*, pages 1245–1250, Dearborn, MI, USA, June 2014.

[92] A. Kraft, C. Maag, A. Neukum, and M. Baumann. Mensch-Maschine-Interaktion bei manuellem und automatisiertem kooperativen Fahren an Auffahrten und Kreuzungen. In *12. Workshop Fahrerassistenz und automatisiertes Fahren*, pages 56–66, Walting, Germany, Sep. 2018.

[93] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, Mar. 1951.

[94] M. Kutila. *Methods for Machine Vision Based Driver Monitoring Applications*. Dissertation, Tampere University of Technology, Tampere, Finland, Dec. 2006.

[95] M. Land and B. Tatler. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, 2009.

[96] 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 International Conference on Robotics and Automation (ICRA)*, pages 3167–3173, Stockholm, Sweden, May 2016.

[97] O. Lappi. Eye movements in the wild: Oculomotor control, gaze behavior & frames of reference. *Neuroscience & Biobehavioral Reviews*, 69:49–68, Oct. 2016.

[98] O. Lappi, P. Rinkkala, and J. Pekkanen. Systematic Observation of an Expert Driver's Gaze Strategy—An On-Road Case Study. *Frontiers in psychology*, 8:620, Apr. 2017.

[99] F. Lethaus and J. Rataj. Do eye movements reflect driving manoeuvres? *IET Intelligent Transport Systems*, 1(3):199, Sep. 2007.

[100] N. Li, J. J. Jain, and C. Busso. Modeling of Driver Behavior in Real World Scenarios Using Multiple Noninvasive Sensors. *IEEE Transactions on Multimedia*, 15(5):1213–1225, Aug. 2013.

[101] Z. Li and V. Willert. Subtleties of extrinsic calibration of cameras with non-overlapping fields of view. *tm-Technisches Messen Plattform für Methoden, Systeme und Anwendungen der Messtechnik*, 86(7-8):433 – 442, June 2019.

[102] Z. Li and V. Willert. Eye-to-eye Calibration for Cameras with Disjoint Fields of View. pages 2631–2638, Nov. 2018.

[103] Y. Liang, M. L. Reyes, and J. D. Lee. Real-Time Detection of Driver Cognitive Distraction Using Support Vector Machines. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):340–350, June 2007.

[104] M. Liebner. *Fahrerabsichtserkennung und Risikobewertung für warnende Fahrerassistenzsysteme*. Dissertation, Karlsruhe, Germany, 2016.

[105] M. Liebner and F. Klanner. Fahrerabsichtserkennung und Risikobewertung. In *Handbuch Fahrerassistenzsysteme*, pages 701–719. Springer Vieweg, Wiesbaden, Germany, 2015.

[106] M. Liebner, F. Klanner, and C. Stiller†. Der Fahrer im Mittelpunkt – Eye-Tracking als Schlüssel zum mitdenkenden Fahrzeug? In *8. Workshop Fahrerassistenzsysteme und automatisiertes Fahren*, pages 87–96. Walting, Germany, Sep. 2012.

[107] Z. Liu, G. Zhang, Z. Wei, and J. Sun. A global calibration method for multiple vision sensors based on multiple targets. *Measurement Science and Technology*, 22(12):125102, 2011.

[108] S. Luthardt, C. Han, V. Willert, and M. Schreier. Efficient graph-based V2V free space fusion. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 985–992, Los Angeles, CA, USA, June 2017.

[109] S. Luthardt, V. Willert, and J. Adamy. LLama-SLAM: Learning high-quality visual landmarks for long-term mapping and localization.

In *IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2645–2652, Maui, HI, USA, Nov. 2018.

[110] S. Luthardt, C. Ziegler, and V. Willert. How to match tracks of visual features for automotive long-term-SLAM. In *IEEE 22nd International Conference on Intelligent Transportation Systems (ITSC)*, Auckland, New Zealand, Oct. 2019. (Submitted).

[111] Y. Ma, S. Soatta, J. Kosecka, and S. S. Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, New York, NY, USA, 2004.

[112] Marius Möller. *Optimierungsverfahren zur Bestimmung geometrischer Abhängigkeiten zwischen zwei Fahrzeugkamerasystemen ohne gemeinsames Sichtfeld*. Bachelor Thesis, Technische Universität Darmstadt, Darmstadt, Germany, Mar. 2013.

[113] M. Martin, J. Popp, M. Anneken, M. Voit, and R. Stiefelhagen. Body Pose and Context Information for Driver Secondary Task Detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 2015–2021, Chanshu, China, June 2018.

[114] S. Martin, E. Ohn-Bar, and M. M. Trivedi. Automatic critical event extraction and semantic interpretation by looking-inside. In *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2274–2279. Las Palmas, Spain, Sep. 2015.

[115] S. Martin, A. Rangesh, E. Ohn-Bar, and M. M. Trivedi. The rhythms of head, eyes and hands at intersections. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1410–1415, Gothenburg, Sweden, June 2016.

[116] S. Martin and A. Tawari. Object of Fixation Estimation by Joint Analysis of Gaze and Object Dynamics. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 2042–2047. Changshu, China, June 2018.

[117] S. Martin and M. M. Trivedi. Gaze fixations and dynamics for behavior modeling and prediction of on-road driving maneuvers. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1541–1545, Los Angeles, CA, USA, June 2017.

[118] P. S. Maybeck. Stochastic models, estimation, and control, 1982.

[119] J. C. McCall and M. M. Trivedi. Driver Behavior and Situation Aware Brake Assistance for Intelligent Vehicles. *Proceedings of the IEEE*, 95(2), Apr. 2009.

[120] B. Metz. Ist der Fahrer aufmerksam? Vorstellung eines Modells zur Beschreibung und Bewertung des Blickverhaltens des Fahrers. In *Der Fahrer im 21. Jahrhundert : Fahrer, Fahrerunterstützung und Bedienbarkeit; 8. VDI-Tagung*, Braunschweig, Germany, Nov. 2015.

[121] T. P. Minka. *A family of algorithms for approximate Bayesian inference*. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2001.

[122] 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 *2016 12th International Conference on Intelligent Environments (IE)*, pages 198–201, London, UK, Sep. 2016.

[123] M. Mori, C. Miyajima, P. Angkititrakul, T. Hirayama, Y. Li, N. Kitaoka, and K. Takeda. Measuring driver awareness based on correlation between gaze behavior and risks of surrounding vehicles. In *15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 644–647, Anchorage, AK, USA, Sep. 2012.

[124] K. P. Murphy. *Machine learning: A probabilistic perspective*. MIT Press, Cambridge, MA, USA, 2012.

[125] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi. Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation. In *IEEE 10th Intelligent Transportation Systems Conference (ITSC)*, pages 709–714, Seattle, WA, USA, Sep. 2007.

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

[127] D. Noton and L. Stark. Eye movements and visual perception. *Scientific American*, 224(6):34–43, June 1971.

[128] M. Nowakowski, M. Sheehan, D. Neal, and A. V. Goncharov. Investigation of the isoplanatic patch and wavefront aberration along the

pupillary axis compared to the line of sight in the eye. *Biomedical Optics Express*, 3(2):240–258, Feb. 2012.

[129] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi. Vision on Wheels: Looking at Driver, Vehicle, and Surround for On-Road Maneuver Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 185–190, Columbus, OH, USA, June 2014.

[130] E. Ohn-Bar and M. M. Trivedi. Are all objects equal? Deep spatio-temporal importance prediction in driving videos. *Pattern Recognition*, 64:425–436, Apr. 2017.

[131] E. Ohn-Bar and M. M. Trivedi. Looking at Humans in the Age of Self-Driving and Highly Automated Vehicles. *IEEE Transactions on Intelligent Vehicles*, pages 90–104, Mar. 2016.

[132] E. Ohn-Bar and M. M. Trivedi. Beyond just keeping hands on the wheel: Towards visual interpretation of driver hand motion patterns. In *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1245–1250. Qingdao, China, Nov. 2014.

[133] D. Orth, D. Kolossa, M. S. Paja, K. Schaller, A. Pech, and M. Heckmann. A maximum likelihood method for driver-specific critical-gap estimation. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 553–558, Los Angeles, CA, USA, June 2017.

[134] 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*, 41(7):1720–1733, 2019.

[135] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara. Learning where to attend like a human driver. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 920–925. Los Angeles, CA, USA, June 2017.

[136] K. B. Petersen and M. S. Pedersen. The matrix cookbook, Nov. 2012.

[137] L. Petersson, L. Fletcher, and A. Zelinsky. A framework for driver-in-the-loop driver assistance systems. In *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005*, pages 771–776, Vienna, Austria, Sep. 2005.

[138] M. Pilu. Self-calibration for an eye-tracker, Feb. 2, 2010. US Patent 7657062.

[139] M. Plavsic, G. Klinker, and H. Bubb. Situation awareness assessment in critical driving situations at intersections by task and human error analysis. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 20(3):177–191, Mar. 2010.

[140] C. M. Privitera. The Scanpath Theory: its definition and later developments. In *Human Vision and Electronic Imaging XI*, volume 6057, pages 87 – 91, 2006.

[141] H. Rahman, S. Barua, and B. Shahina. Intelligent Driver Monitoring Based on Physiological Sensor Signals: Application Using Camera. In *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2637–2642, Las Palmas, Spain, Sep. 2015.

[142] A. Rangesh, N. Deo, K. Yuen, K. Pirozhenko, P. Gunaratne, H. Toyoda, and M. M. Trivedi. Exploring the Situational Awareness of Humans inside Autonomous Vehicles. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 190–197, Maui, HI, USA, Nov. 2018.

[143] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 264–269, Los Angeles, CA, USA, July 2017.

[144] K. Rayner. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychologiacal Bulletin*, 124(3):372–422, 1998.

[145] M. Rezaei and R. Klette. Simultaneous analysis of driver behaviour and road condition for driver distraction detection. *International Journal of Image and Data Fusion*, 2(3):217–236, July 2011.

[146] M. Rezaei and R. Klette. Look at the Driver, Look at the Road: No Distraction! No Accident! In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 129–136, Columbus, OH, USA, June 2014.

[147] M. Roth, F. Flohr, and D. M. Gavrila. Driver and pedestrian awareness-based collision risk analysis. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 454–459, Gothenburg, Sweden, June 2016.

[148] SAE International - On-Road Automated Driving (ORAD) committee. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, June 2018. Technical Report.

[149] F. Sagberg, Selpi, G. F. Bianchi Piccinini, and J. Engström. A review of research on driving styles and road safety. *Human factors*, 57(7):1248–1275, June 2015.

[150] P. M. Salmon, N. A. Stanton, and K. L. Young. Situation awareness on the road: review, theoretical and methodological issues, and future directions. *Theoretical Issues in Ergonomics Science*, 13(4):472–492, May 2011.

[151] S. Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

[152] B. Schlag and G. Weller. Verhaltenswissenschafltiche Aspekte von Fahrerassistenzsystemen. In *Handbuch Fahrerassistenzsysteme*, pages 71–83. Springer Vieweg, Wiesbaden, Germany, 2015.

[153] M. Schreier. *Bayesian Environment Representation, Prediction and Criticality Assessment for Driver Assistance Systems*. Dissertation, Technische Universität Darmstadt, 2015.

[154] M. Schreier, V. Willert, and J. Adamy. Compact Representation of Dynamic Driving Environments for ADAS by Parametric Free Space and Dynamic Object Maps. *IEEE Transactions on Intelligent Transportation Systems*, 17(2):367–384, Feb. 2016.

[155] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. Eye movements and perception: A selective review. *Journal of Vision*, 11(5):9, Sep. 2011.

[156] J. Schwehr, M. Knaust, and V. Willert. How to Evaluate Object-of-Fixation Detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 570–577. Paris, France, June 2019.

[157] J. Schwehr, S. Luthardt, H. Dang, M. Henzel, H. Winner, J. Adamy, J. Fürnkranz, V. Willert, B. Lattke, C. Wannemacher, and M. Höpfl. The PRORETA 4 City Assistant System: Adaptive Maneuver Assistance at Urban Intersections using Driver Behavior Modeling. *at-Automatisierungstechnik Methoden und Anwendungen der Steuerungs-, Regelungs-und Informationstechnik*, 67(9):783 – 798, Sep. 2019.

[158] J. Schwehr and V. Willert. Multi-Hypothesis Multi-Model Driver's Gaze Target Tracking. In *IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1427–1434. Maui, HI, USA, Nov. 2018.

[159] J. Schwehr and V. Willert. Driver's Gaze Prediction in Dynamic Automotive Scenes. In *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Japan, Oct. 2017.

[160] J. Schwehr and V. Willert. Tracking des Aufmerksamkeitsziels des Fahrers mittels eines Multi-Hypothesen Multi-Modell Filters. In *12. Workshop Fahrerassistenz und automatisiertes Fahren*, pages 95–105, Walting, Germany, Sep. 2018.

[161] SensoMotoric Instruments GmbH (SMI). *iViewETG User Guide*. Berlin, Germany, 2016.

[162] N. Shimkin. Estimation and Identification in Dynamical Systems: Multi-Model State Estimation: Lecture Notes, Fall 2009, 2009.

[163] D. Simon. *Optimal state estimation: Kalman, H Infinity, and nonlinear approaches*. Wiley-Interscience, Hoboken, NJ, USA, 2006.

[164] D. J. Simons and C. F. Chabris. Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events. *Perception*, 28(9):1059–1074, Sep. 1999.

[165] D. J. Simons and D. T. Levin. Change blindness. *Trends in Cognitive Sciences*, 1(7):261–267, Oct. 1997.

[166] S. Sivaraman and M. M. Trivedi. Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1773–1795, Dec. 2013.

[167] H. W. Sorenson and D. L. Alspach. Recursive bayesian estimation using gaussian sums. *Automatica*, 7(4):465–479, July 1971.

[168] N. Sprague and D. Ballard. Eye movements for reward maximization. In *Advances in neural information processing systems*, pages 1467–1474, 2004.

[169] N. Sprague, D. Ballard, and A. Robinson. Modeling Embodied Visual Behaviors. *ACM Transactions on Applied Perception*, 4(2), July 2007.

[170] Statistisches Bundesamt. Verkehr: Verkehrsunfälle 2017. Fachserie 8 Reihe 7. 2018.

[171] Statistisches Bundesamt. Verkehr: Verkehrsunfälle 2018. Fachserie 8 Reihe 7. 2019.

[172] Stefan Luthardt. *Erfahrungsbasierte visuelle Lokalisierung für intelligente Fahrzeuge*. Dissertation, Technische Universität Darmstadt, Darmstadt, Germany, in preparation.

[173] B. T. Sullivan, L. Johnson, C. A. Rothkopf, D. Ballard, and M. Hayhoe. The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of Vision*, 12(13):19, Dec. 2012.

[174] B. T. Sullivan, L. M. Johnson, D. H. Ballard, and M. M. Hayhoe. A modular reinforcement learning model for human visuomotor behaviour in a driving task. In *Proceedings of the AISB 2011 Symposium on Architectures for Active Vision*, pages 33–40, York, UK, Apr. 2011.

[175] Y. Sun and X. Yu. An Innovative Nonintrusive Driver Assistance System for Vital Signal Monitoring. *IEEE Journal of Biomedical and Health Informatics*, 18(6):1932–1939, Nov. 2014.

[176] E. Tafaj, G. Kasneci, W. Rosenstiel, and M. Bogdan. Bayesian Online Clustering of Eye Movement Data. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 285–288, Santa Barbara, CA, USA, Mar. 2012.

[177] E. Tafaj, T. Kübler, G. Kasneci, W. Rosenstiel, and M. Bogdan. Online Classification of Eye Tracking Data for Automated Analysis of Traffic Hazard Perception. In *Artificial Neural Networks and Machine Learning – ICANN 2013*, pages 442–450. Springer-Verlag Berlin Heidelberg, 2013.

[178] K. Takagi, H. Kawanaka, M. S. Bhuiyan, and K. Oguri. Estimation of a three-dimensional gaze point and the gaze target from the road images. In *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 526–531, Washington, D.C., USA, Oct. 2011.

[179] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5), May 2011.

[180] A. Tawari and B. Kang. A Computational Framework for Driver's Visual Attention Using A Fully Convolutional Architecture. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 887–894. Los Angeles, CA, USA, June 2017.

[181] A. Tawari, P. Mallela, and S. Martin. Learning to Attend to Salient Targets in Driving Videos Using Fully Convolutional RNN. In *21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3225–3232, Maui, HI, USA, Nov. 2018.

[182] A. Tawari, A. Møgelmose, S. Martin, T. B. Moeslund, and M. M. Trivedi. Attention estimation by simultaneous analysis of viewer and view. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1381–1387, Qingdao, China, Nov. 2014.

[183] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer. Looking-in and looking-out vision for Urban Intelligent Assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 115–120, Dearborn, MI, USA, June 2014.

[184] A. Tawari and M. M. Trivedi. Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 344–349, Dearborn, MI, USA, June 2014.

[185] A. Tawari and M. M. Trivedi. Face Expression Recognition by Cross Modal Data Association. *IEEE Transactions on Multimedia*, 15(7):1543–1552, Nov. 2013.

[186] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT Press, Cambridge, MA, USA, 2006.

[187] M. H. Tong, O. Zohar, and M. M. Hayhoe. Control of gaze while walking: task structure, reward, and uncertainty. *Journal of Vision*, 17(1):28, Jan. 2017.

[188] J. R. Treat, N. S. Tumbas, S. T. McDonald, D. Shinar, R. D. Hume, R. E. Mayer, R. L. Stansifer, and N. J. Castellan. Tri-level study of the causes of traffic accidents: final report. Executive summary. 1979.

[189] M. M. Trivedi, T. Gandhi, and J. McCall. Looking-In and Looking-Out of a Vehicle: Computer-Vision-Based Enhanced Vehicle Safety. *IEEE Transactions on Intelligent Transportation Systems*, 8(1):108–120, Mar. 2007.

[190] Volker Willert. Derivation of the Kalman Filter Equations in a Nutshell. internal document, 2008.

[191] E. A. Wan and R. V. D. Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium*, pages 153–158, Lake Louise, Alberta, Canada, Oct. 2000.

[192] E. A. Wan and R. van der Merwe. *The Unscented Kalman Filter*, chapter 7, pages 221–280. John Wiley & Sons, New York, NY, USA, 2002.

[193] K. Wang, S. Wang, and Q. Ji. Deep Eye Fixation Map Learning for Calibration-free Eye Gaze Tracking. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 47–55, Charleston, SC, USA, Mar. 2016.

[194] Q. Wang, J. Yang, M. Ren, and Y. Zheng. Driver Fatigue Detection: A Survey. In *2006 6th World Congress on Intelligent Control and Automation*, pages 8587–8591, Dalian, China, June 2006.

[195] H. Winner. Quo vadis, FAS? In *Handbuch Fahrerassistenzsysteme*, pages 1167–1186. Springer Vieweg, Wiesbaden, Germany, 2015.

[196] Y. Xia, D. Zhang, A. Pozdnukhov, K. Nakayama, K. Zipser, and D. Whitney. Training a network to attend like human drivers saves it from common but misleading loss functions. *arXiv preprint arXiv:1711.06406*, 2017.

[197] A. L. Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, Boston, MA, USA, 1967.

[198] 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)*, pages 583–588, Los Angeles, CA, USA, June 2017.

[199] S. M. Zabihi, S. S. Beauchemin, and M. A. Bauer. Real-time driving manoeuvre prediction using IO-HMM and driver cephalo-ocular behaviour. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 875–880, Los Angeles, CA, USA, June 2017.

[200] S. M. Zabihi, S. S. Beauchemin, E. A. M. de Medeiros, and M. A. Bauer. Frame-rate vehicle detection within the attentional visual area of drivers. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 146–150, Dearborn, MI, USA, June 2014.

[201] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, Honolulu, HI, USA, July 2017.

Index