OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

Visual recognition systems in a car passenger compartment with the focus on facial driver identification

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von M. Sc. Andrey Makrushin

geb. am 16.02.1980 in Moskau, Russland

Gutachterinnen/Gutachter

Prof. Dr.-Ing. Jana Dittmann
Prof. Dr. Sabah Jassim
Prof. Dr. rer. nat. Reiner Creutzburg

Magdeburg, den 22.01.2014

Andrey Makrushin:
Visual recognition systems in a car passenger compartment with the focus on facial driver identification,
Dissertation,
Otto-von-Guericke University of Magdeburg,
2014.

Acknowledgements

First and foremost I would like to express my deep thanks to my supervisor Prof. Jana Dittmann for motivating me to start working towards a PhD degree and for giving me the opportunity to join her research group. Moreover, I thank her for continuously supporting me not only academically but also financially by enabling me to work on diverse projects and generally for my time with the Advanced Multimedia and Security Lab (AMSL). My deep thanks also go to Prof. Claus Vielhauer for evoking my interest in biometrics and encouraging me to start working on face recognition. I thank both Jana and Claus for all that I learned from them, for putting me on the right track regarding my research activities, for fruitful discussions and insightful advice. I am very grateful to all my colleagues from AMSL and Brandenburg University of Applied Science for the time discussing aspects of pattern recognition, working together on projects and papers and generally for the time we spent together. I especially appreciate the help from Tobias Scheidat, Christian Krätzer, Michael Biermann, Mario Hildebrandt, Ronny Merkel, Kun Qian and Maik Schott.

I am very grateful to Dr. Mirko Langnickel and Dr. Katharina Seifert from Volkwagen Group Research for the joint work on the VisMes project that evoked my interest in visual recognition systems in the automotive domain. I also appreciate the help of my former students Robert Krauße, Enrico Herrmann, Andy Bertz and Sebastian Kleinau. Our joint work on the project provided me with numerous new ideas.

I would like to thank the reviewers Prof. Sabah Jassim and Prof. Reiner Creutzburg, who kindly agreed to read and evaluate my thesis.

I thank all whose who agreed to participate in the experiments and acted as donors by gathering data for our visual databases.

Special thanks go to my English teachers Marcelo Kauer, Tim Sadler and Hans Harald Huber, who helped me to tremendously improve my language skills making it possible to prepare this manuscript.

Last but not least, I would like to express my deep thanks to my family and all my friends for their moral support throughout the long time working on the thesis. They always believed in my potential to finish the PhD study even at moments when I did not. Certainly, I cannot mention everyone. Katharina Franke, Sergei Horbach, Tatjana Bobach-Poltoratski, Dr. Yuriy Tsepkovskiy, Dr. Andriy Telesh, Dr. Dmitry Vlasenko and Valery Makhavikou are those who continuously supported me in struggling with misfortune and frustrations and who were always ready to talk about my ideas and excitements. Frankly, I greatly appreciate this.

Andrey Makrushin
Magdeburg, 2014

Abstract (in German)

Die Ausrüstung der modernen Fahrzeuge mit Kameratechnik ist ein neuer Trend der Automobilbranche. Visuelle Überwachung der Fahrzeugumgebung sowie des Innenraums bietet eine Reihe zusätzlicher Sicherheits- und Komfortfunktionen. Eine Innenraumkamera kann unter anderem eine bedingte Airbagauslösung sowie Müdigkeitskontrolle unterstützen oder zur Verbesserung der Infotainmentsysteme dienen. Die Personalisierung des Fahrzeugs ist ein weiterer Punkt, wo eine Kamera erfolgreich eingesetzt werden kann. Eine Memoryfunktion für die Speicherung von Sitz-, Lenkrad-, Spiegelpositionen, sowie Klimaeinstellungen kann nicht nur über die übliche tastenbasierte Auswahl des Profils oder den Smart Key realisiert werden, sondern auch durch eine biometrische Identifikation des Fahrers. Das Gesicht scheint an dieser Stelle die geeignetste biometrische Modalität zu sein. Dabei wird eine Identifikation und anschließende Anpassung ohne bewusste Interaktion seitens des Fahrers stattfinden, was zu einer Komfortsteigerung führt. Außerdem haben kontaktlose Authentifizierungsverfahren eine hohe Benutzerakzeptanz. Biometrische Modalitäten sind nicht übertragbar und sehr schwer verfälschbar. In Folge dessen steigt die Entführungsresistenz des Autos, indem einen Motorstart nur nach eine erfolgreiche Identifikation erlaubt wird. Diese Arbeit beschäftigt sich mit der Analyse der Gesichtserkennungsmethoden und ihre Anwendbarkeit im automotiven Kontext. Es werden ausgewählte Szenarien für Referenzdatensammlung und Fahreridentifizierung entwickelt sowie ein Konzept des KFZ-Innenraum-Gesichtserkennungssystems entworfen und implementiert. Im Rahmen der Forschung wird eine Menge praktischer Probleme analysiert wie z.B. Kamera- und Lichtquellenpositionierung, erforderliche Bildverarbeitung, Gesichtsdetektierung, Merkmalsextraktion, Klassifizierungsverfahren und quantitative Evaluierung der Ergebnisse. Insbesondere werden darstellungsbasierte und merkmalsbasierte Verfahren zur Gesichtserkennung verglichen. Darüber hinaus werden die Anforderungen an die Aufnahmetechnik formuliert. Die experimentelle Evaluierung wird mit zwei Benutzerdatenbanken durchgeführt. Für den allgemeinen Erkennungsperformanztest stand eine weit verbreitete biometrische Datenbank von 295 Personen (XM2VTS) zur Verfügung. Zur realitätsnahen Untersuchung des Systems wird eine eigene Datenbank von 54 Fahrern in einem realen Fahrzeug erstellt.

Abstract

Equipping a modern vehicle with optical sensors is a new tendency in the automotive industry. The visual observation of car surroundings as well as a car passenger compartment offers a world of additional mechanisms to increase safety, security and comfort. In current cars, an in-car camera supports smart airbag deployment, drowsiness detection and contributes to an advanced level of interaction with infotainment systems. The car personalization is the next step towards a superior car-driver interaction implying the ability of a car to recognize its driver. Such personalization also guarantees a higher level of anti-theft protection granting the engine start exclusively to registered drivers. Nowadays, the personalization is mainly intended for comfort enhancement and associated with a memory function. The memory function activation for the automatic adjustment of the seat, steering wheel, mirrors and air-conditioning can be reached not only using memory buttons or a personal smart key, but also by applying biometric driver identification. Biometric modalities have the advantage that they are non-transferable and very hard to forge. In this context, the face seems to be the most appropriate biometric modality because of the usage of a contact-less acquisition device, high level of user acceptance and the fact that identification arises unconsciously for a driver, making a car more comfortable. This work is devoted to the feasibility analysis of the automatic face recognition techniques in the automotive context. The concept of an in-car face recognition system is developed and implemented in a real car. Several scenarios for collecting biometric data and for driver identification are proposed. A number of practical problems have been addressed. These are positioning of a camera and light sources, image pre-processing, face detection, feature extraction, classification and experimental evaluation of the system performance. In particular, appearance-based and feature-based approaches for face recognition are compared. On top of that, the requirements to the acquisition and identification subsystems are conceived. The empiric study on the recognition performance comprises experiments on two databases. For the evaluation of the general performance of the developed face recognition system (FaceART) and comparison to state-of-the-art systems, the XM2VTS database of 295 persons is utilized. The realistic recognition performance of the proposed system has been evaluated based on the database of 54 car drivers, collected by the author in the real car.

Contents

List of Figures

List of Tables

Abbreviations

EER	Equal Error Rate
ER_d	Driver Error Rate
ER_p	Passenger Error Rate
FDR	False Detection Rate
FMR	False Match Rate
$FNMR$	False Non-Match Rate
$FNOR$	False Non-Occupancy Rate
FNR	False Negative Rate
FOR	False Occupancy Rate
FPR	False Positive Rate
$HTER$	Half-Total Error Rate
k-NN	k-Nearest-Neighbor (algorithm)
$RSAT$	Rotated Summarized Area Table
SAT	Summarized Area Table
TAR	True Accept Rate
ABS	Antilock Braking System
ACC	Adaptive Cruise Control
AdaBoost	Adaptive Boosting
ADAS	Advanced Driver Assistance Systems
AFR	Automatic Face Recognition
AMSL	Advanced Multimedia and Security Lab at the Otto-von-Guericke-University of Magdeburg
AMSLator	Car simulator built in AMSL
ARTMAP	Adaptive Resonance Theory Mapping (neural network)
AUC	Area Under the Curve
AVECLOS	Average Eye Closure (algorithm)
Bagging	Bootstrap Aggregating
BFP	Best-Fit Plane

CAN	Controller Area Network
CART	Classification And Regression Trees
CAS	Collision Avoidance System
CCD	Charge-Coupled Device
CCTV	Closed-Circuit Television
CDF	Cumulative Distribution Function
CIAIR	Center for Integrated Acoustic Information Research at Nagoya University
CMC	Cumulative Matching Characteristic (curve)
CMOS	Complementary Metal-Oxide-Semiconductor
CNN	Classical Neural Networks
COOP	Critical Out-Of-Position
DAS	Driver Assistance Systems
DCT	Discrete Cosine Transform
DET	Detection Error Trade-off (curve)
DSP	Digital Signal Processing
ECU	Electronic Control Unit
EEG	Electroencephalogram
EFS	Electric Field Sensing
EM	Expectation-Maximization
EMG	Electromyogram
ESP	Electronic Stability Program
FAM	Fuzzy ARTMAP
FDR	Facial Driver Recognition
FFCS	Forward-Facing Child Seat
FIR	Far-Infrared
FNN	Fast Neural Networks
FPGA	Field-Programmable Gate Array
FPS	Frames Per Second
FRVT	Face Vendor Recognition Test
GLCM	Gray-Level Co-occurrence Matrix
GMM	Gaussian Mixture Models

HDR	High-Dynamic Range
HMI	Human-Machine Interface
HOG	Histogram of Oriented Gradients
ICA	Independent Component Analysis
ICP	Integrated Center Panel
IR	Infrared
ITS	Intelligent Transportation Systems
IVIS	In-Vehicle Information System
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LDW	Lane Departure Warning
LIDAR	Light Detection And Ranging
LKS	Lane-Keep Support
LoG	Laplacian of Gaussian
LPM	Linear Prediction Model
LTM	Long Time Memory
MB	Memory Buttons
MDL	Minimum Description Length (principle)
MFCC	Mel-Frequency Cepstral Coefficients
NCCC	Normalized Cross-Correlation Coefficients
NHTSA	National Highway Traffic Safety Administration
NIST	National Institute of Standards and Technology
NV	Automotive Night Vision
OOP	Out-Of-Position
PAC	Probably Approximately Correct (learning)
PCA	Principal Component Analysis
PDF	Probability Density Function
PDV	Permanent Driver Verification
PERCLOS	Percentage of Eye Closure (algorithm)
PIN	Personal Identification Number
PMD	Photonic Mixer Device

PR	Pedestrian Recognition
RADAR	Radio Detection And Ranging
RFID	Radio-Frequency Identification
RFIS	Rear-Facing Infant Seat
ROC	Receiver Operating Characteristic (curve)
ROI	Region Of Interest
RPM	Rank Probability Mass (function)
SDI	Single Driver Identification
SDK	Software Development Kit
SFAM	Simplified Fuzzy ARTMAP
SMQT	Successive Mean Quantization Transform
SNoW	Sparse Network of Winnows
SOD	Seat Occupancy Detection
SVM	Support Vector Machine
ToF	Time-of-Flight (camera)
TSR	Traffic Sign Recognition
UDiS	User Discrimination System
VC	Vapnik-Chervonenkis (theory)
VisMes	Research project ‘Visuelle Vermessung im Automobil’

Preface

Let us imagine a vehicle in a Brave New World[1]. The personal car is not a means of transportation anymore, but a living cybernetic organism. It is capable of recognizing you, your fatigue and your emotions. It adjusts driver assistance systems according to your needs automatically and completely transparently. It helps you to drive in critical situations. In case of an accident it adapts the airbag deployment because it is informed about your body dimensions and the current body location. Moreover, it calls the next ambulance because your global position is also known. Isn't it impressive?! Is it science fiction? Do these technologies belong to the future or rather to reality? This work is going to answer some of these questions through analyzing the progress in the domain of visual monitoring systems applied for driving assistance and comfort enhancement.

[1] Allusion to the novel "Brave New World" written by Aldous Huxley

Chapter 1

1 Introduction

This chapter provides an overview of camera-based observation systems in an automotive environment and gives a motivation to develop and integrate such systems. The research challenges regarding the visual observation of a car passenger compartment are designated and research objectives are introduced. The focus is on design, implementation and evaluation of three applications: seat occupancy detection for supporting the conditional airbag deployment, distinguishing of driver and front-seat passenger hands during the interaction with a dual-view touch-screen, and facial driver identification aiming at car personalization.

1.1 Problem description

This section is comprised of the following subsections:

- Motivation (camera in automotive applications)
- Formal view to safety and comfort of a vehicle
- Looking inside and outside of a car
- Car personalization

1.1.1 Motivation (camera in automotive applications)

Automotive industry permanently endeavors to make a vehicle more attractive for a potential customer. Due to growing competition, manufacturers look for possibilities to offer better equipped cars at reasonable costs. First and foremost, new equipment includes abundant driving assistance systems such as antilock braking system (ABS) or electronic stability program (ESP), which makes driving significantly safer. In addition, new vehicles possess in-vehicle information systems (IVIS) and superior human-machine interfaces (HMI) making them more user-friendly and comfortable. This new stream of development opens up fundamentally new challenges for automotive researchers. The number of electronic control units (ECU), the amount of transferred information and the complexity of sensor networks are rapidly growing, bringing to the front the need for information management. Hence, computer science becomes more and more important in modern automotive research.

The basic need for safety relevant driving assistance systems is explained by the inability of an average driver to promptly react in critical situations. The driver's visibility is limited due to blind spots and restricted perception in the dark. The average driver reaction time is not sufficient for avoiding all accidents. Moreover, drivers are often distracted because of paying too much attention to the entertainment facilities or the status of monitors in the vehicle [139]. Thanks to modern driving assistance systems, the number of fatalities in road accidents in Germany is continually decreasing every year [10].

The higher level of safety and comfort is often achieved through the integration of additional sensors for collecting the required environmental information. The examples can be found in a rain sensor for the automatic triggering of wind-screen wipers, in a laser proximity sensor warning of approaching an obstacle, or in a pressure sensitive mat integrated into a seat for sitting posture analysis and subsequent adjustment of airbag deployment. On the one hand, getting rid of those sensors reduces the comfort and safety making the car less attractive for customers. On the other hand, the endless addition of new sensors notably increases the complexity and the end price of the car. One solution for this problem is the integration of inexpensive and versatile sensors which might substitute several other sensors and offer the data input for new applications.

Due to the rapid development of the optical technologies, a camera seems to be the best choice for such an inexpensive and versatile sensor. So far, vision-based systems have been rarely adopted in cars due to the high costs, lack of robustness and considerable dimensions of the acquisition devices. Recently, a high-quality complementary metal-oxide-semiconductor (CMOS) sensor or a charge-coupled device (CCD) sensor, which is the main element of a digital camera, has become considerably cheaper as well as digital signal processing (DSP) boards. Furthermore, the recently introduced field-programmable gate array (FPGA) technology allows implementation of a complete image processing system on one low-cost microchip. Using this technology along with the optical sensors allows manufacturers to produce relatively small smart cameras with embedded image processing software at reasonable costs.

However, it is not verified yet that cameras are able to completely substitute the standard sensors for the most prominent outside car and inside car applications. Therefore, the feasibility study regarding whether the visual analysis is a promising alternative of a standard solution for a particular automotive application is the central question of this work. The next question is, if a camera can make it possible to implement fundamentally new automotive applications never implemented before using alternative sensors.

Since an output of a camera is an image and an image is an array of pixel intensities containing no semantic information about represented objects, image processing algorithms play the most important role for the interpretation and understanding of the visual information. The algorithms are expected to fill the semantic gap between an object in the real world and its representation on the image. From this perspective, the main challenge for developers of visual recognition systems consists in robust detection and extraction of objects from images using conventional pattern classification approaches which couple image processing operations with statistical machine learning methods.

1.1.2 Formal view to safety and comfort of a vehicle

The term safety is commonly defined as protection of human beings from recognized dangers to achieve an acceptable level of risk. The term automobile safety refers to the practices to avoid automobile accidents and protect passengers and pedestrians in case an accident has happened. Regarding vehicle safety, these practices include design and construction of a vehicle as well as its safety-related equipment. Traffic safety, in contrast, refers to safe automobile infrastructure and addresses the design of roadways and legal issues. As shown in Figure 1.1, automobile safety includes “active” and “passive” safety components. Active safety is the term combining the mechanisms for crash prevention. These mechanisms are often referred to as Driver Assistance Systems (DAS). “Passive safety” is the term describing the components of the vehicle for protecting occupants during a crash. While active safety addresses three main risk factors: driver, vehicle and infrastructure, passive safety exclusively addresses the vehicle [65].

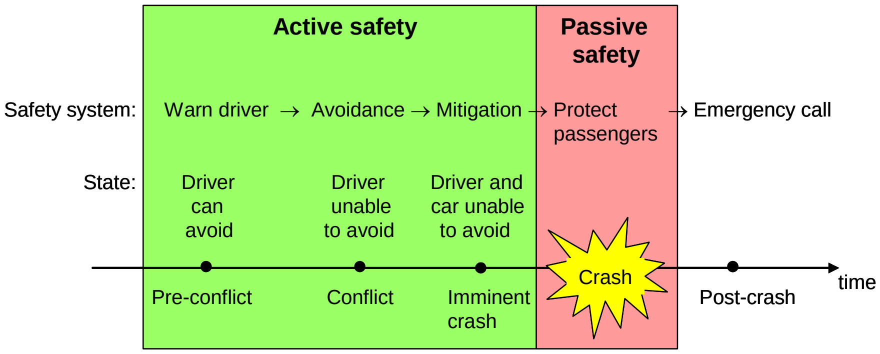


Figure 1.1: Active vs. passive safety

The term comfort is not formally defined in an automotive context. However, the intuitive conception about comfort is quite comprehensible. Basically, a high level of comfort is distinguished by the high number of automatically adjustable components such as, climatic system, seats, mirrors, steering wheel, windshield wipers, headlights etc. Langnickel et al. [143] have extended the conception of comfort by including the transparency of use, meaning that the less the driver's activity is needed for the component activation, the more comfortable the component is.

1.1.3 Looking inside and outside of a car

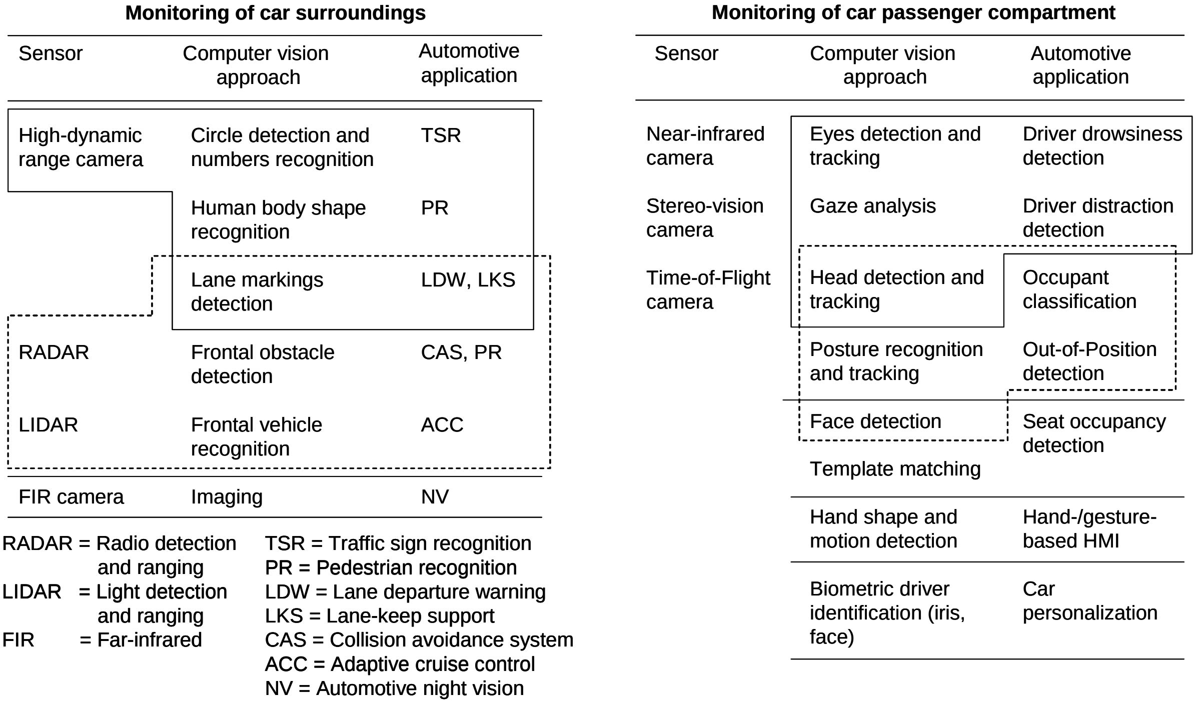
Cameras are versatile devices making them useful for both outside car and inside car observations. They are theoretically able to support not only comfort oriented systems but also "active" and "passive" safety components of an automobile. Examples of automotive systems which may make use of an optical sensor are introduced in Table 1.1.

Table 1.1: Examples of automotive applications making use of visual recognition

	Outside looking camera	Inside looking camera
Active safety	- Lane-keep support (LKS) - Lane departure warning (LDW) - Collision avoidance (CAS) - Pedestrian recognition (PR) - Automotive night vision (NV) - Traffic sign recognition (TSR)	- Driver drowsiness detection - Driver distraction detection
Passive safety		- Seat occupancy detection - Occupant classification - Out-of-position detection
Comfort	- Adaptive cruise control (ACC)	- Hand-/gesture-based HMI

The observation of vehicle's surroundings is reasonable only before the crash. This is why all applications potentially making use of an outside looking camera do not address passive safety and can be combined under the term DAS.

According to research done by the European Commission, the overall number of fatalities in the European Union by 2020 can be reduced by half thanks to the Advanced Driver Assistance Systems (ADAS). This is why the European Parliament prescribed in 2012 the mandatory fitting of Advanced Emergency Braking Systems in commercial vehicles from 1 November 2013 for new vehicle types and from 1 November 2015 for all new vehicles in the European Union [66]. This requirement basically includes the implementation of Lane Departure Warning, Anti-Collision Warning and Pedestrian Recognition systems.



Monitoring of car surroundings

Sensor	Computer vision approach	Automotive application
High-dynamic range camera	Circle detection and numbers recognition	TSR
	Human body shape recognition	PR
	Lane markings detection	LDW, LKS
RADAR	Frontal obstacle detection	CAS, PR
LIDAR	Frontal vehicle recognition	ACC
FIR camera	Imaging	NV

RADAR = Radio detection and ranging
LIDAR = Light detection and ranging
FIR = Far-infrared

TSR = Traffic sign recognition
PR = Pedestrian recognition
LDW = Lane departure warning
LKS = Lane-keep support
CAS = Collision avoidance system
ACC = Adaptive cruise control
NV = Automotive night vision

Monitoring of car passenger compartment

Sensor	Computer vision approach	Automotive application
Near-infrared camera	Eyes detection and tracking	Driver drowsiness detection
Stereo-vision camera	Gaze analysis	Driver distraction detection
Time-of-Flight camera	Head detection and tracking	Occupant classification
	Posture recognition and tracking	Out-of-Position detection
	Face detection	Seat occupancy detection
	Template matching	
	Hand shape and motion detection	Hand-/gesture-based HMI
	Biometric driver identification (iris, face)	Car personalization

Figure 1.2: Sensors and computer vision approaches applied in automobiles for the monitoring of surroundings and passenger compartment

Figure 1.2 introduces automotive systems along with the required computer vision approaches. The left columns contain commonly applied sensor technologies. These are radio detection and ranging (RADAR), light detection and ranging (LIDAR) often referred to as laser sensor, far-infrared (FIR) camera also called thermal camera and high-dynamic range (HDR) camera for the observation of car surroundings. The monitoring of a car passenger compartment is usually done with conventional CCD or CMOS camera in the shorter end of the near-infrared spectrum from 750 to 1000 nm. For the range imaging, which has recently become a trend technology, a standard monocular camera is replaced by a stereo camera or by a Time-of-Flight (ToF) camera. Let us take a look at what these automotive systems do and which sensors are currently applied in particular:

- Lane-Keep Support (LKS) system is designed to reduce the number of road accidents caused by driver distractions or drowsiness. The system includes the Lane Departure Warning (LDW) system and active steering component. The LDW system warns a driver when the vehicle begins to move out of its traffic lane on properly marked roads (e.g. highways or arterial roads). The active steering system automatically leads the vehicle back to the traffic lane. The LDW systems are currently developed by almost all prominent car manufacturers and based on all four kinds of sensors: laser sensors [184], radar sensors, infrared sensors and optical video sensors [75].
- Collision Avoidance System (CAS) warns a driver of any dangers located ahead on the road. The system provides autonomous braking to avoid a forward collision in case the driver

does not respond to the warning signal in time. The CAS systems are geared to vehicles in surroundings and especially helpful in restricted sight conditions caused by fog, for example. CAS systems use proximity measuring devices such as laser or radar sensors.

- Pedestrian Recognition (PR) system is an extension of CAS oriented to pedestrians. For instance, the system developed by Volvo [198] tracks the road in front of the vehicle, detects objects and determines distances to them. The system relies on two sensors. The first one is the radar integrated into the front grille for the detection of objects and the determination of distances. The high-resolution camera is installed behind the rear-view mirror for the recognition of object types detected by the radar. Based on the camera images, the system is capable of predicting the pedestrian movement trajectories and, thus, estimating the risk of an accident. Similar to CAS, in case of the high risk of an accident the driver is warned and the automatic braking is applied if necessary. Currently, the system operates at speeds below 30 km/h which is the upper limit of the robust pedestrian detection.

The next three systems, namely Adaptive Cruise Control, Automotive Night Vision and Traffic Sign Recognition belong to Advanced Driver Assistance Systems and are currently offered as optional equipment on certain premium vehicles.

- Adaptive Cruise Control (ACC) is designed to automatically adjust the vehicle's speed to keep a safe distance from the vehicle ahead. The vehicle slows up when approaching the anterior vehicle and accelerates when the distance to that vehicle increases. ACC systems use either radar [221] or a laser sensor for the distance determination.
- Automotive Night Vision (NV) is a technology to improve a driver's visual perception in darkness or poor weather conditions. The road in front of the vehicle is observed and pictures are transferred to the driver display. Active and passive systems are in use. Active systems make use of a conventional CCD camera and active near-infrared illumination. In passive systems a thermal camera is used for capturing thermal radiation emitted by the objects. Thanks to that, the seeing distance spread out beyond the reach of the vehicle's headlights and the sight clarity increases.
- Traffic Sign Recognition (TSR) is the most immature technology developed for the detection and recognition of traffic signs on the road. The first generation of TSR systems is capable of recognizing European round speed restriction signs [60]. The second generation of TSR systems is able to recognize overtaking restrictions as well. The currently introduced TSR systems make use of a high-dynamic range CMOS camera integrated into the rear-view mirror.

The vehicle's surroundings are mainly observed with radar or laser sensors measuring the proximity to the ambient objects. Here the robustness of measurements plays the most important role. For the majority of applications there is no need for high-resolution sensors because the distance to an obstacle is the only required information, but this distance has to be very precisely estimated. Exceptions can be found in PR and TSR systems, where the high-resolution camera is unexpendable for the object type recognition. Due to the natural limitations of cameras such as an insufficient dynamic range [17] leading to overexposure and discrepancies between the real world and the sensor information, cameras can not replace proximity sensors for the far-distance obstacles detection. In contrast to outside car imaging, the inside car imaging does not require such high-dynamic range images and the background is relatively constant, making the camera very promising observation device.

The observation of a vehicle's passenger compartment supports both active and passive safety. Driver drowsiness detection is an example of the system to **avoid accidents**; the seat occupancy detection is an example of the system supporting smart airbag and, thus, protecting passengers from injuries **during the crash**. Moreover, the observation of a car passenger compartment can be helpful in reducing crash effects. The intelligent car can make an emergency call disclosing the number of passengers and their current conditions.

- Driver drowsiness detection systems are designed to monitor driver's eyes and evaluate blinking behavior. Eyelid closure time and speed as well as blinking frequency are considered to be parameters determining the degree of drowsiness [101, 99]. In 2006, Toyota was the first company to introduce this safety system as a part of so-called driver monitoring system [232]. The system utilizes active near-infrared illumination and a conventional camera with CCD sensor to monitor driver attentiveness. The camera is placed on the top of the steering column. The system is capable of eye tracking due to specific specular reflection of a pupil under infrared illumination. If the driver is not paying attention to the road ahead and a dangerous situation is detected, the system will warn the driver by flashing lights and warning sounds. If no action is taken, the vehicle will apply the brakes (a warning alarm will sound followed by a brief automatic application of the braking system).
- Seat occupancy detection systems are designed to prevent the deployment of airbags at unoccupied seats. This helps to avoid the considerable cost imposed by the replacement of airbags after an accident. Moreover, airbag deployment must be prevented at seats occupied by child restraint systems or small children (under 66 lbs) because of serious injury risks. Traditional seat occupancy detection systems include pressure sensitive (weight) sensors and transponders in child restraint seats. Optical sensors could be a perfect alternative to the current detection methods especially for rear seats where adaptive restraint systems have not reached a high saturation rate yet. The crucial advantage of an optical sensor is that one camera can be used to determine the occupancy of all seats. Furthermore, a camera-based setup allows upgrading a seat-occupancy detection system to an occupant classification system without hardware changes. The only required adaptation is the implementation of more sophisticated image processing and pattern classification algorithms.
  In fact, a seat occupancy detection system can be considered as a degraded case or rather as a part of an occupant classification system where the system chooses between two alternatives: "occupied seat" and "empty seat". A complete occupant classification system is able to identify if a seat is occupied by an object, by a child restraint system or by a small, average or large person [68]. This allows selecting the airbag deployment stage regarding its firmness and velocity. The further improvement of this system is the out-of-position (OOP) classification. The system identifies situations in which the airbag can seriously injure or even kill a passenger because of his/her unorthodox body position and prevent the deployment.

Apart from safety related applications, an inside car camera is proposed to be used for the improvement of infotainment systems providing superior human-machine interface (HMI). The range of applications which can potentially make use of a camera is very wide. Just to mention a few, these are the driver emotions recognition [109], gestures based communication [3], gaze tracking for the driver attention control [74]. However, the focus is on only one HMI system distinguishing driver's and front-seat passenger's hands, further called User Discrimination System (UDiS).

- Distinguishing between driver and front-seat passenger during their interactions with dual-view touch screen integrated to the automobile center console is one example, where visual recognition seems to be better than other technologies. Dual-view devices are designed to allow users to see different contents on the same display depending on the viewing angle. For example, the driver can follow the directions given by the navigation system while the passenger watches a film or scrolls through the play list in a music player. With regard to

touch-screen based interaction, it must be precluded that the passenger's interactions (e.g. with the music player) affect the driver's display content and vice versa. The dual-view displays are currently available on the market e.g. in Range Rover [123].
Two user discrimination concepts currently exist. The first one is electric field sensing (EFS) [194]. This approach is based on capacitive coupling of the user's body with the touch panel and can easily discriminate users based on the individual frequency of the transmitter electrode integrated into the seat. However, capacitive technology still has some serious limitations regarding the passenger clothing and the lack of user acceptance. Occupants fear potential health hazards often attributed to electromagnetic fields. Moreover, automotive industry has been making serious efforts to reduce electromagnetic fields in a vehicle and, therefore, additional electric fields are hardly acceptable. The second technology is based on the optical sensing of the passengers' activities and in contrast to the first technology is widely accepted by users [108]. The camera that is mounted in the roof observes the center console and registers motion in the driver and passenger regions. Analyzing this information along with shapes of the moving hands, the decision is made whether a driver or passenger is interacting with the touch-screen.

The attention of established industrial concerns is currently focused on the observation of car surroundings for the improvement of active safety components. The currently applied proximity sensors (millimeter-wave radars and lidars) have already fulfilled the criteria of the feasibility study, so cameras are not considered promising alternative here. For this reason, outside car camera applications are beyond the scope of this work. The focus is on the visual recognition systems inside a car. Here, I see a great potential for theoretical research because of general appropriateness and versatility of visual sensors and few examples of available industrial systems. Figure 1.3 schematically shows the range of observation systems looking inside and outside of a vehicle. The marked blocks are addressed in detail. Computer vision algorithms are analyzed and prototypic systems are implemented and evaluated.

Driver drowsiness detection is also beyond the scope of this work because the camera-based driver monitoring system has already been developed by Toyota in 2006 and successfully integrated in premium models of Lexus and Toyota. Currently, drowsiness detection is offered by several vehicle suppliers, but those systems are either based on an analysis of driver behavior signals [45, 137] or are derivates of the LDW system [75, 244]. The major focus of this work is the facial driver identification system which can be categorized as an essential part of car personalization making use of an optical sensor.

1.1.4 Car personalization

A higher level of comfort, safety and security may be achieved through the personalization of a car which can be understood as an individual adjustment of automotive systems to a driver. Comfort enhancement is associated with a memory function. After the automatic selection of a stored passenger profile, the seat, steering wheel, rear mirrors, air conditioning and car radio are adjusted to a passenger. The safety implication can be found in the driver-specific adaptation of driver assistance systems. One possible application is speed restriction depending on the driver's age or experience. Security is improved through the anti-theft protection system assuming that access to a car (or car ignition) is granted only to a registered driver.

The key issue for implementing personalization is the ability of a car to recognize its driver. Currently existing systems imply that drivers explicitly report their identities. This is done either by means of memory buttons (MB) or through a radio-frequency identification (RFID) tag integrated into an individual smart-key. Both of these means of identification have a serious disadvantage

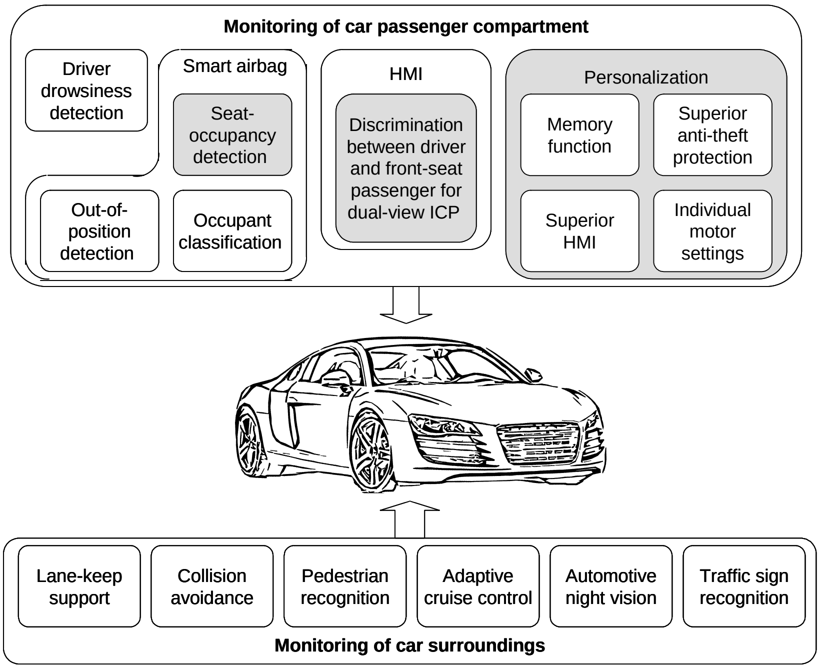

Monitoring of car passenger compartment
Driver drowsiness detection
Smart airbag
Seat-occupancy detection
Out-of-position detection
Occupant classification
HMI
Discrimination between driver and front-seat passenger for dual-view ICP
Personalization
Memory function
Superior anti-theft protection
Superior HMI
Individual motor settings
Lane-keep support
Collision avoidance
Pedestrian recognition
Adaptive cruise control
Automotive night vision
Traffic sign recognition
Monitoring of car surroundings

Figure 1.3: Visual monitioring systems looking inside and outside of a vehicle

because they do not provide the direct link between an enrolled identity and a person sitting in front of the steering wheel. The smart-key can be easily transferred to another person making all individual adjustments meaningless. Using memory buttons does not improve the situation. A button with the wrong profile can be selected activating the inappropriate comfort or driver assistance settings. In other words, key-based or more generally token-based systems are linked to the vehicle and not to the driver. The only way to provide the link between the potential driver and the registered profile is biometric identification.

Despite evident advantages, biometric authentication implicates risks evoked by possible misclassifications and the need to preserve sensitive biometric data. Therefore, the demands on an in-car biometric system have to be formally revealed from the user's point of view as well as from the manufacturer's point of view. For a passenger, user acceptance and usability aspects are of the highest importance, just as for a manufacturer the costs and the technical implementation aspects are essential.

The integration of biometric systems into a car has been studied since the late 1990s. However, only one biometric modality has been practically addressed so far, namely fingerprint. For instance, a fingerprint for immobilizer deactivation and driver's profile activation was implemented by Daimler AG in luxury series cars about ten years ago. Volkswagen Group has taken up this initiative through the "one-touch memory" option in Audi A8 [6]. Currently, several custom fingerprint sensors are widely presented on the market, so that every car can be retrofit with such device. Despite technical advantages and brilliant performance of fingerprint devices, fingerprint-based driver identification has not achieved a high saturation level in cars. This is why the automotive industry takes a great interest in face and iris recognition. Volkswagen AG is currently working on facial driver identification, but the developed system is declared not mature enough for series production [243]. Sarnoff Corporation announced an adaptation of its iris recognition system for use inside of motor vehicles [19]. Nonetheless, to the best of my knowledge, currently there are no patented and implemented face-based or iris-based driver authentication systems in consumer market cars still making the topic attractive for academic investigations.

Regarding the complexity of the acquisition system, the face modality is preferable compared to iris. Iris acquisition requires active illumination and a high-quality camera (incl. high-quality image sensor and objective) providing a high-resolution image. In contrast, face acquisition can do with a simple, small, inexpensive and low-power-consuming camera with a wide view angle objective. Remaining biometric modalities which can be captured with an optical sensor are far less appropriate for the utilization in a car passenger compartment than aforementioned ones. This is why the focus is on face modality and its applications to biometric driver identification.

1.2 Research challenges

Let us summarize the challenges of automotive researchers regarding the inside car observations.

The first and foremost challenge is to indicate:

- the existing applications, which can benefit from the inside car camera, but currently use alternative sensors;
- the new applications, which can be realized using the inside car camera.

After outlining potential applications, the next research challenge is to answer questions which are derived from practice:

- To what extent can a camera improve those applications in terms of reducing costs, increasing safety/comfort?
- Is the utilization of a camera feasible?
- What additional risks emerge together with the integration of new systems?

These questions primarily address hardware components of the vision-based system. The positive answers let us qualify the system as promising.

For the promising applications software development becomes the main focus. From this perspective, the challenge lies in answering technical questions:

- How robust can the implementation be?
- What is the operational scenario?
- Which image processing and machine learning algorithms are appropriate?

In fact, adopting computer vision to the automotive field is a very challenging task because of natural limitations of vision-based systems. On the one hand, the theory of digital image processing is mature including algorithms for detection, segmentation and recognition of objects; on the other hand, the object representation on an image is ambiguous often resulting in inaccuracies and recognition mistakes. For example, a single camera is not capable of robust distance measuring. Another typical problem is the low-dynamic range of a sensor leading to overexposure in images while driving along a tree-lined avenue or coming out from under a bridge on a sunny day.

Using non-visual sensors, a signal processing algorithm is usually trivial. The applications entirely rely on the incoming sensor signal revealing the physical characteristics of a measured object. In case of a vision-based system there is a semantic gap between the object in the real world and its representation on a camera image. This is why non-trivial image processing is required. From this point of view, the aforementioned technical questions can be joined to the following question:

- Is it possible to robustly fill the semantic gap between the object in the real world and its representation on a camera image by an intelligent image processing algorithm for the particular application under certain scenario-related restrictions?

The rule of thumb states, the simpler a signal processing algorithm is the more robust the system is. Complex algorithms lead to slow and non-robust systems. Therefore, while adopting computer vision algorithms for automotive applications, researchers look for the trade-off between the complexity of the algorithm and recognition performance.

1.3 Research objectives

Based on the attainments from the joint research project VisMes (Visuelle Vermessung im Automobil) with the industrial partner and numerous discussions with automotive experts, three promising applications of an inside car camera have been recognized:

1. Seat occupancy detection using an omni-directional camera for supporting conditional airbag deployment;
2. Distinguishing driver hand from front-seat passenger hand during interaction with a dual-view touch-screen using a ceiling camera facing the center console;
3. Facial driver identification using a dashboard camera monitoring the driver's face through the steering wheel aperture for car personalization.

While selecting the applications, the endeavor has been made to properly answer the questions from the previous subsection and to pick out only those which greatly benefit from visual sensing, but are still in the "proof of the concept" development stage implying that they are not presented in series cars yet. For instance, driver drowsiness detection is the most important safety-related application because drowsy driving is the prime cause of fatalities and serious injuries in traffic accidents [67]. Nonetheless, visual drowsiness detection has already been thoroughly investigated by several research groups [122, 101, 16, 99] so that theoretical as well as practical solutions exist. Moreover, since 2006 the driver monitoring system for visual drowsiness detection has been offered by Toyota in several models of Lexus [232].

Apart from driver drowsiness detection, this thesis does not cover the following applications, which, however, significantly benefit from an in-car camera and have been fairly well addressed in the works of other researchers:

- Driver distractions detection [227, 74]
- Driver emotions recognition [109]
- Occupant classification [136, 68]
- Out-of-position detection [233, 135]

The selection of the application (1) is motivated by the fact that the 360 degree camera is proposed for simultaneous observation of all seats in a car. A similar hardware setup has been used only by Wender et al. [249]. However, the proposed algorithm relies exclusively on face detection and the experimental evaluation engages a very limited collection of 600 images (300 with occupants and 300 empty). Despite the solid correct detection rate of 84.45% while having 0.075 false detections per frame, as reported in this paper, these results cannot be extrapolated to a real car because the test frames have been manually selected. For instance, in the framework of the VisMes project, the in-car videos have been collected in "get in to the car" and "get out of the car" scenarios [160]. Passenger faces are covered in more than half of video frames. Hence, face detection alone is insufficient to successfully verify seat occupancy [158]. Advanced template matching is proposed for the detection of occupied seats supported by face detection to verify that seats are occupied by passengers.

The application (2) implies the usage dual-view touch screen, which is, to the best of my knowledge, not implemented in cars yet. Due to this fact, the application proposed in this work is the first of its kind and has no alternative vision-based solutions. There are three clues to distinguish driver and front-seat passenger hands: the amount of motion in a particular region, the shape of the moving object and the beginning and final positions of the moving hand.

The idea of facial driver identification, which is the proposed application (3), has been discussed since the late 1990s, but the concrete implementation concepts are still not established. There are several academic studies [20, 142, 14, 229, 63, 116, 174, 34] addressing parts of the system, but never considering the problem completely including the choice of hardware and software components and the description of application and operational scenarios. In most studies the focus is on the improvement of face recognition algorithms and their adaptation for a car passenger compartment [14, 229, 63, 116, 174, 34]. Other studies propose only the theoretical argumentation of biometric driver identification giving no clue about practical implementation [20, 142]. In contrast, my objective is to develop a comprehensive facial driver identification system from scratch. This includes retrofitting the car with the acquisition device, active illumination modules and the signal processing hardware, development of application and operational scenarios, development of the face recognition algorithm.

Figure 1.4 schematically shows the image processing and pattern recognition approaches applied in targeted applications.

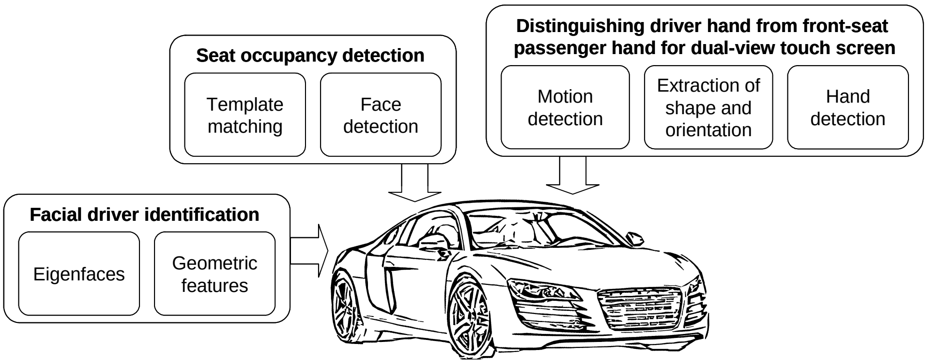

Seat occupancy detection
Template matching
Face detection
Distinguishing driver hand from front-seat passenger hand for dual-view touch screen
Motion detection
Extraction of shape and orientation
Hand detection
Facial driver identification
Eigenfaces
Geometric features

Figure 1.4: The targeted applications

Taking the selected applications into account, the research objectives are:

- to propose the system design including hardware components, application and operational scenarios, and image processing algorithms;
- to implement the proposed recognition systems;
- to evaluate the recognition performance of the implemented systems in a statistically significant way.

The hardware problem is common to all three applications. The research objective here is to identify the optimal type and position of a camera and active illumination sources.

For the applications (1) and (2) the description of the task implicates the application scenario. In contrast, the driver identification can be used for several purposes. In this work, three application scenarios improving comfort, safety and security of the vehicle are proposed:

- Biometric memory function for the improvement of comfort, meaning that after driver identification, the seat, steering wheel, rear mirrors, air conditioning and car radio are automatically adjusted to a driver;

- Mandatory driver assistance for the improvement of safety, meaning that after the driver is identified, the engine power is restricted and driver assistance systems are compulsively switched on for young and inexperienced drivers;
- Biometric anti-theft protection for the improvement of security, meaning that the car ignition is blocked until a registered driver is identified.

While the operational scenarios for the applications (1) and (2) are trivial, meaning that the system has to make a decision for each camera frame, the operational scenarios for (3) include two separate procedures: the registration/update of biometric data and driver authentication. Moreover, the authentication can be done either only once directly after getting into the car (for comfort- and safety-related application scenarios) in identification mode (see Subsection 2.1.4) or permanently for each camera frame (for the security-related scenario) in verification mode (see Subsection 2.1.4). The study on how these operational scenarios influence identification accuracy is the next objective of this work. Figure 1.5 combines all discussed components of a facial driver identification system, which is the main focus of this thesis.

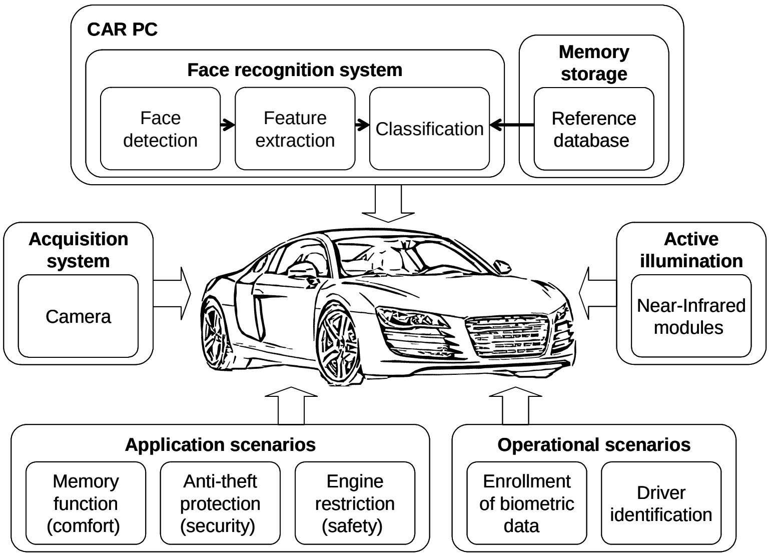


Figure 1.5: Main components of a facial driver recognition system

Regarding the face recognition system, there are additional objectives:

- Argumentation regarding why face modality is most appropriate for driver identification
- Comparison of approaches for face detection and image normalization
- Comparison of approaches for face modeling and classification

The experimental part of this research aims at determining if the reliable face recognition in a car is possible with the current state of computer vision technologies.

1.4 Summary of main contributions

Here, the achievements regarding three proposed camera-based monitoring applications in a car passenger compartment are summarized.

1.4.1 Seat occupancy detection (SOD)

The development of the seat occupancy detection system is based on the hardware setup of the VisMes project. The Audi A6 Avant has been retrofit with a DSP system for video capturing, five near-infrared lighting modules, and a panoramic camera (Sony RPU-C3522) including a 360-degree annular lens and a CCD sensor. Test images have been collected from 39 persons including males and females of different ages and various body sizes as well as different objects such as child restraint systems or pieces of baggage. The test persons were asked to get into the car in an arbitrary manner, fasten the seat belt and remain seated for at least three seconds, then change the seat or get out of the car. The total number of test frames amounts to 53928. Figure 1.6 shows the camera setup and an exemplary camera image.

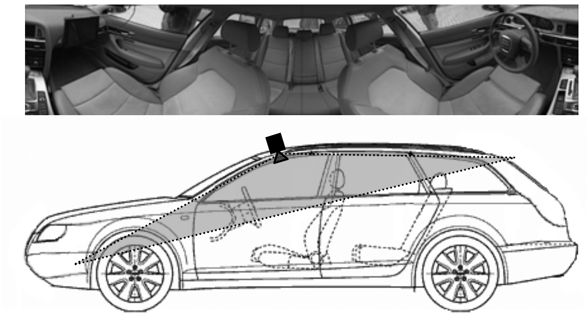


Figure 1.6: Omni-directional camera mounted on the ceiling next to the windshield for the monitoring of the whole car passenger compartment

There are two approaches proposed for seat occupancy detection:

- Template matching
- Face detection

Template matching Template matching is applied to the seat occupancy detection to solve the problem in an inverted way, namely not to detect occupied seats, but to detect empty seats. Templates of empty seats are cut and stored, and then consecutively searched in camera frames. I study the impact of image pre-processing approaches such as local normalization and edge detection, the impact of template selection strategy (small template, large template, several small templates) as well as the impact of multi-algorithm and temporal fusion of matching-scores to the recognition performance of template matching. Different combinations of pre-processing, template selection and fusion strategies are analyzed and compared. The experiments yield respectable detection performance expressed in terms of equal error rates (EER) which at best do not exceed 2.60% for front seats, 6.23% for rear side seats, and 11.07% for the rear center seat. The straightforward conclusion about which combination leads to the optimal performance cannot be made due to high variations in results. The way to define templates and their quantity remains an open question since the trivial multi-template approach with only two templates has generally performed best. Template selection is based on intuition. Such a selection does not ensure that the templates can be reliably found in other images of an unoccupied vehicle, especially if illumination conditions dramatically differ from the reference. A more methodological approach is hence required, e.g. using template candidates from several images of an empty vehicle with various illuminations. Or, in other words, template matching can be extended to statistical pattern classification where images of empty seats (templates) build the set of positive examples and images of occupied seats build the set of negative examples. However, the resulting accuracy of occupancy detection is not competitive when compared to weight measuring systems where false detections of persons practically never occur.

Face detection The most plausible way to check whether a seat is occupied by a passenger is the detection and localization of heads or, more precisely, faces. In order to prove or disprove this hypothesis, three commonly used and widely available face detection algorithms (the algorithm of Viola and Jones [242], the algorithm of Kienzle et al. [127] and the algorithm of Nilsson et al. [180]) have been applied for the images of a car passenger compartment. Experimental evaluation has shown that a face detection system cannot reliably detect faces of vehicle occupants. This is due to the fact that occupants' faces are often angled away from the camera while detection systems are primarily geared to almost frontal faces. Another hindering factor, especially for the rear seats, is the covering of faces by interior components such as head-restraints of front seats. The test indicates that without adaptation the considered algorithms can be helpful only on rear seats with true detection rates varying from 12.52% to 20.13% while corresponding false detection rates vary from 0.15% to 1.58% for the best algorithm. Comparing these results with the results of Wender et al. [249], who used a similar hardware setup, it can be asserted that for the ideal case when still frames with normally posed occupants are used, the solid face detection performance can be achieved. In the general case, a face detection system used alone is not sufficient for seat occupancy detection. However, face detection may complement standard weight-based occupancy detection or even the aforementioned visual approach based on template matching for distinguishing of persons and objects.

1.4.2 Distinguishing between driver and front-seat passenger hands during the interaction with the center console (UDiS)

The user discrimination system has been completely reproduced in the car simulator (later referred to as AMSLator) in a laboratory including a passenger compartment with a genuine dashboard, front seats, rear seats, touch screen mounted to the center console, monochrome CCD-camera (Imaging Source DMK 31BU03.H) and two near-infrared lamps (880 and 940 nm), so that different illumination conditions can be simulated. This acquisition system allows automatic and unambiguous registration of all touch screen interactions and their synchronization with a video stream. Figure 1.7 shows the camera setup along with two exemplary camera images.

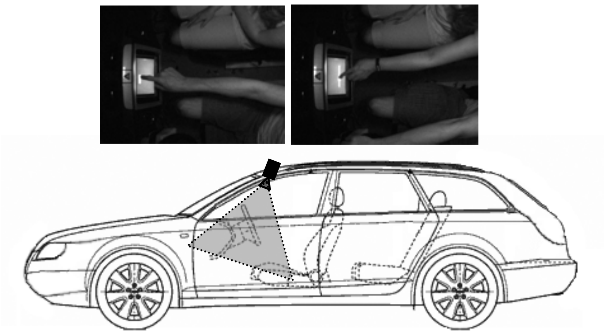


Figure 1.7: Monochrome CCD-camera mounted on the ceiling next to the windshield for the monitoring of driver and passenger hands

Approaches applied in UDiS:

- Motion detection in the driver and front-seat passenger regions
- Motion-based segmentation of the forearm and determination of its orientation
- Hand detection

Motion-based analysis Based on the hypothesis that the arms of driver and front-seat passenger are the only moving objects within the center console region, the segmentation of the arm is performed based on motion detection. Hence, the imaging system can segment the forearm while moving towards the touch screen based on smart image differencing. The differential image is derived from the current video frame subtracting the previous adjacent frame if they are different enough or subtracting the preceding frame with the different arm position. The shape of the arm (incl. forearm and hand) is extracted from the differential image by applying adaptive thresholding and morphological operations. The resulting binary shape allows the calculation of the amount of movement in driver and front-seat passenger regions as well as the detection of the principal direction of an arm which is pointing at the acting person.

Hand detection In addition to the motion-based forearm and hand segmentation, texture-based hand detection is engaged. Hand patterns of the driver and the front-seat passenger are learned applying the object detection algorithm of Viola and Jones [242]. Based on the frame sequence containing a localized hand, the hand moving trajectory is estimated and the acting person is recognized based on the collocations of the first and the last trajectory points.

The system based on solely hand detection is not reliable enough. The fusion of the motion-based and texture-based approaches leads to acceptable recognition accuracy. The error rates do not exceed 11% in case of uniform illumination and at worst yield 14.60% in case of uniform non-illumination. However, the resulting recognition errors do not allow the practical use of the proposed user discrimination system, but show that optical sensing is a very promising technology for this purpose.

1.4.3 Facial driver recognition (FDR)

For the purpose of face monitoring and facial driver identification an experimental automobile (Opel Vectra B) has been equipped with a low-cost CCTV camera mounted on the dashboard, an analog-digital video converter, and a car personal computer (CarPC) for the video capturing. Night acquisition is possible due to near-infrared light-emitting diodes (NIR-LED) integrated into the camera body. Figure 1.8 shows the camera setup together with an exemplary camera image.

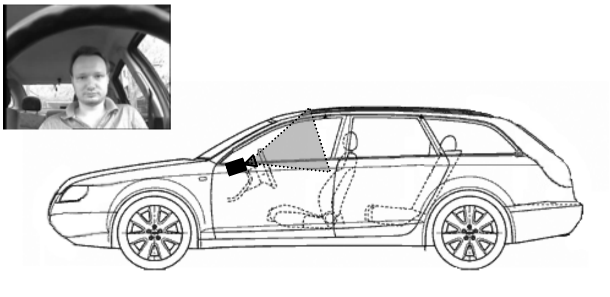


Figure 1.8: CCTV camera mounted on the dashboard and facing the driver through the steering wheel aperture for the monitoring of the driver's face

Components of the facial driver identification system:

- Application scenarios (see section 1.3)
  - Biometric memory function (comfort)
  - Mandatory driver assistance (safety)
  - Biometric anti-theft protection (security)
- Operational scenarios
  - Manual, semi-automatic and automatic enrollment (scenario with head rotations)
  - Single driver identification in an interactive mode to support the biometric memory function and mandatory driver assistance
  - Permanent driver identification to support biometric anti-theft protection
- Face recognition system (FaceART)
  - Pre-processing (face detection and light normalization)
  - Feature extraction (eigenfaces)
  - Classification (k-nearest neighbor algorithm vs. ARTMAP)

Operational scenarios Manual driver registration implies that a qualified salesman or an authorized car repair shop engineer instructs a new driver to take a required position for taking reference shots and manually controls their quality. Semi-automatic (interactive) registration implies that the car instructs a new driver to take an appropriate position for a shot and controls image quality. In the automatic mode, the system asks a new driver to provide several actions such as head rotations or head inclinations. Then, the system automatically learns all possible appearances of the face building a face model.

Single driver identification takes place directly after getting into a car and prior to engine ignition. The system asks the driver to look at the camera and captures a frontal face image for subsequent matching. In contrast, permanent driver identification implies the regular capturing and matching of a driver face either after a certain period of time or after each door closing.

Face recognition system As a part of my research, I have implemented a face recognition system called FaceART [161]. The pre-processing subsystem includes a face detection algorithm of Viola and Jones [242] as well as best-fit-plane subtraction and histogram equalization for light normalization. The eigenfaces approach [234] is utilized for feature extraction. The classification of feature vectors can be done using either the k-nearest neighbor algorithm (k-NN) or using an adaptive resonance theory mapping neural network (ARTMAP). The FaceART as a representative of appearance-based approaches is opposed to the commercial face recognition system Luxand FaceSDK [155] which can be considered a representative of feature-based approaches.

Regarding single driver identification, the identification accuracy of the Luxand FaceSDK is higher than 99% in the collected videos. In contrast, the identification accuracy of the FaceART surpasses only 90%. In case of permanent driver identification the accuracy drastically declines for both systems. The Luxand FaceSDK correctly identifies the driver in more than 68% of test frames and the FaceART in more than 63% of test frames. The experiments have shown that even non-optimal face recognition algorithms such as FaceART can achieve satisfactory identification accuracy making practical use of face recognition in a car very promising. However, in order to provide high identification accuracy, the operational scenario has to guarantee capturing of frontal faces.

1.4.4 Conclusion

Summarizing the results of three proposed applications, the different grades of success can be stated. Seat occupancy detection based on template matching achieves high detection rates only on front seats. The sufficient detection rates on rear seats can be achieved by combining template matching and face detection. Further efforts should be made towards occupant classification and out-of-position detection. The implemented system for the discrimination of driver and passenger hands has poor recognition performance and requires the adaptation of the utilized algorithms. However, the applied techniques have the potential for further adjustments. The facial driver identification system performs well when the operational scenario prescribes capturing of frontal faces. The recognition results in the unrestricted mode, meaning that arbitrary frames are used for driver identification, are insufficient for considering the system mature for practical use.

1.5 Thesis outline

The introduction is followed by the second chapter including fundamental definitions from both automotive research and computer vision. Related works considering inside car and outside car monitoring systems are addressed in detail. Special attention is given to the studies on vision-based biometric driver recognition.

The third chapter describes a theoretical concept of applying computer vision approaches for the monitoring of a car passenger compartment providing links between camera, image processing algorithms and automotive applications. The methodology comprises all basic components of pattern classification: pre-processing, feature extraction, feature selection, learning of a statistical model and matching. Several selected algorithms are also addressed in detail. These algorithms are later on used as construction elements in the introduced automotive applications (seat occupancy detection, discrimination of driver and passenger hands and facial driver identification) to implement template matching, face and hand localization and face recognition.

The fourth chapter addresses the design and implementation of the proposed automotive applications. This includes the choice of type and position of camera and active illumination sources as well as the development of software components under application constraints.

The fifth chapter is devoted to the experimental evaluation of the proposed algorithms. The strengths and weaknesses of the applied computer vision approaches are identified regarding the selected applications. Based on that, the limitations of the vision-based recognition techniques in application to the automotive environment are revealed.

In the sixth chapter, the results of experiments are discussed and critically analyzed in order to provide the generalization of the achievements in the framework of the proposed theoretical concept.

The seventh chapter concludes the thesis summarizing challenges in automotive research and my achievements and contributions. The chapter also provides an outlook of the ongoing automotive research emphasizing the topics of special interest. Here, I also try to forecast future research directions regarding visual in-car monitoring and car personalization.

The thesis is finalized by a list of references and an appendix to the experimental part including tables and diagrams not included in the chapter.

Chapter 2

2 Thesis fundamentals and summary of related works

In this chapter, the main definitions from domains of optical systems and image processing are introduced together with the fundamentals of statistical pattern recognition. Furthermore, an overview of current achievements in the field of automotive-related vision-based sensing as well as of established alternative sensing technologies is provided.

2.1 Image understanding

This section is comprised of the following subsections:

- Image sensing technologies (CMOS vs. CCD)
- Semantic gap between an object and its appearance on the image
- Pattern recognition
- Biometrics as a special case of statistical pattern classification

2.1.1 Imaging sensors

There are two established image sensing technologies: complementary metal-oxide-semiconductor (CMOS) and charge-coupled device (CCD). Both sensors consist of metal oxide semiconductors and rely on the photoelectric effect of silicon to convert incoming photons to electrons generating an electric charge. The charge is accumulated proportionally to the illumination intensity. The crucial difference is in how the sensor elements accumulate a charge and convert it to voltage (see Figure 2.1). In a CMOS sensor the conversion happens in each single pixel thanks to several layers of semiconductors. In contrast, pixels of a CCD sensor collect a charge and, after exposure is complete, transfer it to a common output structure, where the charge is converted to voltage.

The capabilities and limitations of sensors are predetermined by these different readout techniques. At this point, there is no proof of the superiority of one or another technology. In [151], Litwiller mention cases in which CMOS sensors naturally fit better. These are security cameras, videoconferencing, bar-code scanners, fax machines, consumer scanners, biometrics, and automotive in-vehicle monitoring. He also declares that CCD sensors are more suitable for high-end imaging applications. These are, for example, digital photography, broadcast television, industrial imaging, and some scientific and medical applications. This is why, contrary to some beliefs, both technologies will coexist in the future in a complimentary manner [121].

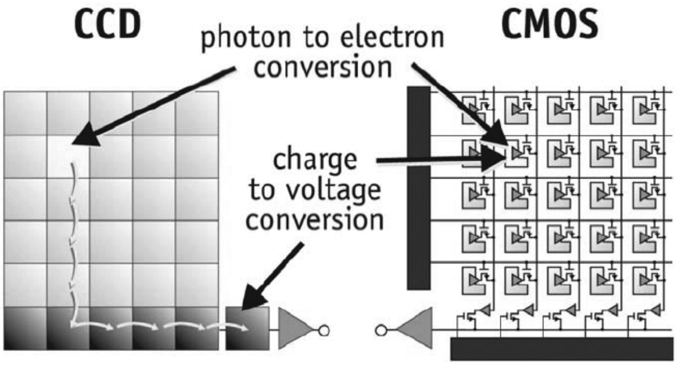


Figure 2.1: Principal difference of CCD and CMOS imaging sensors: CCD moves photogenerated charge from pixel to pixel and converts it to voltage at an output node. CMOS converts charge to voltage inside each pixel (modified from [152])

Let us take a look at the advantages and disadvantages of each sensor. Litwiller has proposed eight performance characteristics which can be used for the formal comparison: responsivity, dynamic range, uniformity, shuttering, speed, windowing, anti-blooming, biasing and clocking [151].

- Responsivity is often referred to as sensitivity and implies the amount of signal created from the optical energy in each pixel.
- Dynamic range is the ratio of a pixel's saturation level to its signal threshold. High dynamic range is important for preserving details in both bright and dark regions of an image.
- Uniformity is the consistency of response for different pixels under identical illumination conditions.
- Shuttering is the ability to start and stop exposure arbitrarily. A global electronic shutter allows capturing crisp images of fast-moving objects.
- Speed is a characteristic describing the readout time, namely how fast the signal is transferred from a sensor.
- Windowing is the ability to read the signal from a part of sensor pixels.
- Blooming is an effect of overexposure diffusion to adjacent sensor pixels destroying details around the bright spot on an image.
- Biasing and clocking can be understood as operating on different voltage biases and implicating corresponding power consumption.

The undisputed advantages of CCD technology are higher pixel uniformity resulting in lower fixed pattern noise (FPN) and therefore clearer images and superior electronic shuttering. A uniform synchronous shutter is hardly implementable for CMOS, but is of vital importance for scenarios with fast-moving objects where object motion can lead to a distorted image.

Park [190] compares the CCD and CMOS sensors regarding in-car application. He designates the CMOS technology as more promising due to lower power consumption, better fit for the implementation of high dynamic range sensing (up to 140 dB), absence of blooming effect, possibility to operate in high temperature environment (up to 125-degree C), possibility of on-chip signal processing and windowing option.

In fact, mainstream CCD-based and most of CMOS-based sensors provide an optical dynamic range from 40 to 60 dB [132]. For reference, the dynamic range of a human eye reaches 105 dB [106]. These dynamic ranges are sufficient for uniformly illuminated scenes without extreme contrasts. In automotive applications non-uniformly illuminated scenes with extreme contrast are often presented. Some image areas can therefore be overexposed or underexposed losing important image details. The overexposure is especially harmful for CCD sensors due to the blooming effect. This is why the current tendency in automotive imaging is to promote CMOS technology for developing high dynamic range cameras. Such cameras have been introduced for example in [132] for vehicle occupant classification.

There are two ways to overcome the problem of the limited dynamic range. On the one hand, a sensor can be improved to extend the dynamic range by adjusting the exposure time for each pixel individually (so-called intelligent pixel) [220] or by assembling images taken under different exposure times [112, 162]. On the other hand, the dynamic range of a scene can be reduced by screening ambient illumination using artificial light sources emitting specific wavelengths and applying an optical cut-off filter passes through only a restricted range of wavelengths. The near-infrared (NIR) spectral range (750-1000 nm) is most suitable due to the natural imager sensitivity in this range and invisibility of NIR light to a human eye [132].

2.1.2 Semantic gap between an object and its appearance on the image

Regardless of sensor type, the camera output is a sequence of images. Hence, the major task of all image analysis systems is to extract knowledge from images. Having a noiseless highly detailed image, this task seems to be trivial for human beings, but not for computers. For automatic recognition systems there is a “semantic gap” between an object and its appearance on the image. In fact, an object may have different appearances on an image due to various illumination conditions, diverse poses, or partial occlusion. At the same time, two different objects may have the same appearance due to lack of illumination, overexposure, covering by shadows or other objects. This phenomenon is schematically illustrated in Figure 2.2.

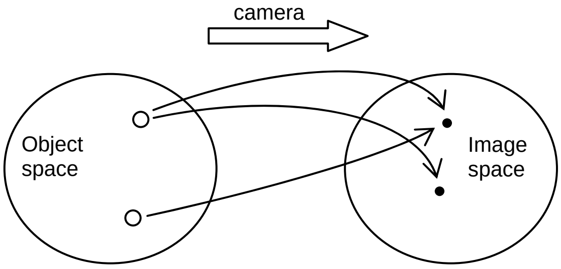


Figure 2.2: Ambiguous representation of an object on an image

The only possible solution is to find a pose and illumination invariant representation of the object in some imaginary space, let us call it feature space, by defining formal object descriptors, let us call them features. Ideally, each particular object is unambiguously described in the feature space by feature vectors representing the pattern of the object. The aim of a classifier is to determine object margins in the feature space for distinguishing objects. This trick is presented in Figure 2.3. The direct transformation from the object space to the image space is done by a camera and the inverse transformation should be accomplished by feature extraction and pattern classification algorithms.

Often feature extraction is addressed as a part of pattern classification. This inconsistency in terms is caused by the fact that the term classification is sometimes used to address not the complete pattern recognition process, but only a matching part of it. Sometimes, however, the term pattern classification is used equally to statistical pattern recognition designating the combination of classifier training (also referred to as machine learning) and actual classification. In this work, I

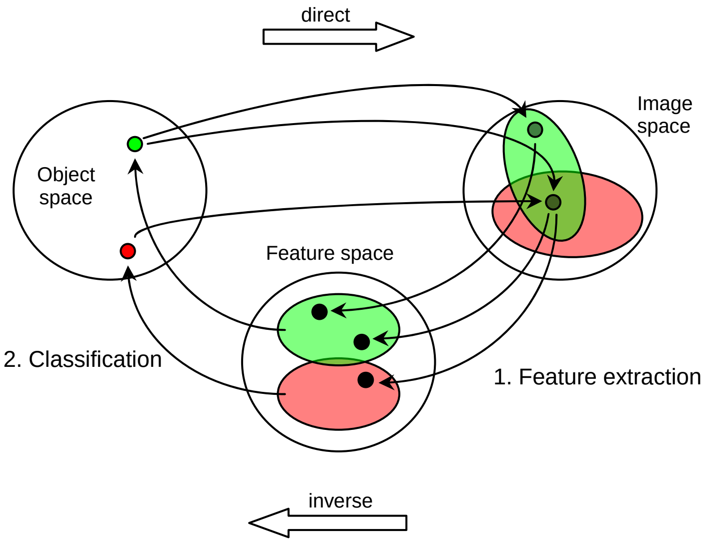


Figure 2.3: Two step approach to solve the ambiguousness

follow the terminology as it is used in the pattern classification book of Duda et al. [55].

2.1.3 Pattern recognition

Under the term pattern recognition one understands an automatic process of determining the class of a test sample under the condition of predefined classes. In our case classes are registered objects and the test sample is an image showing the object of interest. Generally, pattern recognition consists of five steps: image acquisition, image pre-processing, feature extraction and selection, matching, and decision making. Depending on the particular application some stages can be joined or become void.

There are two ways of classifying patterns: semantic (or syntactic) pattern recognition and statistical pattern recognition. Talking about semantic pattern recognition, heuristics-based approaches are referred to. The decision to assign an object to one or another class is done based on expert knowledge. In other words, decision rules are derived from heuristics. Examples of this method include using a rule of thumb, an educated guess, an intuitive judgment, or common sense [191]. In contrast, statistical pattern recognition engages machine learning approaches to derive decision rules from positive and negative examples. In other words, statistical pattern recognition is focused on the statistical properties of the training patterns expressed in probability densities and the decision boundaries are the result of an example-based cluster building process. Here, the training of a classifier or in the degraded case collecting reference data has to be provided prior to classification. The set of references (or mathematical description of class boundaries) represents a (statistical) model of a class. While expert knowledge is an indispensable attribute of semantic pattern recognition, the presence of positive and negative samples is a prerequisite of statistical pattern recognition. Figure 2.4 demonstrates the difference between semantic and statistical approaches. Let us briefly go through all stages in order to deeply understand the pattern recognition process.

Image acquisition is a hardware-related problem that is sufficiently discussed in the first subsection of this chapter. The output of the acquisition stage is either an image or a sequence of images.

The pre-processing stage addresses the image processing task where the raw image, captured by a camera, is adapted and passed on to the next stage aiming at emphasizing important details

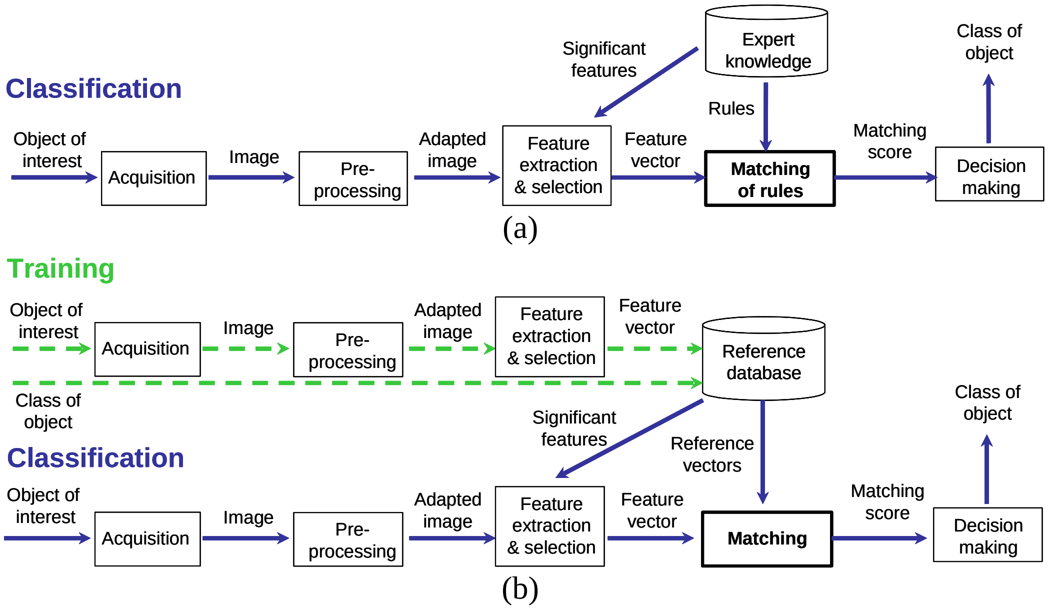


Figure 2.4: General pattern recognition workflow: (a) semantic vs. (b) statistical classification

(by e.g. scale change, rotation, contrast enhancement, edge detection etc.) and/or segmenting them. Inputs and outputs of pre-processing are images whereby after adaptation all original visual information about an object of interest is preserved. The major objective of pre-processing is maintaining feature extraction.

A feature is a characteristic of an object which can be formally measured so that the obtained feature values distinguish the object (or class of objects) from other objects (or classes of objects). A good feature possesses high inter-class variability and high intra-class similarity meaning that feature values are very similar for objects in the same class, and very different for objects in different classes. This property of a feature is also referred to as discrimination power. It is important to notice that features have to be invariant regarding irrelevant transformations of the input data [55]. For imaging, examples of such irrelevant transformations are changes in illumination, scale, rotation and translation. In fact, the environmental illumination, distance to a camera, viewing angle and captured region have significant influence to the object's appearance, but the object itself does not change.

Defining features is less formalized and a rather creative process which relies on specific domain knowledge or on expert's intuition. In cases when the domain knowledge is not presented, there is no guarantee that any proposed features properly describe object characteristics. However, there are features widely used in different imaging applications. These are for instance coefficients of the Fourier transformation [182], gray-level co-occurrence matrix (GLCM) [100], statistical invariant moments [111], chain codes [212], responses to Gabor filters [46] or Laws filters [144], local binary patterns (LBP) [185] and Haar-like features [242].

Single feature is seldom sufficient for accurate discrimination of several classes or even only two classes. Providing further features, the decision space becomes two-dimensional, three-dimensional and so on, so that decision boundaries between classes become more legible. However, the high number of features can lead to so called "curse of dimensionality" [13]. Points in feature space are then characterized by a low level of density so that the minimum distance between any two points becomes too close to the maximum distance between any two points and the difference between minimum and maximum distances becomes negligible in relation to the minimum distance. In application to statistical pattern recognition this means that the proper definition of decision boundaries requires an enormous amount of training samples (to reduce the sparseness), and this amount exponentially grows with the number of dimensions. The question of how many features are required for the proper distinguishing of all presented classes is the central question of feature extraction.

Nonetheless, the usual strategy is to define so many features as possible and then eliminate all irrelevant or redundant ones. This process is referred to as feature selection. Kira in [129] gives the following definition: "Feature selection is the problem of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept". According to [124] irrelevant features are those that are not class specific and redundant features are those that depend on other features or on a combination of other features.

To cut a long story short, formally, the feature extraction stage transforms an input image to a feature vector containing values of all considered features. During this process the visual information gets lost, but all information relevant for discriminating objects is preserved in feature vectors.

The first three steps of the pattern recognition workflow (acquisition, pre-processing and feature extraction) are common for semantic and statistical pattern classification. The fourth step however is fundamentally different. Rules-based matching is trivial. Feature values are individually checked against some criteria and a matching score is generated reflecting a measure of discrepancy.

In contrast, the fourth step of statistical pattern classification becomes complex. In the training phase this stage is represented by supervised machine learning. The main concern here is how to partition the feature space into regions so that all patterns in one region belong to one particular class. The result of partitioning is the set of decision boundaries or more general decision functions which are permanently stored in the reference database and referred to as class models.

Due to the fact that the image generation process is generally considered to be non-deterministic and even if one assumes it to be deterministic it is often noisy and during the training only incomplete information is available. So the learning problem is fundamentally ill-posed. The learning algorithms providing complex decision boundaries tend to overfitting. This phenomenon is shown in Figure 2.5. Suppose we have two features and therefore two-dimensional feature space. The task is to distinguish objects of two classes (circles and triangles). The distributions of training and test feature vectors are shown in Figures 2.5a and 2.5d correspondingly. Which decision boundary is optimal? The complex decision boundary perfectly splits training samples of the classes (see Figure 2.5b), but looks erroneous for splitting test samples (see Figure 2.5e). In contrast, the simple line as a decision boundary is not perfect for splitting training samples (see Figure 2.5c), but looks significantly better for splitting test samples (see Figure 2.5f).

The problem of proper selection of the complexity of the model based on a limited set of training samples (which often provide incomplete information about class distributions) is referred to as the generalization problem. Complex models provide bad generalization. The trade-off between the complexity of the model (number of decision rules) and a proper description of training data may be solved by the minimum description length principle (MDL) [210]. Grünwald writes [90]: "When two models of the data are equally well, MDL will choose the one that is the 'simplest' in the sense that it allows for a shorter description of the data. As such, it implements a precise form of Occam's Razor - even though as more and more data becomes available, the model selected by MDL may become more and more 'complex'!"

The same idea appears in the structural risk minimization approach introduced by Vapnik [240] and several statistical approaches for non-parametric inference. Here it is acknowledged that the machine generated data can be infinitely complex (e.g. not describable by a finite degree

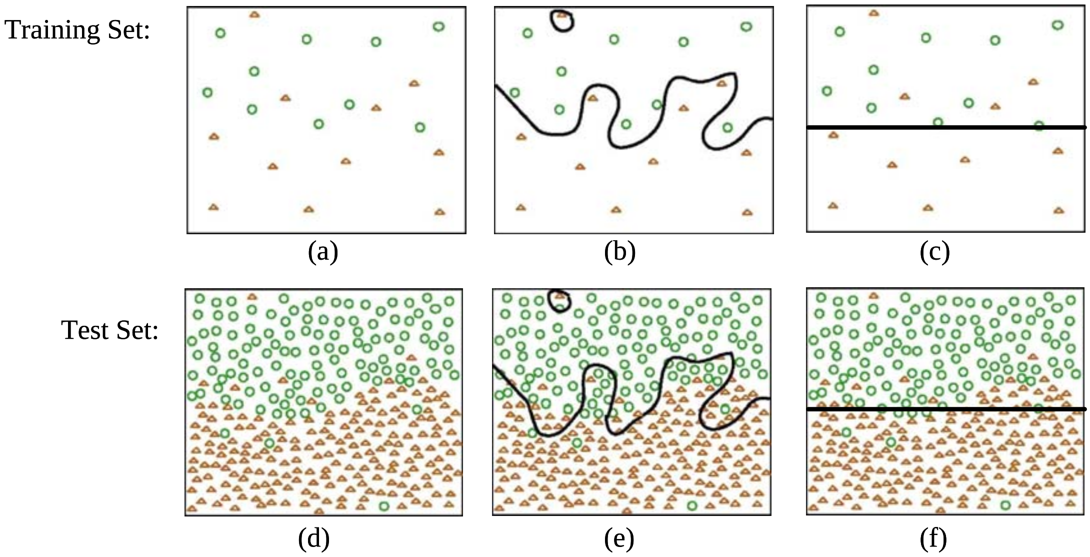


Figure 2.5: Generalization problem while having a small amount of training data: (b), (e) complex model is selected; (c), (f): simple model is selected (modified from [241])

polynomial). Nevertheless, it is still a better strategy to approximate it by simple hypotheses (e.g. low-degree polynomials) as long as the training set is small. The sophisticated machine learning algorithms apply the MDL principle as well as possess embedded feature selection routines. Wu [254] describes the top 10 machine learning algorithms identified by the IEEE International Conference on Data Mining in 2006. These are C4.5 [203], k-Means, Support Vector Machine (SVM) [239], Apriori, Expectation-Maximization (EM), PageRank [4], Adaptive Boosting (AdaBoost) [78], k-Nearest-Neighbor (k-NN) [44], Naive Bayes, and Classification and Regression Trees (CART) [26]. The wide range of machine learning algorithms can be found in the data mining software WEKA [248].

Independently from an applied machine learning approach, the most natural way to improve class models and to better estimate true underlying characteristics of classes is to get more training samples. However, in most pattern recognition applications collecting training data is restricted by scenario or even too time and resource intensive so that the amount of training samples is usually very limited. There are resampling methods for the artificial enlargement of a training set. Bootstrap aggregating (Bagging) is one example [25]. Given an original training set of N samples, bagging creates several subsets of M samples ($M < N$) through uniform selection with replacements. Clearly, by following this selection strategy, some samples in the subsets are duplicated. If N is large enough and M approaches N, then each subset is expected to contain approximately 63% of unique samples from the original set and remaining samples are duplicates. Such a subset is referred to as a bootstrap sample. Each bootstrap sample is used for training a classifier. The classification is then provided applying all classifiers independently which is followed by voting on the results. Bagging improves the stability and accuracy of machine learning algorithms and helps to avoid overfitting [59].

In the classification phase, the matching stage is the comparison of a current test sample with a model. The model can be represented by a set of decision rules (decision boundaries in the feature space), by the set of reference feature vectors or by a single feature vector. In the last two cases, the machine learning degrades to collecting reference data. The matching stage produces the matching scores representing discrepancy between the test samples and all class-models stored in the reference database. This discrepancy can be either a measure of similarity or dissimilarity. In case the class is represented by a single feature vector (e.g. mean vector) the matching requires the formal definition of a distance function (e.g. Euclidean distance). In case of multiple feature vectors representing a class, a distance function has to be extended by a matching rule (e.g. k-NN [44]). In case of the model-based representation, the matching consists in the successive application of decision rules providing probabilities that the test sample belongs to one or another class. These probabilities can even be considered as similarity scores.

The last stage of the pattern recognition workflow is decision making. An incoming matching score is a scalar value which is compared to a predefined threshold. Assuming the matching score is a measure of similarity then exceeding the threshold means the test sample can be assigned to the corresponding class. Otherwise the origin of the test sample remains undefined.

Concluding the subsection it is important no notice that there is no strict boundary between feature extraction and classification (matching). An ideal feature extractor yields a representation that makes the job of the classifier trivial. An omnipotent classifier does not need the help of a sophisticated feature extractor [55]. The same can be said about pre-processing and feature extraction. In the case that pre-processing perfectly prepares an image, feature extraction becomes trivial and the sophisticated feature extractor does not need any pre-processing. The boundary between matching and decision making is also imaginary. Some classifiers make an exact decision of class-membership based on non-metric decision rules (e.g. Decision Tree [202] dissembling membership probabilities, which makes the decision stage obsolete. Distinction of processing stages is done because of practical, rather than theoretical reasons. For instance, while feature extraction is considered to be a domain specific problem requiring expert knowledge for the definition of features, pre-processing and matching can be commonly used for any kind of images and any feature vectors correspondingly.

2.1.4 Biometrics as a special case of statistical pattern classification

The term "biometrics" is a combination of the words "bios" and "metro" which come from the Greek and mean life and metrics correspondingly. Thus biometrics literally means the measurement of life. In "Guide to Biometrics" [22], the authors introduce biometrics as science: "Biometrics is the science of identifying or verifying the identity of a person based on physiological or behavioral characteristics." The National Science and Technology Council proposes in their glossary [181] dual usage of the term biometrics to describe either a biometric characteristic, or a person authentication process, namely: "Biometric is a measurable biological (anatomical and physiological) and behavioral characteristic that can be used for automated recognition", or "Biometrics is a set of automated methods of recognizing an individual based on measurable biological (anatomical and physiological) and behavioral characteristics". The similar definition of biometrics as a process is given in the ISO/IEC standard [117]: "Biometrics is the automated recognition of individuals based on their behavioral and biological characteristics".

In order to avoid confusion with terms, biometric characteristics are further addressed as (biometric) modalities or (biometric) traits and the aforementioned process as biometric user authentication.

A physiological modality has a genotypic or phenotypic nature meaning that its appearance is predefined by genes or is caused by biological processes during an early embryo development. Examples of physiological modalities are: DNA, fingerprint, face, iris, retina, hand geometry, hand vein structure and ear. These modalities are considered to be static because they are unlikely to change over time and can be captured by a single shot. A behavioral modality reflects an activity of a user and arises from training. Through the repetition of same routines over the long period of time, muscles reproduce actions in the exactly same way. Examples of behavioral modalities are: voice, handwriting, gait and keystroke. These modalities are considered to be dynamic because people need time to reproduce it. Consequently, the acquisition has to be done over a certain period of time (e.g. video sequence for a gait).

Not every characteristic of human beings can be considered a proper biometric modality. Jain [119] poses seven requirements to biometric modalities:

1. Universality: occur in as many people as possible
2. Distinctiveness: as unique as possible, at best do not appear the same for any two persons
3. Permanence: do not change over time
4. Collectability: can be measured quantitatively
5. Performance: can be effectively calculated having limited resources
6. Acceptability: widely accepted by people in daily life, easy and comfortable to measure
7. Circumvention: cannot be (or can be very hardly) forged or duplicated

The modalities lacking distinctiveness are called soft biometrics [120]. Used for clustering of data in a reference database.

According to Miller [173], users can be authenticated based on "what they possess", "what they know" or "who they are". The possession-based user authentication implies that a user owns a physical object such as a key, an ID card, or radio-frequency identification (RFID) chip that grants the user access to the secret. The knowledge-based authentication implies that a user is aware of secret information granting access to the secret, for instance PIN or password. The third way to authenticate users refers to the unique biometric characteristics of a person. Bolle et al. [22] describes positive and negative properties of authentication methods which are summarized in Table 2.1. For security enhancement, the means of user authentication can be combined. One can use an ID card with an assigned PIN and/or a fingerprint stored on that.

Table 2.1: Properties of different approaches to user authentication (modified from [22])

Authentication method	Examples	Properties
Possession-based	Key, ID card, RFID chip, etc.	Can be shared, duplicated, or may be lost or stolen
Knowledge-based	PIN, password, etc.	Can be shared, or may be forgotten
Biometric-based	Fingerprint, face, iris, etc.	Not possible to share, can be hardly forged or duplicated, cannot be lost or forgotten

The main motivation to use biometrics for user authentication is the assumption that it is more secure then possession- and knowledge-based authentication simply because of its nature. Biometric modality cannot be lost, stolen, forgotten or even shared because it is directly linked to the person. The no-sharing property is a key factor increasing security, but eventually reducing comfort of use.

Biometric user authentication considered as a process is a special case of statistical pattern recognition. Human beings are placed in focus and the recognition of objects transforms to the recognition of biometric traits. The training phase is called enrollment, here the biometric samples are collected and stored in the reference database as biometric templates of users, see Figure 2.6.

According to Bolle et al. [22], authentication can be provided either in verification or in identification mode. During verification the claimed identity can be verified or not verified based on a biometric sample. During identification the identity of an originator of the biometric sample is determined. Figure 2.7 shows the difference between two modes.

Returning to the in-vehicle imaging applications, it has to be noticed that only biometric characteristics are under consideration that can be captured by a camera. These are face, iris and

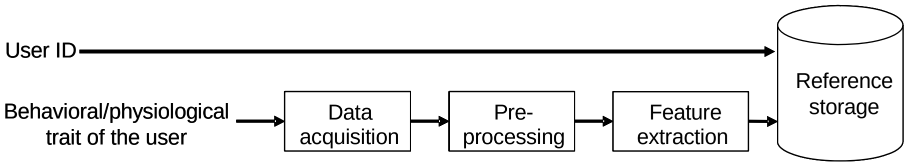


Figure 2.6: Enrollment of biometric data

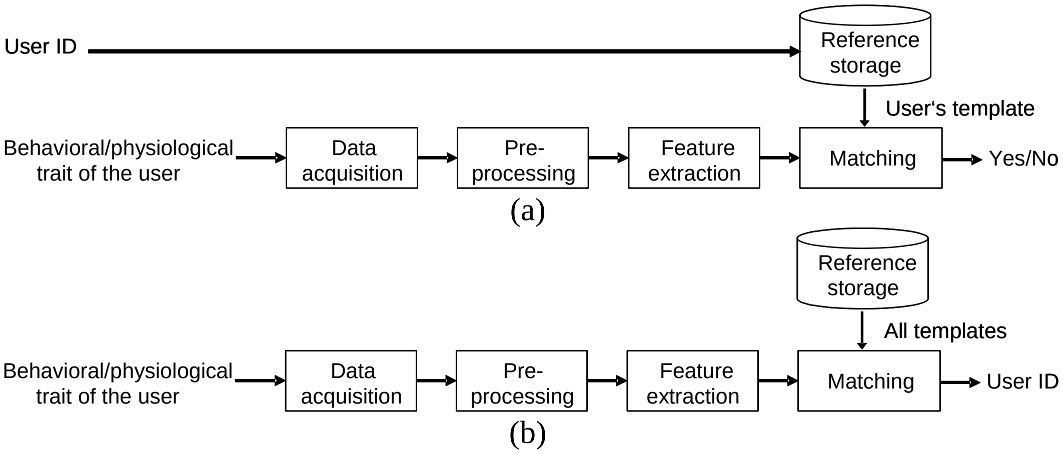


Figure 2.7: Biometric user authentication: (a) verification mode - the claimed identity is verified or not verified based on the biometric sample, (b) identification mode - the identity of an originator of the biometric sample is determined (rank list)

ear.

2.2 State-of-the-art, automotive related part

This section is comprised of the following subsections:

- Requirements to in-vehicle computer vision systems
- Drowsiness detection
- Support of smart airbag
- Discrimination of driver and front-seat passenger hands
- Car personalization
- Facial driver identification

2.2.1 Requirements to in-vehicle computer vision systems

Numerous sensors are currently integrated in modern vehicles. Fleming [73] provides an overview of automotive sensors and categorizes them to powertrain, chassis and body sensors. In accordance with the proposed categorization, body sensors are responsible for so-called body control functions, namely occupant safety, security, comfort, convenience and information. The author also gives examples of applications making use of body sensors: crash avoidance, crash worthiness, anti-theft, anti-intrusion, advanced airbags and occupant sensing. According to Fleming, computer vision approaches are mostly used in trucks for occupant monitoring, determination of pre-crash position and lane-keep assistance. He states that a vision sensor consists of a digital camera with typically

100K pixels and 120dB dynamic range, mounted on the windshield inside the cab, as well as image recognition software that incorporate lane detection and vehicle trajectory recognition algorithms.

Computer vision approaches have not reached a high saturation rate in automotive systems yet. This is explained by the complexity to interpret images and to derive the information about the presented objects. In the automotive field no assumptions can be made on the scene illumination or lighting contrast, which is a fundamental problem for all imaging devices, such as conventional digital cameras, structured light devices, or ToF cameras, because they basically measure the reflected light. The dynamic illumination variations (caused by e.g. direct sunlight, shadows, rain, or fog) need to be compensated by intelligent image processing algorithms. Hence, the subsequent processing must be robust enough to adapt to different environmental conditions and to their dynamic changes (such as transitions between sun and shadow, or the entrance or exit from a tunnel). Nonetheless, vision-based monitoring approaches become a trend in the development of modern driver assistance systems. Bertozzi et al. [17] provide a comprehensive survey of state-of-the-art approaches and development perspectives for vision-based systems in intelligent transportation systems (ITS). Besides other concepts, the authors pose general requirements to such on-board systems supporting autonomous driving, whereby these requirements appear to be general demands on automotive electronic control units (ECU):

- Robustness: The system must be robust enough to adapt to different conditions and changes of environment, road, traffic, illumination, and weather. Moreover, the hardware system needs to be resistant to mechanical and thermal stress.
- Reliability: A high degree of reliability is required especially for safety critical systems. Consequently, the project has to be thorough and rigorous during all its phases, from the requirements specification to the design and implementation. An extensive phase of testing and validation is therefore of major importance.
- Cost: Strict cost criteria need to be considered during the system design. As a rule of thumb the system should cost no more than 10% of the vehicle price. The system cannot be based on expensive processors and therefore is restricted to either off-the-shelf components or low-cost ad-hoc solutions. Operative costs such as power consumption need to be kept low as well.
- Compact size: The hardware components incl. sensors need to be kept compact in size and should not disturb car styling.
- User-friendly interface: The interaction between human beings and the system should not be complicated for a user. A driver should be able to switch the system off.

Regarding the design, Wikander [251] poses functional requirements to in-vehicle occupancy verification systems which can be generally adopted to all on-board vision-based systems:

- to operate at speeds ranging from 0 to 150 km/h;
- to operate in all kinds of weather, light, roadway, and traffic conditions;
- to operate invisibly or at least unobtrusive to users;
- to operate in a fully automatic mode requiring no action from vehicle occupants and eventually give a feedback to the driver;
- to be easily retrofit to existing vehicles or at least to demand minimal additions/changes to vehicle equipment.

Langnickel et al. [143] poses the requirements for biometric systems integrated into a car. The authors suggest considering the problem not only from the user's but also from the manufacturer's point of view. Regarding a user, eight criteria are mentioned:

- high authentication speed,

- low level of user interaction,
- low misclassification rates,
- high transparency,
- low invasiveness,
- low costs,
- anonymous and secure storage of personal data,
- high level of usability and administration.

A manufacturer focuses on seven criteria:

- high grade of automatic acquisition (incl. adaptation and liveness recognition),
- appropriate technical properties of system components (size, weight, power supply),
- high operating speed,
- high robustness of sensors against attacks and vandalism,
- low production costs,
- reasonable size of the reference database,
- effort to maintain the system.

Bertozzi et al. [17] divide automotive vision-based sensors into two categories: active and passive. For instance, laser-based sensors and millimeter-wave radars are classified as active sensors because they detect the distance to obstacles by measuring the travel time of a signal emitted by the sensor and reflected by the obstacle. Millimeter-wave radars are more robust to rain and fog than laser-based sensors but more expensive. To the drawbacks of both technologies one can assign the low spatial resolution and slow scanning speed. Imaging sensors acquire data in a non-invasive way by measuring the reflected light and therefore are defined as passive sensors. Imaging devices are less robust than millimeter-wave radars or laser-based sensors due to the drastic image distortions caused by varying lighting conditions (e.g. fog, darkness, or direct sunlight).

Current automotive compliant cameras include new important features that enable the solution of some basic problems directly at sensor level:

- image stabilization performed during acquisition,
- extension of camera's dynamic range,
- enhanced spatial resolution of a sensor,
- independent processing of pixels by CMOS sensors,
- integration of a CMOS sensor and a processing chip,
- extension of the frame rate.

From dozens of in-vehicle monitoring applications four can be distinguished as the examples for which the feasibility of visual sensing is confirmed. These applications include: driver drowsiness detection, occupancy detection to support smart airbag, discrimination of the front row occupants accessing the infotainment components and facial driver identification for car personalization. Moreover, the vision-based monitoring seems to be the most suitable sensing technology for these applications.

2.2.2 Drowsiness detection

Driver drowsiness, also called driver fatigue, has been widely recognized as a major contributor to highway accidents [99]. Although the official statistical information about the fraction of crashes caused by drowsy driving differs from study to study, the importance of driver drowsiness detection for safe driving is undisputed.

Based on the data from Crashworthiness Data System collected in 2001, it is reported that 4.4% of crashes can be linked to driver drowsiness [58]. Based on the statistic of National Highway Traffic Safety Administration (NHTSA), 1-3% of all police reported crashes and approximately 4% of fatalities are caused by drowsiness [131]. A study [69] from Australia reports the rate of approximately 6%. English researchers report rates of up to 16% [110]. Campbell et al. [29] report that in the USA at least 8% of so-called run-off-the-road crashes have happened because of drowsiness. The recent studies indicate even higher rates. A study from 2006, called 100-car Naturalistic Driving and sponsored by NHTSA, links 22-24% crashes and close to crash situations to driver drowsiness [130]. Another study from 2006, which is based on interviews with Canadian drivers, reports that nearly 60% of drivers have experienced drowsy driving and 15% confess falling asleep while driving [238].

The drowsiness detection problem has been studied since the late 1970s. According to Erwin [64], there are several behavioral indications preceding the onset of sleep which can be utilized for drowsiness detection:

- slow eyelid closures,
- increased number and duration of eye blinks,
- reduced pupil diameter,
- head nodding,
- excessive yawning,
- slowdown of breathing and heart rate,
- decline of muscle tone and body temperature,
- electromyogram (EMG) shift to lower frequencies and higher amplitude,
- increase of electroencephalogram (EEG) alpha waves,
- struggling to fight sleep and stay awake.

In 1976, Erwin [64] stated that slow eyelid closure is the best indicator of driver drowsiness. Since then, the focus has been laid on eyes and especially on blink behavior. The vision-based automatic systems make an effort to detect eye blinks and to measure blink frequency, blinking rate and the eye closure rate. These indicators are used for example in [235, 176] to represent driver awareness while operating a vehicle. There are laboratory experiments and field operational studies [53, 205] confirming that the eye closure rate also known as percentage of eye closure (PERCLOS) is the most reliable and valid indicator of driver drowsiness. In fact, the eyelids of a drowsy driver tend to droop covering a part of iris and pupil. Considering an image of an eye, the PERCLOS algorithm estimates the fraction of the covered part of the eye in proportion to the fully exposed eye for each video frame. However, the PERCLOS calculation is not trivial. It requires robust image processing approaches for the localization of eye components, namely eyelids, pupil and iris and deriving statistics from these components such as width-height ratio of a pupil or an iris. There are several studies endeavoring robust detection, localization and tracking of eyes [122, 245, 16, 20, 99].

Ji et al. [122] propose using artificial near-infrared sources for the frontal illumination of eyes. The lamp is composed of two LED rings of different diameters which are sequentially switched on. Two images are taken over these different illumination conditions. Due to the specific specular reflection of the pupil under infrared illumination, pupils become bright having very high contrast to the remaining image components. The image resulting from the subtraction on of two aforementioned images contains contrast peaks in pupil locations. This idea is taken over in [16] for locating pupils. Besides the PERCLOS, the authors calculate eye closure duration, blink frequency, nodding frequency, face position and fixed gaze. The registration of the head nodding is a very natural extension to eyes tracking whereby the whole head is in focus. This helps to cover the situations where the eyes are not visible because of head rotation. The robust computation of PERCLOS is very sensitive to noise and requires high-resolution and high-contrast images implying high camera cost. Hammoud et al. [99] propose the estimation of average eye closure (AVECLOS) to reduce the complexity and to increase the robustness of the system. Here, an eye is judged to be open or closed over a certain period of time.

Currently, vision-based drowsiness detection has become standard and already implemented on FPGA [255]. Such systems are offered by several vehicle suppliers [99, 137] and therefore can be considered as well-proven technology requiring no further fundamental research. Moreover, Toyota already introduced camera-based driver drowsiness detection system in 2006 and since then has integrated it in premium models of Lexus and Toyota. The sketch of this system is presented in Figure 2.8.

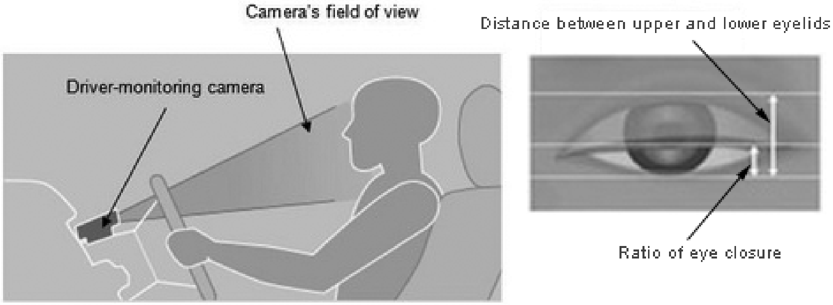


Figure 2.8: Toyota eye monitor (modified from [230])

Vision-based tracking of eyes and/or a head is not the only way to estimate driver drowsiness. Papadelis et al. [189] shows that the changes in electroencephalogram (EEG) reflect changes of brain activity when a driver becomes tired. Lin et al. [149] developed the EEG-based driver drowsiness detection system making use of the independent component analysis (ICA).

Recent effort toward robust drowsiness detection is linked to fusing multiple sensors and different modalities. These are for instance: pupil variations, head nodding, EEG data, steering wheel angle, vehicle standard lateral deviation, pulse waves, gaze direction and blink duration [85, 113, 102]. However, while the robustness increases, the system becomes more expensive and often less practical to be utilized in a car.

Recently, there has been a shift of paradigm in the struggle for safety on the road [74]. The vehicle related safety systems seem to approach the limit of their effectiveness. Sensing of the vehicle surroundings as well as monitoring of vehicle dynamics and driver behavior need to be extended by infrastructure related clues, namely a driver and a vehicle have to be considered as a part of an infrastructure. Fletcher [74] proposes to monitor driver's eye gaze and to check how it correlates with road events. This helps to immediately determine driver inattentiveness in critical situations and warn him or take over the vehicle if required. The system developer must be aware of "look but not see" case meaning that even if a driver looks in the same direction where a road event has happened, it is not guaranteed that he perceives it. However, Fletcher shows that warning a driver of missed road events significantly improves driving safety.

2.2.3 Support of smart airbag

The development of adaptive restraint systems is an important step towards the improvement of passive safety. The term adaptive means that safety belts and airbags are adjusted to passenger dimensions as well as that the airbag deployment is prevented at unoccupied seats and seats occupied by child restraint systems or small children. The airbag inflation velocity and firmness also depend on the severity of the impact. Currently used airbag control units possess up to 10 deployment stages. Hence, smart airbag requires data about the vehicle status at impact and also detailed meta-data about passengers.

Seat occupancy detection can be considered a subtask of occupant classification. While the objective of occupancy detection is to determine whether a seat is occupied or not, the occupant classification system determines the type of occupant. Generally, the seat state is assigned to one of six major groups: (1) 95% male, (2) 50% male, (3) 5% female, (4) forward-facing child seat (FFCS) with a 3-6 y.o. child, (5) rear-facing infant seat (RFIS) with an infant and (6) empty seat. The group membership is determined based on weight, stature and sitting height. An airbag must be deployed with normal power only for 95% and 50% male occupants. For 5% female and 3-6 y.o. child the low-risk deployment is prescribed. For RFIS and empty seat the airbag deployment is prohibited [7]. The exact regulations can be found in the NATHA safety standard [178].

In the synthesis report of the Texas Transportation Institute, Wikander [251] enumerates all sensory systems ever applied for the automated vehicle occupancy detection and provides the current state of the technology along with the development perspectives. The addressed sensory systems include:

- weight sensors,
- capacitive and electric field sensors,
- ultrasonic sensors,
- thermal infrared imaging,
- optical/NIR sensors,
- monocular (2D) systems,
- omni-directional imaging,
- stereo imaging,
- imaging using structured lighting,
- imaging using volumetric modeling,
- Time-of-Flight imaging,
- biometric sensors,
- smart cards and readers,
- telematics.

Reif [206] proposes to split the currently applied occupancy detection systems to four groups:

- weight-based systems measuring the force exerted on the seat rails or on a seat mat,
- capacitive systems measuring electric or magnetic fields,
- imaging systems using conventional cameras,
- imaging systems using optical proximity sensing by Time-of-Flight, structured light or stereo-vision principles.

In current vehicles, occupancy detection is exclusively implemented using occupant weight measuring devices. Such devices include seat mats incorporating expansion-measurement strips, force-sensing resistors or piezo-electric elements that change their resistance or voltage depending on the amount of force exerted. Weight sensors can be installed in the seat rails. The force applied deforms metal springs inside the sensor or causes pins to penetrate Hall elements. However, these weight-based systems can be fooled by heavy objects and moving persons while lifting from the seat leading to false alarms and inaccurate differentiations between objects and persons. Advanced systems are available for child-seat detection using transponders or resonators integrated in the child seats. Antennas in the passenger seat detect changes of the transponder field or modulations of the resonator signals, thus deactivating the airbag. All manufacturers must offer special car seats complying with the different child seats offered all over the world. Non-compliant child seats cannot be detected. This is why the NHTSA safety standard [178] does not accept these tagging systems and addresses exclusively weight measuring. The safety standard comes as a result of the NHTSA study [177] reported 175 fatalities between 1990 and 2008 which were caused by airbags, among them 104 children. Hence, the safety standard requires airbag suppression at seats occupied by child restraint systems or children less than 66 lbs. Figure 2.9 illustrates the functional principles of currently used seat occupancy detection systems. Both systems have the disadvantage of requiring components integrated in child and/or in passenger seats.

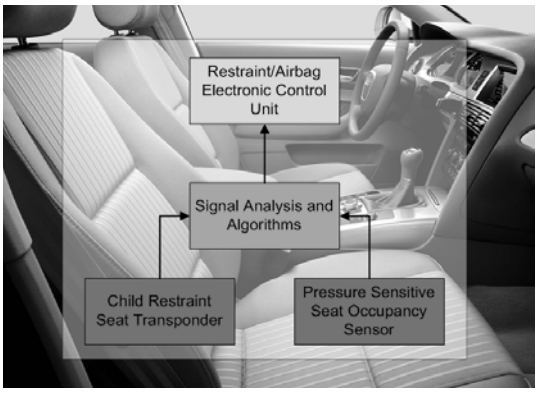

Restraint/Airbag Electronic Control Unit
Signal Analysis and Algorithms
Child Restraint Seat Transponder
Pressure Sensitive Seat Occupancy Sensor

Figure 2.9: Functional principle of seat occupancy detection (reprinted from [160])

Since the installation rate of side and head airbags in the rear seat area increases, occupancy detection becomes relevant also for the rear seats. Vision-based seat occupancy detection is a promising alternative to the currently used weight-based sensors especially for rear seats where adaptive restraint systems have not yet reached a high saturation rate.

Various suppliers and automobile manufacturers have recently given more attention to vision-based sensing. The systems are supposed to monitor individual seats or the complete interior and facilitate classification by image processing algorithms [258, 7]. However, such systems are not used in mass production yet, but very intensively developed in academic domain. Krumm and Kirk [136] put an emphasis on video-based occupancy detection. They propose the classification algorithm using a monochrome camera and a pair of monochrome cameras for stereo vision. The classification rates yield 99.5% and 95.1% for monochrome and stereo vision correspondingly. Koch [132] introduces a camera for high-dynamic range (HDR) imaging and shows its applicability in an automotive environment for real-time occupant classification. Park [190] studies and analyzes camera-based approaches for real-life scenarios and proposes face detection and tracking for improved distinguishing between passengers and objects.

Another important task closely related to occupant classification is the detection of so-called out-of-position (OOP) situation. In case of an unorthodox sitting posture, the airbag can seriously injure or even kill the passenger. So the airbag must not deploy in OOP situations. Vision-based OOP detection is addressed in [35]. The authors propose successive distinction of empty seats from persons, infants from adults, and finally adults from children. Farmer and Jain [68] consider OOP as a part of occupant classification concentrating on occupant tracking for the distinguishing between adults and infants. Krotosky et al. [135] propose the head detection algorithm based on stereo vision which is used for occupant detection. The reconstruction of disparity maps is done in real-time and is successful in more than 91% of frames. Trivedi et al. [233] take over the idea of head detection and tracking for occupant posture analysis and apply it in stereo and thermal infrared images. They introduce critical OOP regions (COOP) that can be seen in Figure 2.10

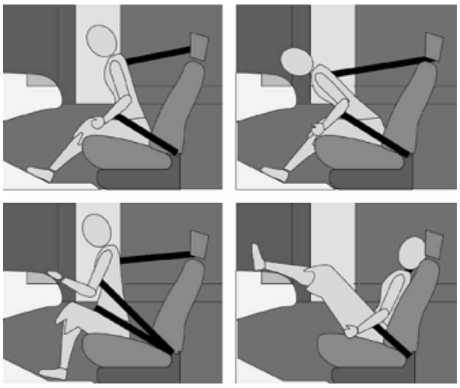


Figure 2.10: Examples of passenger out-of-position (OOP) situations; right column includes critical OOP situations when the airbag deployment may lead to fatality (modified from [233])

and provide the algorithm for determining all typical passenger postures. The aforementioned techniques are quite successful providing a recognition rate of at least 90% in cases when the camera faces one particular seat. However, it is not feasible to use multiple cameras, one for each seat, due to the extremely increasing cost.

The approach developed in this thesis makes use of an omni-directional camera to cover all seats simultaneously. The similar camera setup is introduced in [249]. The omni-directional camera is used to detect passenger faces and hence to indicate which seats are occupied by subjects. The proposed algorithm is based on the face detection approach of Viola and Jones [242] applying Haar-like features and the AdaBoost classification cascade trained from ten 30-second videos including 300 frames each and tested on 600 pre-selected frames. The approach yields the detection rate of 85% with 0.075 false detections per frame (45 false detections in 600 test frames). In our work [158], different face recognition approaches are compared to examine the general applicability of such algorithms for 360-degree in-vehicle images and to indicate the most suitable one. For detection of an empty seat face detection is complemented by advanced template matching which is described in [160].

Most recent occupant classification systems are based on proximity sensing. Fritzsche et al. [80] introduce approaches toward the implementation of a photonic mixer device (PMD) sensor which is the most well-known time-of-flight camera today. Devarakota et al. [49] propose very similar techniques recognizing empty seat, RFIS, FFCS and an adult. Despite relatively low recognition rates provided by current systems, range sensing slowly gains market positions and will dominate the scene soon because of the plausibility of the measuring technique and the availability of sensors.

2.2.4 Distinguishing between driver and front-seat passenger hands

Recently, the number of control switches on the center panel of an automobile has been continuously increasing. Considering that one button or switch usually handles only one particular function, the number of switches in modern automobiles becomes immense. Consequently, the usability of the center panel drastically decreases. In order to reduce the costs and to increase the comfort of center console handling, the conventional switches will be replaced by integrated center panel (ICP). A modern ICP combines new infotainment functions (e.g. DVD-player, web browser) with classical automotive functions (e.g. air conditioning, car radio). The interaction with an ICP is often provided by a knob manipulator on the armrest (e.g. BMW iDrive or Audi MMI) or by a touch screen. A recent modernization step is the manufacturing of dual-view displays for the ICP.

These devices allow users to see different contents on the same display depending on the viewing angle (see Figure 2.11). For example, the driver can follow the directions given by the navigation system, while the passenger watches a film or scrolls through the play-list in a music player. This novel interaction concept implies the personalization of the ICP manipulator, which means that the automotive system is aware of whether a driver or a passenger is currently interacting with the ICP. Moreover, it must be precluded that the passenger's interactions (e.g. with the music player) affect the driver's display content and vice versa.

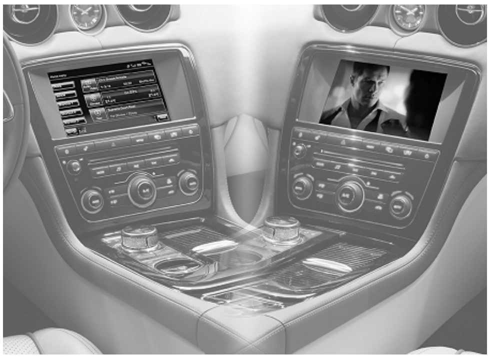


Figure 2.11: Changing display contents depending on the viewing angle (modified from [211])

Due to the innovativeness of the dual-view displays, there is a lack of scientific research in this area. However, there are studies on user discrimination not related to automotive environment. Two user discrimination technologies have commonly been used so far. The first one is based on electric field sensing (EFS) and has been already used in the domain of multi-touch tabletops for a long time. The DiamondTouch multi-user table was presented by Dietz and Leigh [52] in 2001. The location dependent electric field is generated here by the touch panel. The users' bodies capacitively couple the touch panel containing transmitter electrode with receiver electrodes installed in the working environment. Pickering et al. [194] propose a very similar solution for a car environment. Transmitter electrodes are integrated into the car seats and each console element contains a receiver electrode. The touching of a receiver electrode by a passenger changes an electric field. The capacity of an electric field and its changing can be measured on receiver electrodes. Each seat has an individual frequency in the transmitter electrode. Therefore, through the measuring of frequency, it is possible to identify which passenger interacts with the consol element. Since electric fields are rather a problem than an advantage of modern automobile and the addition of new electric fields is very undesirable, is better to conduct the user discrimination with alternative sensing technologies.

The second technology is vision-based sensing. Cameras are currently integrated into a passenger compartment of luxury series cars and will become a standard in middle-class vehicles in the near future. A video camera permanently captures images of the center console and its surroundings. User discrimination is done by means of computer vision algorithms.

Askar et al. [5] describe a user discrimination approach to video conferencing. Based on the skin color model, the algorithm looks for skin regions on an image and provides hand segmentation based on a pre-calculated threshold value. The segmentation result is a binary image with blobs on the hand position. Furthermore, hand tracking is done. Indeed, it is pointed out that overlapping of hands leads to discrimination problems. Due to the color segmentation the presented algorithm is limited to the color images and moreover has the same reaction to every near-skin color item. Thus, it is reasonable to use other characteristics for hand segmentation.

Althoff et al. [3] suggest using gray-scale images captured under active near-infrared illumination. Nevertheless, the system presented in this paper does not deal with user discrimination but with recognition of hand gestures in a car passenger compartment. However, the computer vision techniques used in [3] can be applied for driver/passenger discrimination. Due to active infrared illumination, a hand in the middle area of a console is brighter than a background. The gray value histogram of an image has a peak in case a hand is presented on the image. Thus, the hand can be easily detected and segmented by means of a gray value threshold. Additionally, the authors involve a motion based algorithm, which measures motion entropy on image sequences. This technique greatly supports the forearm and hand detection and segmentation. Geometric algorithms use center point and direction vectors to distinguish between forearm and hand.

Another study on in-vehicle gesture recognition is provided by Reiner [207]. He investigates gesture-based HMI interfaces and gives an overview of finger and hand gesture recognition systems which can be applied in a car passenger compartment. Among other technologies he mentions: data gloves, RFID tags, colored markers, accelerometers, multi-touch surfaces, conventional camera, range camera, capacitive proximity sensing, thermal imaging and ultrasonic tracking. Two technologies are emphasized as most promising: capacitive proximity sensing and optical range sensing with a Microsoft Kinect camera. The objective of both systems is to provide a link between the stretched index finger of the right hand and the cursor position on the display mounted on the center console. The capacitive system makes use of an antenna mounted on the gearshift to track finger movements in the gearshift area. The range imaging system observes the gearshift from above and recognizes both static and dynamic finger and hand gestures so that automotive systems can be operated using predefined gestures.

The same objective, namely to discriminate which of the front row occupants is accessing the infotainment controls, is targeted by Cheng and Trivedi [37]. The proposed real-time vision-based user determination system captures visible and near-infrared images of the front row seat area in a vehicle and decides who is currently active: driver, passenger, or no one. The modified histogram-of-oriented-gradients (HOG) descriptor is used to represent the image area over the infotainment controls. Support vector machine is applied to classify each image to one of the three classes. In order to improve the classification result the median filter is applied over time. The average correct classification rate yields 97.9%.

The system presented in this thesis has been developed since 2008 at AMSL lab of the University of Magdeburg. Herrmann et al. [108] proposes to analyze the amount of motion in pre-selected driver and front-seat passenger regions. So the discrimination system can recognize the direction from which the forearm comes. The evaluation is done in a real car and based on numerous video frames with variable lighting conditions, whereby the illumination is not entirely controlled. Same to Cheng [37], the decision who is currently active (driver, passenger, or no one) is made for each single frame, which is definitely not the best way for the evaluation of the recognition performance. In contrast, our next work [159] investigates well-defined illumination conditions in the laboratory using car simulator and provides the more appropriate action-based evaluation. There are two image processing workflows proposed. The first one determines motion regions based on difference images resulting from subtracting two subsequent frames. The second one extracts the forearm shape based on the edge detector and estimates the direction of the principal axis in the resulting shape. In our last investigation [107] the motion-based approach is supplemented by hand localization based on the Viola-Jones object detector [242]. The finite state machine is utilized to weight motional and spatial decision components and provide the more reliable user discrimination.

In contrast to Cheng [37], the effort here is made toward user discrimination for accessing the dual-view touch screen. Hence, the discrimination system operates only while the contact between a finger and a touch screen arises and therefore chooses from two alternatives: driver and front-seat passenger.

2.2.5 Car personalization

The term car personalization refers to the ability of a car to recognize its driver. This option gives a boost to car comfort, safety and security. An example of comfort improvement is the fully transparent memory function which automatically adjusts the seat, steering wheel, rear-view mirrors, air conditioning and radio for the driver. Safety is improved by the activation of driver-specific assistance systems e.g. speed restriction for young or inexperienced drivers. Security is improved due to biometric immobilizer which allows the access or ignition exclusively to a registered driver. The recently wide-spread infotainment systems can also greatly benefit from car personalization due to, for instance, automatic activation of the user profile for internet services.

The currently used authentication systems include memory buttons (MB) and a radio-frequency identification (RFID) tag integrated into an individual smart-key. In both cases a driver explicitly reports his identity that cannot be verified by the system. Hence, the link between the stored profile and a person in the driver seat cannot be confirmed. This is why biometric authentication seams to be the best way to recognize a driver and to activate the registered profile. Despite evident advantages, biometric authentication implicates risks evoked by possible misclassifications and the need for preserving sensitive biometric data. Therefore, different biometric modalities are studied and compared.

Intensive academic research has been provided on driver recognition based on speech, face and driving behavioral signals. With the support of the Japanese government, the Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University started a project on collecting and analyzing biometric data in an automotive environment [126]. They simultaneously captured spoken text, video of a face, speed, acceleration, accelerator pedal pressure, break pedal pressure, steering wheel angle and engine RPM. Finally the database of 812 drivers was completed with a total recording time of over 600 hours.

Igarashi et al. [114] explore the uniqueness of three driving behavioral signals (pressure to accelerator and brake pedals, vehicle acceleration) based on the subset of CIAIR database of 30 drivers. The authors apply the linear prediction model (LPM) and Gaussian Mixture Models (GMM) to each signal for modeling individual driver behavior. The performance is formulated in terms of prediction accuracy and intra-driver prediction error. Finally, the identification rate of 73.3% is reported. Further, Benli et al. [14] examine five signals (brake pedal pressure, accelerator pedal pressure, engine RPM, vehicle speed, steering wheel angle) on the CIAIR subset of 100 drivers. As before, the GMM is applied for feature extraction and several fusion methods including fixed rules and trainable combiners for classification. In the best case, the observed identification rate amounts to 99.65%. Generally, the experiments show that driving behavioral signals are not discriminative enough to indicate the person, but they can be a useful part of a multi-modal biometric signature for driver authentication.

In further experiments on CIAIR database, Erdogan et al. [63] go beyond driving behavior signals and fuse them with canonical modalities, namely speech and face. The standard feature extraction techniques are used in all three cases, namely GMM for driving signals, Mel-Frequency Cepstral Coefficients (MFCC) for speech and principal component analysis (PCA) [234]) for face. Although the considered CIAIR subset contains only 20 drivers, the reported results are very impressive. The authors report a classification accuracy of 100% after the fusion. The modality contributing the most to the driver recognition is speech with an accuracy of 98%, just as the face falls behind with accuracy of 89%. Stallkamp et al. [229] take the same CIAIR subset and concentrate exclusively on the face modality. Two local appearance-based approaches are compared. The first one is based on PCA and the second one uses discrete cosine transform (DCT) for the face representation. The authors make an effort on the comparison of single frame authentication and video-based authentication. The DCT-based approach outperforms PCA as well as video-based recognition outperforms the single-frame-based one. The reported accuracy rates of the DCT approach amount to 88.5% for single frame and to 96.3% for video.

Scheidat et al. [219] suggest fusing face, speech and soft biometrics (body weight, body volume). They propose a family scenario and collected the database of four persons in a car simulator. The tests are done in verification mode and the results are provided in terms of equal error rate (EER). The selected algorithm for face is PCA and for speech the combination of MFCC and GMM. Experiments show very poor results with EER over 15% even for this small database. The fusion of these biometric signals with body weight and body volume measured by cushion-based pressure sensor improves the EER to 6.25%.

Although most authors assert that the integration of additional modalities is essential for achieving high reliability and high identification rates, the performance indices show that even single biometric modality is able to provide sufficient recognition performance. We believe that the amplification of the recognition system is the deadlock in the development. Considering the costs, possible sensor breakdowns, convenience and complexity of use, a unimodal biometric system has immense advantages.

There are several unconventional recognition systems such as driver recognition from sitting postures [208] and driver recognition from dynamic handgrip on steering wheel [34]. The test results show a lack of uniqueness of these biometric characteristics. However, the low recognition performance does not stop the authors from reporting the aforementioned technologies to be promising for in-car application. Table 2.2 summarizes the proposed approaches for biometric driver identification. Following the ideas of Langnickel [143] and Büker [20], our study indicates the face as the most appropriate and probably only feasible biometric modality for convenient and reliable driver authentication.

Table 2.2: Biometrics in a car (IA - Identification Accuracy, EER - Equal Error Rate, TAR - True Accept Rate, TRR - True Reject Rate)

Study	Biometric modalities	Algorithms	Users in DB	Recognition performance
Igarashi et al., 2005 [114]	driving signals	LPM + GMM	30	IA = 73.3%
Benli et al., 2008 [14]	driving signals	GMM + score fusion	100	IA = 99.65%
Erdogan et al., 2005 [63]	driving signals + speech + face	GMM (driving signals) + MFCC (speech) + PCA (face) + fusion	20	IA = 100%
Stallkamp et al., 2007 [229]	face	PCA + DCT	20	IA = 96.3%
Scheidat et al., 2009 [219]	face, speech and soft biometrics	PCA (face) + MFCC&GMM (speech)	4	EER = 6.25%
Riener et al., 2008 [208]	sitting postures	Statistical features	34	IA < 25%

Chen et al., 2011 [34]	dynamic handgrip on steering wheel	Likelihood-ratio-based classifier	?	TAR = 85.4%
				TRR = 82.65%

Almost all studies on driver authentication imply the integration of DSP board into the vehicle's electronic network. However, the personal smart phone can serve as the driver's identity bearer as well. The identities of passengers can be determined from the registered phone numbers, or the biometric recognition can be done directly on a smart phone [186] wirelessly connected to the car.

2.2.6 Facial driver recognition

Over the last time, the face modality has become the highest research interest for the automotive community. Different optical devices as well as pattern recognition algorithms have been practically examined in a car passenger compartment for both face detection and recognition.

Fukumi [82] applies the frontal near-infrared camera for locating the face and determining of its rotation. The template matching algorithm looks for face candidates with one of five possible rotations (frontal, left, right, 45-degree left, 45-degree right), and then the neural network approves or disapproves the result.

Wu and Trivedi [253] introduces the facial landmark detection and global geometric constraints for robust face localization. The authors evaluate the approach on the database of five subjects, whereby the images are captured with the in-car camera facing the driver.

Ishak et al. [116] presents a complete face recognition system in a car. The face detection is done by means of Classical and Fast Neural Networks (CNN/FNN) straightforwardly operating on grey-value images. The face recognition is done based on PCA as well as on the combination of PCA and LDA (linear discriminant analysis [12]). The experiments are carried out on the database of 24 subjects in verification mode. In the best case, the true accept rate amounts to 91.43% while the false accept rate amounts to 0.75%. According to the supervised scenario where a driver is asked to look straight ahead at the camera, the performance cannot be judged as sufficient for practical realization.

Moon and Lee [174] propose a 3D face recognition system for driver authentication. Several algorithms for alignment, fitting and comparison of 3D meshes are considered. Further, the 3D data is combined with texture. The system is declared to be robust to pose and illumination changes. In fact, the reported classification accuracy gained on texture is exactly the same to the accuracy gained on texture and shape. Moreover, the 3D faces for the evaluation of the recognition performance are not collected in an automotive environment. Although the number of test persons is 110, the reported performance of 90.4% accuracy for frontal faces and 85.7% accuracy for pose-variant faces can be judged only as hypothetical.

Chen et al. [36] propose face matching for driver exchange detection as a part of a drunk-driving prevention system. For the face detection the authors use Viola-Jones algorithm [242] and for the face matching the combination of PCA and LDA, which has become standard recently. Considering automotive environment, the performance has been evaluated based on the ARTC driver's face database including 21 persons. The reported accuracy of the driver exchange recognition is 100%. This work shows that standard face detection and recognition methods can be successfully applied to an in-car environment when the matching is done in supervised conditions and the verification mode is used.

In this thesis, appearance-based and feature-based approaches for face recognition are compared. As a proponent of appearance-based approach the FaceART system developed in [161] is used. The localized and normalized facial images are represented by coefficients of PCA transform and classified by k-NN or Adaptive Resonance Theory Mapping (ARTMAP) neural network. As a proponent of feature-based approaches the commercial face recognition system Luxand FaceSDK [155] is used. This system is accessible on the market and has a solid recognition performance on common face databases. Here, fiducial points are localized and the face is modeled by an elastic graph of these points.

The performance testing is carried out on three databases with 6, 51 and 54 drivers whereby the data was collected in a real car at different locations and under various illumination conditions.

Table 2.3: State-of-the-art systems in facial driver recognition (IA - Identification Accuracy, EER - Equal Error Rate, TAR - True Accept Rate, FAR - False Accept Rate)

Study	Camera position	Task	Algorithm	Users in DB	Recognition performance
Fukumi, 2005 [82]	frontal camera	locating the face and determining of its rotation	neural networks	1379 images for one person	IA = 85.5%
Wu and Trivedi, 2005 [253]	half-side camera	facial landmark detection for face localization (eyes localization)	SIFT descriptors + KNN (Gabor features + AdaBoost)	5 (70)	IA = 90.9%
Ishak et al., 2006 [116]	frontal camera	face recognition	CNN/FNN + combination of PCA and LDA	24	TAR/FAR = 91.43%/ 0.75%
Moon and Lee, 2007 [174]	not in car	3D face recognition	fitting of 3D meshes	110	IA = 90.4%
Chen et al., 2011 [36]	frontal camera	Pair-wise face comparison	Viola-Jones + combination of PCA and LDA	21	IA = 100%
Stallkamp et al., 2007 [229]	half-side camera	face recognition	Viola-Jones + PCA + DCT	20	IA = 96.3%

2.3 Evaluation methodology

This section introduces the general evaluation methodology together with certain statistical methods for the estimation of the recognition performance of pattern recognition systems. Concepts are adopted for the most part from the domain of biometric user authentication.

2.3.1 Empirical study

A pattern recognition system represents an intricate chain of image processing algorithms which cannot be expressively described in form of mathematical equations and even if they could, the equations will have a very complex structure implying that outcomes have highly nonlinear dependency on incomes. Hence, solving such systems analytically is not possible. In other words, the recognition system is a black box providing chaotic relations between inputs and outputs and any effort to formalize the internal structure by a third party will not succeed. So, the performance evaluation of a pattern recognition system can be carried out only empirically.

The term empiricism is originated from ancient Greek and refers to the paradigm of perceiving phenomena through observations or experience, and rejecting dogmatic beliefs. In other words, according to empiricism, knowledge can be gained only from experience and evidence. Therefore, an empirical study consists of: (1) gathering observable data using scientific instruments, (2) inducting the research hypothesis, (3) deducting consequences of the hypothesis as testable predictions, (4) testing the hypothesis with new empirical material, and (5) evaluating outcomes. According to Groot [105] these steps form a loop. An act of the observation is referred to as an experiment and the result of an experiment is empirical evidence. Empirical evidence is usually analyzed quantitatively meaning that the numerous outcomes of experiments statistically prove or disprove the research hypothesis. Empirical evidence is never strict. The results of statistical testing are expressed in probabilities saying that the hypothesis is likely true or likely false.

Experimental data is comprised of test samples (incomes) together with the ground truth (expected outcomes). A recognition system receives test samples as an input and returns classes assigned to the samples. The more assignments match the ground truth, the higher the recognition performance is. From this description it becomes obvious that the recognition performance depends not only on algorithms but also on the set of test samples. The favorable test set can lead to the overstated recognition performance and vice versa. The main question of every experimental study is how to choose this limited set of test samples, so that the estimated performance approaches the "real" performance of the system.

2.3.2 Two-class pattern recognition

All pattern recognition problems can be assigned to one of two groups: two-class problems and N-class problems. Solving a two-class problem, a pattern recognition system chooses from two alternatives. Usually the question is posed in the form: "Is it an object of interest or not?" In an N-class problem, the selection is made from N alternatives. The typical question is "Which object do you see?" Any N-class problem can be divided up into the sequence of two-class problems as in the following example. Assume there is a blue ball in a picture. The picture is presented to a color recognition system. The problem can be posed as an N-class problem with the question: "What is the ball's color?" or as a sequence of two-class problems with questions: "Is this ball red?", "Is this ball green?", "Is this ball blue?" and so on looking over all known alternatives. Note that the number of alternatives is always fixed. In biometrics, user verification is a typical two-class problem and user identification is a typical N-class problem.

Hypothesis testing

For the evaluation of two-class decision making systems, there is an established and deeply studied statistical model called statistical hypothesis testing. Since it is important for the understanding of the error rates indicating the recognition performance, the basics of this model are briefly introduced. A detailed description can be found for instance in [213]. Hypothesis testing starts with posing a null hypothesis H_0. In opposition to the null hypothesis, the alternative hypothesis H_A is posed. Both hypotheses together form the complete decision space meaning that the rejection of one hypothesis immediately leads to the acceptance of another. Hypothesis testing consists of examining the null hypothesis by making experiments. The null hypothesis is considered critical and, therefore, can be rejected only in case of strong evidence against it. An experiment can provide two true decisions and two false decisions as summarized in Table 2.4. If the null hypothesis is true, it can be correctly accepted or falsely rejected. If the null hypothesis is false, it can be falsely accepted or correctly rejected.

Table 2.4: Confusion matrix of a two-class problem (hypothesis testing)

Ground truth Decision of the system	$\mathrm{H_0}$ is true ($\mathrm{H_A}$ is false)	$\mathrm{H_0}$ is false ($\mathrm{H_A}$ is true)
$\mathrm{H_0}$ is rejected ($\mathrm{H_A}$ is accepted)	False positive, Type I error, $P(\mathrm{H_A} \mid \mathrm{H_0})$	True positive, $1 - P(\mathrm{H_0} \mid \mathrm{H_A})$
$\mathrm{H_0}$ is accepted ($\mathrm{H_A}$ is rejected)	True negative, $1 - P(\mathrm{H_A} \mid \mathrm{H_0})$	False negative, Type II error, $P(\mathrm{H_0} \mid \mathrm{H_A})$

Historically, hypothesis testing is characterized by two errors. The type I error (also referred to as a false positive rate (FPR)) is the conditional probability to accept the alternative hypothesis given the null hypothesis is true $P(H_A|H_0)$. The opposite value $1 - P(H_A|H_0)$ is called the specificity. The type II error (also referred to as a false negative rate (FNR)) is the conditional probability to accept the null hypothesis given the alternative hypothesis is true $P(H_0|H_A)$. The opposite value $1 - P(H_0|H_A)$ is called sensitivity. In most practical applications the error rates are not equal. In biometric login systems, for instance, false positive decisions (called false accepts) are graded as critical errors evoking security risk of granting access to an intruder. In contrast, false negative decisions (called false rejections) only reduce convenience by prohibiting access to a genuine user.

The errors considered in hypothesis testing can be easily adopted for the proposed automotive applications. For the simplified model of seat occupancy detection, the type I error is that the occupied seat is recognized to be empty and the type II error is that the empty seat is recognized to be occupied. At a glance, missing an occupant and suppressing an airbag seems to be much more critical than false airbag deployment on an empty seat. However, in a real system the choice is usually made from the other two alternatives: a seat occupied by an adult (normal airbag deployment), and a seat is empty or occupied by RFIS, or by an out-of-position adult (airbag suppression). Here, at first sight, both errors seem to be equally critical because suppressing an airbag on an occupied seat as well as deploying an airbag on a seat occupied by RFIS or out-of-position person can lead to serious injuries or even to fatality. However, from the manufacturer's point of view, airbags killing people are much more critical than falsely suppressed airbags failing to protect people. Thus, the adoption of errors of hypothesis testing is not trivial here. Occupant classification is rather an N-class problem due to varying stages of airbag deployment in relation to the occupant's body dimensions.

For user discrimination, two operational scenarios are proposed in Section 4.2: frame-based and action-based decision making. In case of the frame-based scenario, the system simultaneously decides for a driver and a passenger whether the person is interacting with the center panel or not. Hence, the errors are: "an interaction with the center panel is missed", and "a missing interaction is detected". The critical moment is when contact between a hand and a control component occurs. Missing the hand may lead to an incorrect reaction of the system. In contrast, if the system falsely recognizes an action when no interaction occurs, this decision can be ignored. In case of the action-based scenario, the system chooses between driver and passenger hands only during the contact with the center panel. Here, both kinds of misclassification are equal and the errors can be formulated as: "a driver is recognized as a passenger" and "a passenger is recognized as a driver".

The verification scenario within the facial driver recognition system is a typical example of adopting hypothesis testing for a biometric system. The type I error (critical error) is recognition of an impostor as a genuine driver thus unblocking the engine ignition. The type II error is blocking the engine ignition for a genuine user who was misclassified as an impostor. Driver identification is the N-class problem which is beyond standard hypothesis testing.

Recognition errors (FAR vs. FRR)

Hereafter, hypothesis testing is applied to a biometric system as in [22]. In doing so I stick to particular hypotheses to better demonstrate this evaluation concept. It is done without loss of generality and can be applied to any other pattern recognition system with different hypotheses, as shown above. So, the hypotheses are posed as follows:

H_0: the test person is an impostor,
H_A: the test person is a genuine user.

Practically, the decision to accept or to reject the null hypothesis is made based on the test statistic revealed in an experiment. In my considerations the matching score coming from the matching/classification module of a pattern recognition system plays the role of the test statistic. The matching score is a measure of either similarity or dissimilarity between a test sample and a reference sample in a training-based classification or a formal description of how well the test sample matches rules in a rule-based classification. However, without loss of generality, the matching score can be interpreted as a probability of matching and represented by a real value within the interval [0,1]. The higher the matching score, the higher the similarity between the biometric test sample of the person and his/her biometric template. The decision, whether to accept or to reject the null hypothesis is made based on the result of the comparison of the matching score and the predefined threshold. If the matching score s exceeds the threshold τ, the null hypothesis is rejected, and accepted otherwise:

$$\begin{aligned} &s \geq \tau \Rightarrow rejectH_0 = acceptH_A \\ &s < \tau \Rightarrow acceptH_0 \end{aligned}$$

After a series of experiments the quality of the recognition system can be identified in terms of the aforementioned error rates: false positive rate (α) and false negative rate (β). Considering a biometric system as a matching system or as a system granting access to a secret, it is more convenient to interpret false positive and false negative decisions as false matches and false non-matches or rather as false accepts and false rejects correspondingly. Formally, false match rate (FMR) denotes the probability of getting a matching score higher than the threshold in imposter trials and false non-match rate ($FNMR$) denotes the probability of getting a matching score less than the threshold in genuine trials. So the FMR and $FNMR$ are the errors of a matching algorithm. False accept rate (FAR) denotes the probability of accepting an imposter and false reject rate (FRR) denotes the probability of rejecting a genuine user. So, the FAR and FRR are the errors of a biometric system. In a positive access control system e.g. PC-login, a high matching score indicates a user and access is granted. Hence, the match is equal to acceptance and the non-match is equal to rejection. In a negative access control system e.g. border crossing, a high matching score indicates that the person is in a travel ban list and access is denied. Here, the match is equal to rejection and the non-match is equal to acceptance. Although considering $FMR/FNMR$ is more accurate, most researchers operate with FAR/FRR because in terms of biometrics they are more plausible. Hereafter, FAR/FRR are used when addressing error rates and contemplate a positive access control system where $FAR = FMR$ and $FRR = FNMR$.

After including the matching score to the definition of error rates, it becomes clear that the error rates are monotonic functions of the threshold. Formally, the error rates are given by the following equations:

$$\begin{aligned} FAR(\tau) = Pr(H_A|H_0) = Pr(s \geq \tau|H_0) \\ FRR(\tau) = Pr(H_0|H_A) = Pr(s < \tau|H_A) \end{aligned} \tag{2.1}$$

The error rates are entirely defined by probability distributions of genuine and impostor matching scores often referred to as probability density functions (PDF). Most biometric systems generate normally distributed matching scores in genuine as well as in impostor trials. In a perfect case the distributions do not overlap leading to the zero error rate when the threshold is located between the distribution curves. However, in practice the distribution curves always overlap as shown in Figure 2.12a. Considering a decision threshold τ, the FRR is the area under the PDF of genuine scores taken from minus infinity to τ and the FAR is the area under the PDF of impostor scores taken from τ to plus infinity. In statistics these functions are referred to as cumulative distribution functions (CDF). They are presented in Figure 2.12b.

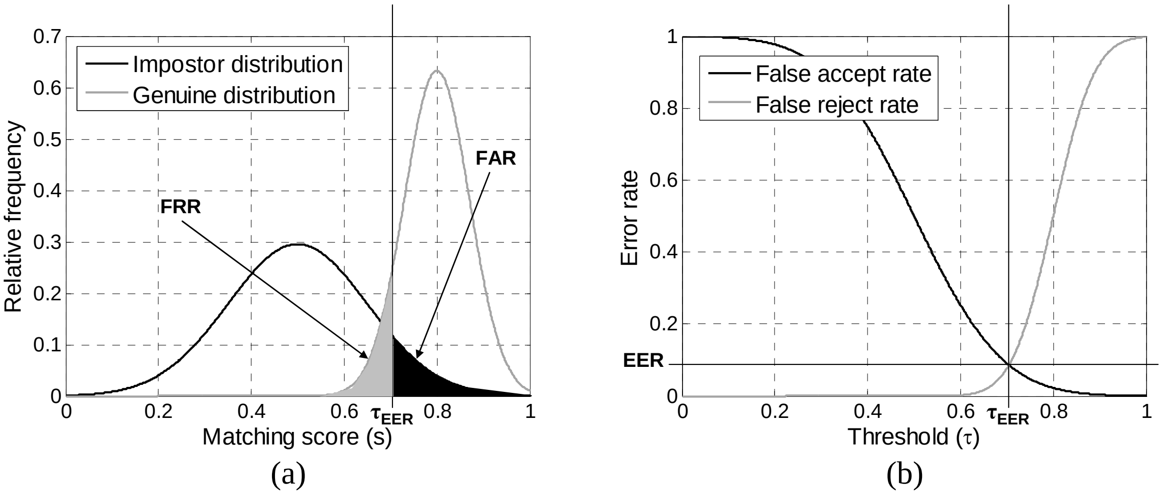


Figure 2.12: (a) Probability density functions (PDF) and (b) cumulative distribution functions (CDF) of genuine and impostor matching scores

Equal error rate as a scalar performance indicator

$FAR(\tau)$ and $FRR(\tau)$ are independent reciprocal curves. Shifting the threshold left, increases the FAR and decreases the FRR and vice versa. The separate consideration of only one curve leads to a misconception and makes absolutely no sense. On margins of the domain of a threshold the errors usually yield their extreme values of either 0 or 1. However, some systems produce matching scores that reach extreme values (0 or 1). In this case the error functions will not reach extrema on one of the margins. For example, if $s = 1$ for several impostor and genuine samples, then the $FAR(1) > 0$ and $FRR(1) < 1$ correspondingly. This phenomenon is caused by the unequal consideration of margins caused by the alteration of strict and non-strict inequality signs in the definition (see Equations 1). An easy solution to this end is considering $\tau \in [0 - \delta, 1 + \delta], \delta \to 0$ which guarantees that $FAR(0 - \delta) = FRR(1 + \delta) = 1$ and $FRR(0 - \delta) = FAR(1 + \delta) = 0$. In doing so, the error curves undoubtedly cross at some point that is the reference point for the performance evaluation in biometrics. Since both error rates are equal, the ordinate at this point is denoted as an equal error rate (EER) and the abscissa is referred to as a threshold where the EER is reached τ_{eer} (see Figure 2.12b).

$$FAR(\tau_{eer}) = FRR(\tau_{eer}) = EER \tag{2.2}$$

The EER is a scalar value indicating the recognition performance of a two-class pattern recognition system. Utilizing the EER for the comparison of biometric systems has become standard. However, the EER indicates the optimistic recognition performance reached on a training set. The decision threshold τ must be defined prior to performance testing. This is why at least two independent sample sets are required. The first set is used for training the system and determining the system parameters such as for example τ_{eer}. The second set is used for the actual evaluation. Note that using the τ_{eer} as the reference point for the performance estimation, the $FAR(\tau_{eer})$ and $FRR(\tau_{eer})$ might be not equal regarding the test set. Aiming at a single scalar value as a performance indicator, the half-total error rate ($HTER$) is often applied as an average between the FAR and FRR values obtained after testing.

In summary, the procedure of determining recognition performance can be formalized as follows:

1. Using the training set to extract the threshold τ_{eer} and to declare the $EER = FAR(\tau_{eer}) =$

$FRR(\tau_{eer})$ as an optimistic recognition performance expected in further runs;
2. Using the test set to calculate the $HTER(\tau_{eer})$ and to declare it as a true recognition performance.

In high-level security biometric systems the risk of intrusion is of major importance. Here, the recognition performance is estimated slightly differently:

1. Define the maximal value of the critical error which is FAR, e.g. $FAR \leq 0.001$;
2. Using the training set to extract the threshold $\tau_{fmr-max}$ so that $FAR(\tau_{fmr-max}) = 0.001$ and to permanently store $\tau_{fmr-max}$ as a decision threshold;
3. Using the test set and $\tau_{fmr-max}$ to calculate real $FAR(\tau_{fmr-max})$ and $FRR(\tau_{fmr-max})$. $FAR(\tau_{fmr-max})$ is expected to yield 0.001 and $FRR(\tau_{fmr-max})$ indicates the performance.

Receiver operating characteristic

Another common way of comparing two-class pattern recognition systems is utilizing a receiver operating characteristic (ROC) curve. This curve presents the FRR in relation to FAR [22]. Recognition performance is better the closer the ROC curve approaches the origin. Systems are compared by plotting ROC curves on the same diagram (see Figure 2.13. ROC curves can be presented for both the training set and the test set.

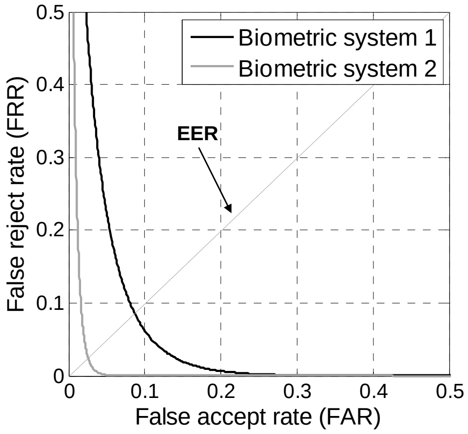


Figure 2.13: Receiver operating characteristic (ROC) curves of two biometric systems: biometric system 2 significantly outperforms biometric system 1

Sometimes FRR is substituted by the true accept rate (TAR): $TAR = 1 - FRR$. This is often used in detection systems to clearly show the detection rate indicated by TAR in relation to the given limits of false detections indicated by FAR. The ROC is then referred to as detection error trade-off (DET) [22].

2.3.3 Statistical significance of hypothesis testing

In practice, the error rates are set equal to relative frequencies of the corresponding outcomes in an experiment as described by Equations (2.3).

$$\begin{aligned} FAR(\tau) &\sim \text{Number of impostor samples with } s \geq \tau / \text{ Number of all impostor samples} \\ FRR(\tau) &\sim \text{Number of genuine samples with } s < \tau / \text{ Number of all genuine samples} \end{aligned} \tag{2.3}$$

This can be done since a probability is understood in the "Frequentist" sense as a relative frequency of occurrence of an experiment's outcome when repeating the experiment.

Relative frequency of an error as a point estimation

Hereafter, an error of an experiment is described in a general sense with no relation to FAR or FRR. Assume that an experiment consists of N trials with two possible outcomes in each: non-error (0) or error (1). In statistics, this experiment is referred to as a Bernoulli experiment. The number of errors e in this experiment is given by Equation (2.4) where $e_j = 1$ if an error occurs in j^{th} trial and $e_j = 0$ otherwise.

$$e = \sum_{j=1}^{N} e_j \tag{2.4}$$

The observed relative frequency of errors is:

$$P^* = \frac{e}{N} = \frac{1}{N} \sum_{j=1}^{N} e_j \tag{2.5}$$

In statistics, P^* is the common point estimate of the true error probability P. This estimate is considered optimal because it is unbiased and efficient [213].

Confidence interval

However, the most important question is, how well the observed relative frequency of errors P^* approximates a true value P which is the mathematical expectation of the error in further trials. A point estimate is perfect in case the trials are independent and the number of trials approaches infinity. In practice, the number of trials is very limited. Moreover, from a statistical point of view, 3 errors in 100 trials is not equal to 30 errors in 1000 trials and not equal to 3% of the actual error rate. In order to address this problem, the interval estimate is used to estimate the maximal deviation of the observed relative frequency from the true value at a desired confidence level. The confidence level is given in percent (say 95%) and expresses the confidence of an observer that the true value of the parameter is within the confidence interval, or that the 95% of the observed confidence interval includes the true value of the parameter. The margins of the confidence interval are calculated from percentiles of the normal distribution.

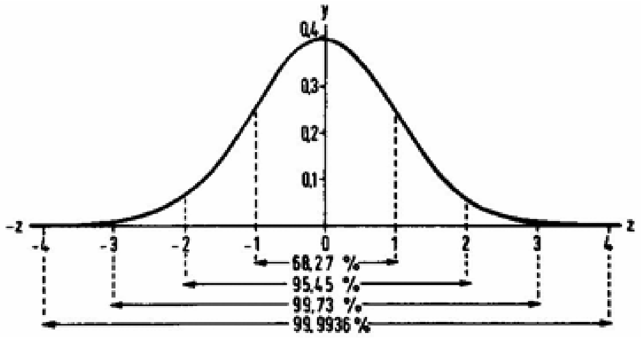


Figure 2.14: 68.27th, 95.45th, 99.73rd and 99.993th percentiles of the normal distribution $\mathcal{N}(0, 1)$ (modified from [213])

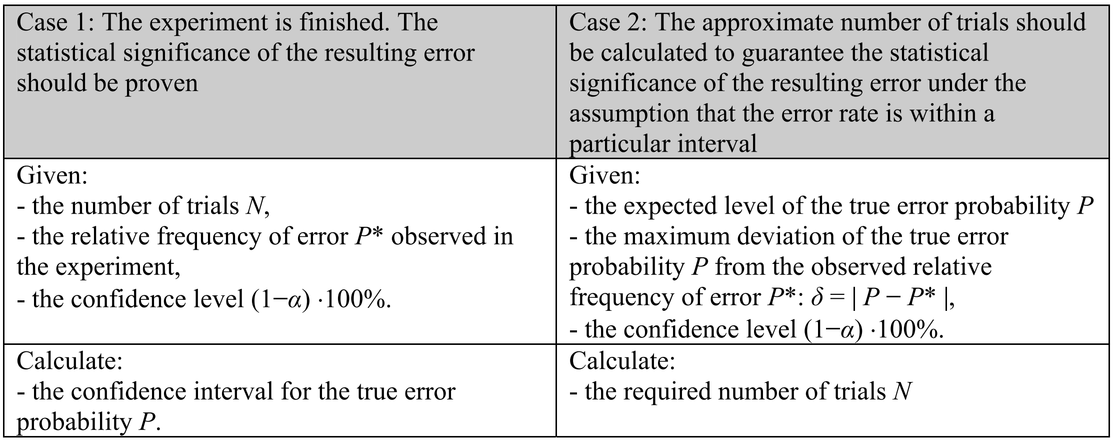
The number of errors e undergoes the binomial distribution $e \sim B(N, P)$. According to the de Moivre-Laplace theorem, if N is large enough, the binomial distribution of e approaches the normal distribution $\mathcal{N}(NP, NP(1-P))$ with the expected value $E(e) = NP$ and the variance $Var(e) = NP(1-P)$ [70]. Hence, the distribution of the mean $P^* = e/N$ is also normal with the expected value $E(P^*) = P$ and the variance $Var(P^*) = P(1-P)/N$. Various rules of thumb may be applied to decide whether N is large enough. The most common rule is that both NP and $NP(1-P)$ must be greater than 10 [250]. The confidence interval of P is given by Equation (2.6) changing distribution parameters to their estimates.

$$\left[P^* - z_{1-\frac{\alpha}{2}}\sqrt{\frac{P^*(1-P^*)}{N}}, P^* + z_{1-\frac{\alpha}{2}}\sqrt{\frac{P*(1-P^*)}{N}}\right] \tag{2.6}$$

$$\begin{gathered}100\%(1-\alpha) = 90\% \Rightarrow z_{1-\frac{\alpha}{2}} = 1.68\\100\%(1-\alpha) = 95\% \Rightarrow z_{1-\frac{\alpha}{2}} = 1.96\\100\%(1-\alpha) = 99\% \Rightarrow z_{1-\frac{\alpha}{2}} = 2.58\\100\%(1-\alpha) = 99.9\% \Rightarrow z_{1-\frac{\alpha}{2}} = 3.29\end{gathered}$$

In practical research two situations occur as illustrated in Table 2.5. Either after finishing an experiment, the confidence interval is calculated, or prior to an experiment, the number of trials in an experiment is calculated so that the resulting estimation of the error probability is statistically significant.

Table 2.5: Carrying out of an experiment



Case 1: The experiment is finished. The statistical significance of the resulting error should be proven	Case 2: The approximate number of trials should be calculated to guarantee the statistical significance of the resulting error under the assumption that the error rate is within a particular interval
Given: - the number of trials *N*, - the relative frequency of error *P** observed in the experiment, - the confidence level (1−*α*) ·100%.	Given: - the expected level of the true error probability *P* - the maximum deviation of the true error probability *P* from the observed relative frequency of error *P**: *δ* = | *P* − *P** |, - the confidence level (1−*α*) ·100%.
Calculate: - the confidence interval for the true error probability *P*.	Calculate: - the required number of trials *N*

In Case 1, the interval margins can be calculated using (2.6). If no errors occur, the margins of confidence interval are calculated differently. As described in [250], the lower margin P_{min} is obviously 0 and the right margin P_{max} can be derived from the probability estimation of mutually exclusive events $e_j = 0$ (no error occurs). Since the probability of getting no error in j^{th} trial is $Pr(e_j = 0) = 1 - Pr(e_j = 1) = 1 - P$, the probability of getting no errors in N independent trials is $(1-P)N$. For the given confidence level $\beta = 1 - \alpha$, the confidence level for mutually exclusive events is $1 - \beta$. Combining these two facts in (2.7), the upper margin P_{max} can be calculated using (2.8).

$$(1 - P_{max})^N = 1 - \beta \tag{2.7}$$

$$P_{max} = 1 - \sqrt[N]{1-\beta} \tag{2.8}$$

In Case 2, the number of trials N can be derived from the following equations:

$$Pr\left(P^* - z_{1-\frac{\alpha}{2}}\sqrt{\frac{P^*(1-P^*)}{N}} \leq P \leq P^* + z_{1-\frac{\alpha}{2}}\sqrt{\frac{P^*(1-P^*)}{N}}\right) = 1-\alpha \tag{2.9}$$

$$Pr\left(|P - P^*| \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{P^*(1-P^*)}{N}}\right) = 1-\alpha \tag{2.10}$$

with δ to denote $|P - P^*|$ and the confidence level of 95% ($z_{1-\alpha/2} = 1.96$):

$$\delta \leq 1.96\sqrt{\frac{P^*(1-P^*)}{N}} \tag{2.11}$$

then the approximate value of N can be derived from (2.12):

$$N \approx 1.96^2 \frac{1}{\delta^2} P^*(1-P^*) \tag{2.12}$$

If the observed error frequency P^* yields zero, the approximate number of trials N is derived from (2.7) where $\delta = P_{max} - P^* = P_{max}$.

$$N \approx \frac{log(1-\beta)}{log(1-\delta)} \tag{2.13}$$

Rules of thumb for the confidence interval estimation

There are two rules of thumb used to provide a simple way of stating an approximate confidence interval for the true error probability P:

- The Doddington's rule of 30 [199] states: "Test until 30 errors occur, then the true error probability P is within 30% of the estimated one P^* at a confidence level of 90%" (in other words for $\alpha = 0.1$, $|P - P^*| \leq 0.3P^*$).
- Rule of 3 [246] states: "If no errors occur in N tests, then the error probability P is less than $3/N$ at a confidence level of 95%".

2.3.4 N-class pattern recognition

The results of an N-class pattern recognition problem can be represented in the form of the confusion matrix. An example of the confusion matrix with three categories A, B and C is given by Table 2.6.

Table 2.6: Confusion matrix of a 3-class problem

Ground truth Decision of the system	A	B	C
A	#(A as A)	#(B as A)	#(C as A)
B	#(A as B)	#(B as B)	#(C as B)
C	#(A as C)	#(B as C)	#(C as C)

The correct decisions (successes) are on the main diagonal: A is recognized as A, B is recognized as B, and C is recognized as C. The false decisions (misclassifications) lie above and below the main diagonal: A is recognized as B or C, B is recognized as A or C, and C is recognized as A or B.

The ground truth and decisions of the system can be generally interpreted as two raters A and B who/which select categories. The confusion matrix (also called a two way contingency table) characterizes the extent of agreement between raters.

Classification accuracy

Providing the classification accuracy is the simplest and the most plausible way of evaluating pattern classification systems. Formally, the classification accuracy (further referred to as $Pr(\alpha)$) is the probability of correct category assignment. In practice, the probability is approximated by the relative number of correct decisions observed in some experiment. With regard to the confusion matrix, the accuracy is the sum of the main diagonal elements divided by the whole number of samples in the test set. Since the observed classification accuracy is the relative frequency of successes, the confidence interval can be obtained in the similar way as described in section 2.3.3.2 assuming that the successes undergo binomial distribution (success or failure) with the normally distributed mean that represents the classification accuracy.

Classification accuracy is a naïve measure of the extent of agreement because it does not take into account the number and distributions of test samples. Assume a classification system assigns all samples to one class (say A) out of three possible classes (say A, B and C). If class A contains 1000 samples and classes B and C contain 5 samples each, the classification accuracy is approximately 99% because the system has classified 1000 out of 1010 samples correctly. In fact, this system is absolutely useless producing very low accuracy rates on a dataset of uniformly distributed samples. So, if the accuracy rate is solely used to evaluate an identification system, the numbers of samples in classes need to be very similar.

Kappa statistic

The kappa statistic [42] helps to overcome the limitation of accuracy by measuring an inter-rater agreement. Formally, the classification accuracy measures the extent of agreement while kappa also takes into account the agreement occurring by chance. The value of the kappa statistic κ is given by (2.14):

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{2.14}$$

where $Pr(a)$ is the probability of an agreement among raters, and $Pr(e)$ is the probability of a chance agreement. $Pr(a)$ is estimated by the observed relative frequency of agreements (equally to the classification accuracy). $Pr(e)$ is estimated based on the observed relative frequencies of each observer randomly choosing each class. The perfect agreement is indicated by $\kappa = 1$, while no agreement is indicated by κ approaching zero or even $\kappa < 0$. Landis and Koch [140] have proposed the human-friendly categorization for interpreting kappa: $\kappa \leq 0$ - poor agreement, $\kappa \in (0, 0.2]$ - slight agreement, $\kappa \in (0.2, 0.4]$ - fair agreement, $\kappa \in (0.4, 0.6]$ - moderate agreement, $\kappa \in (0.6, 0.8]$ - substantial agreement and $\kappa \in (0.8, 1)$ - almost perfect agreement. Practically, the $Pr(a)$ is calculated straightforwardly

$$Pr(a) = \frac{1}{n}\sum_{i=1}^{q} n_{ii}$$

and $Pr(e)$ is calculated using different heuristics e.g. for Cohen's κ -statistic [42]

$$Pr(e) = \sum_{i=1}^{q} p_{A_i} p_{B_i}$$

or for Scott's π -statistic [223]

$$Pr(e) = \sum_{i=1}^{q} \left(\frac{p_{A_i} + p_{B_i}}{2} \right)^2$$

where

$$p_{A_i} = \frac{n_{A_i}}{n} = \frac{1}{n} \sum_{j=1}^{q} n_{ij}, p_{B_j} = \frac{n_{B_j}}{n} = \frac{1}{n} \sum_{i=1}^{q} n_{ij}$$

are derived from the confusion matrix (2.15). A value n_{ij} is the number of samples assigned by Rater A to i^{th} class and by Rater B to j^{th} class, q is the number of classes and n is the total number of samples.

$$\begin{array}{cc|cccc|c}
 & & RaterB & & & & \\
 & & 1 & 2 & \dots & q & Sum \\ \hline
RaterA & 1 & n_{11} & n_{12} & \dots & n_{1q} & n_{A1} \\
 & 2 & n_{21} & n_{22} & \dots & n_{2q} & n_{A2} \\
 & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
 & q & n_{q1} & n_{q2} & \dots & n_{qq} & n_{Aq} \\ \hline
 & Sum & n_{B1} & n_{B2} & \dots & n_{Bq} & n
\end{array} \tag{2.15}$$

Referring back to pattern recognition where Rater B is the ground truth and Rater A is the recognition system it becomes obvious that n_{B_i} denote the number of samples in i^{th} category and n_{A_i} denotes the number of samples assigned to i^{th} category. If all categories contain the same number of samples (say m) then $p_{B_i} \equiv m/n$ and the $Pr(e)$ in the κ -statistic's equation also becomes m/n:

$$Pr(e) = \sum_{i=1}^{q} p_{A_i} p_{B_i} = \frac{m}{n} \sum_{i=1}^{q} p_{A_i} = \frac{m}{n} \sum_{i=1}^{q} \frac{n_{A_i}}{n} = \frac{m}{n}$$

Substituting $Pr(e)$ by m/n in (2.14), the κ -statistic equals the classification accuracy $Pr(a)$.

Gwet [96] states that the κ -statistic (as well as π -statistic) has serious conceptual flaws making it very unreliable. The chance-agreement $Pr(e)$ can reach the value of 1 meaning that two raters agree by chance with a probability of 1. This contradicts common sense because the chance-agreement probability is not expected to exceed 0.5. Hence, Gwen proposes the AC1-statistic as the conceptually well-founded alternative heuristic for the estimation of $Pr(e)$ [96, 97]. In pattern recognition, the κ -statistic yields extremely low values (even when the extent of agreement is high) only in cases of a highly unequal number of samples in different categories and/or if almost all samples are assigned to only one category. This behavior of κ does not contradict the general concept but indicates either the deficiently composed test set or the inefficiency of the classifier.

In the example from the previous subsection where all samples are assigned to class A and the classification accuracy approaches 99%, the κ -statistic yields 0 confirming the inefficiency of the classifier.

Cumulative match characteristic

In biometrics, where an N-class problem is identical to user identification, there is an alternative indicator of the recognition performance called cumulative matching characteristic (CMC). In large-scale applications that include a high number of users even very sophisticated recognition systems are not able to return the correct identity as the best match for the test sample. This is caused by the fact that different persons sometimes possess very similar biometric patterns (e.g. faces of twins or doppelganger). In pattern recognition this problem is referred to as the inter-class similarity. However, the correct identity is often on one of the first places in the sorted list of the returned identities. This is why the decision making process of identification systems dealing with a high number of users is often supported by a human supervisor. An identification system returns a rank-list of identities and the supervisor picks the correct person from the list. This is called rank-based identification. The items in the rank-list are ordered regarding observed matching scores. Hence, returning correct identity on the first place of the rank-list becomes inessential. It is important that the outgoing list (also called watch-list) includes the correct identity. CMC answers the question: "What is the optimal length of the watch-list?" CMC is derived from the rank probability mass (RPM) function that gives the probability $Pr(r)$ that the correct identity occurs at the position r in the rank-list. CMC is a cumulative sum of RPM showing for all possible lengths of the rank-list the probability that the correct identity is in the list. CMC is a monotonically increasing function (see Figure 2.15). The higher the probability of the first entry and the earlier the CMC curve approaches 100%, the better the identification performance. According to Bolle et al. [22], CMC is the standard performance statistic of biometric search engines.

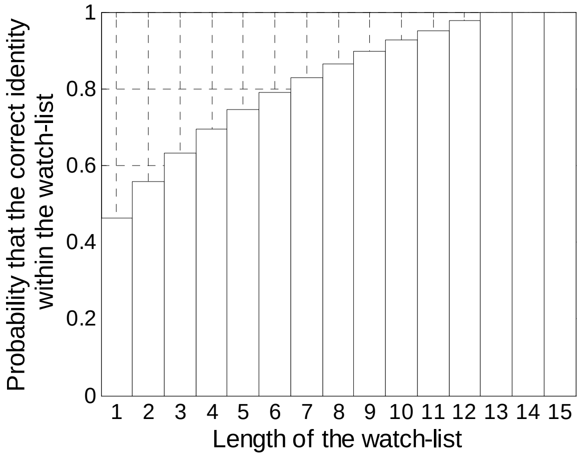


Figure 2.15: Cumulative match characteristic (CMC) curve of a rank-based identification system providing the watch-list of 15 categories. The classification accuracy yields 46.35% while the list of 13 categories includes the correct identity with the probability of 100%

2.3.5 Relation between verification and identification rates in biometrics

Since the majority of biometric systems can operate in both modes: verification and identification, in practice important questions are: "What relation between verification and identification indicators exists?" and "Should a researcher provide both, or is one enough?"

Bolle et al. [23] assert that CMC can be computed from FAR and FRR and does not offer any additional information but represents another way of displaying data. In case two users are registered, the CMC curve can be directly derived from ROC. This is also affirmed by Grother and Philips [93] who show that the classification accuracy is equal to an area under the ROC curve when ROC is understood as DET defined in 2.3.2. In case more than two users are registered, the relation between CMC and distributions of matching scores is not so trivial. However, the equations for calculating CMC for score distributions are provided in both aforementioned studies [93, 23].

An experimental study on the question whether a “good” verification system can be a “poor” identification system and vice versa is made by DeCann and Ross. [48]. The authors examine the relationship between the area under the DET curve (AUC) and the weighted rank-M identification accuracy derived from the CMC curve. An AUC value indicates the verification performance. The weighted rank-M value indicates the identification performance. The results show that usually “good” verification systems are “good” identification systems and vice versa. However, the high variance of matching scores in impostor distribution can drastically reduce the identification performance and only slightly the verification performance. The opposite case arises when the variance of impostor matching scores is very low and the difference between means of impostor and genuine distributions is also low.

In the proposed driver identification scenario there is no need for CMC because of the absence of a human supervisor. Hence, the focus is on the rank-1 accuracy. The accuracy is supported by the kappa statistic to address the distribution of test samples between drivers and the eventual tendency of the recognition system to choose one particular driver. In the driver verification scenario, the focus is on the EER which is calculated together with its confidence interval.

Chapter 3

3 Methodology and concept

This chapter consists of three parts. Firstly, the designing concept of automotive applications is introduced. Secondly, the important algorithms from domains of image processing and pattern recognition are explained in detail. These algorithms are used as construction elements in computer vision approaches applied for the introduced automotive applications. The chapter finishes with the introduction of the evaluation methodology which is for the most part adopted from the domain of biometric systems but can be generally used for any kind of pattern recognition systems.

3.1 Methodology of designing automotive applications

This section introduces my vision of the designing process of automotive applications which perform vision-based monitoring of a car passenger compartment.

In the concept proposed here, an automotive application consists of five components: application scenarios, operational scenarios, an acquisition system, illumination sources and computer-vision algorithms. Application scenarios define purposes of an application namely why the car retrofit with the new application becomes more attractive and which benefits a driver and/or a passenger gets. Operational scenarios define how the recognition system operates namely how the system interacts with a driver or a passenger and at what moments the system must generate outcomes based on the camera's video-stream. Operational scenarios are responsible for the realization of application scenarios. Moreover, they pose requirements to hardware components which include an acquisition system and illumination sources. The acquisition system captures images and passes them onto image processing algorithms. So, it requires at least an imaging sensor, an analog-digital converter and a DSP board or a conventional PC for video storage and processing. The selection and positioning of illumination sources need to be brought into conformity with the type, location and the angle of view of the camera. This is why illumination sources and a camera actually belong together. Splitting them into two different components is motivated by the fact that they can be independently substituted. The acquisition system together with illumination sources poses demands on computer vision algorithms just as the utilization of particular software solutions requires specific hardware. This is why hardware and software components are interconnected. The proposed concept is schematically illustrated in Figure 3.1 with the aforementioned connections between components.

Denoting an automotive application as AA, the set of application scenarios as AS, the set of operational scenarios as OS, the set of cameras as C, the set of illumination sources as IS and the set of computer vision algorithms as CV, an AA can be formally described by a tupel AS, OS, C, IS, CV where each of the addressed sets contains one or more members.

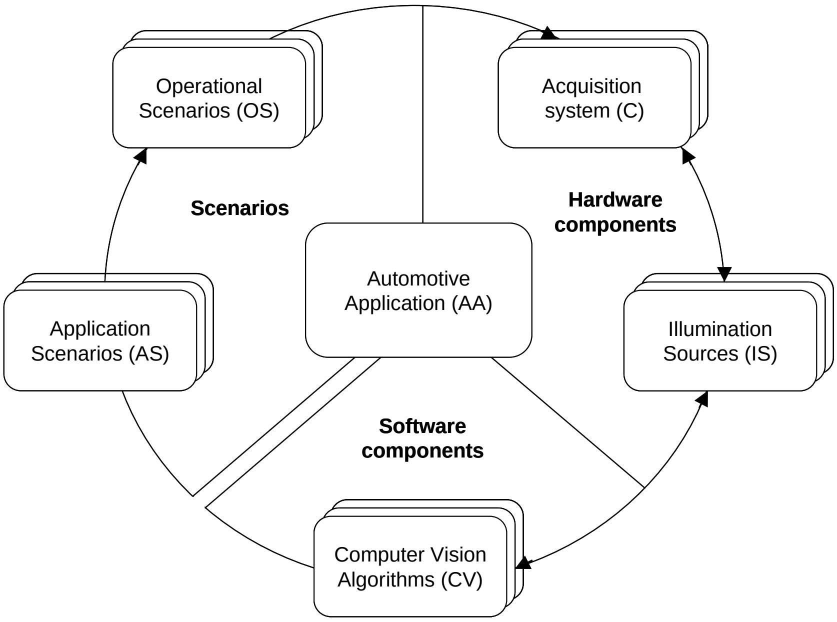

Operational Scenarios (OS)
Acquisition system (C)
Scenarios
Hardware components
Automotive Application (AA)
Application Scenarios (AS)
Illumination Sources (IS)
Software components
Computer Vision Algorithms (CV)

Figure 3.1: Schematic representation of an application and its components

The design process can be formalized as follows:

1. Discover a new application benefiting from vision-based in-car observation (e.g. biometric memory function) and define the application scenarios (e.g. facial driver recognition for the activation of the memory function).
2. For each application scenario define an operational scenario (e.g. single driver identification direct after getting into a car) and derive requirements to hardware components (e.g. small low-cost NIR-sensitive camera frontally facing driver's face which is supported by active NIR light sources).
3. Choose and install appropriate hardware components namely choose camera type and position, position of light sources, type of DSP board etc., and derive restrictions to software components (e.g. refuse color-based skin segmentation or get rid of machine learning approaches).
4. Select/design and implement software solutions and evaluate their recognition performance regarding the proposed operational scenarios.

The brief description of all three vision-based in-car automotive applications proposed in this thesis is introduced in Table 3.1. The table summarizes application components in accordance with the introduced designing methodology. The detailed description of each single component is given in Chapter 4 which is devoted to practical designing of applications.

Computer vision approaches applied in the proposed automotive applications include several common image processing and pattern recognition algorithms. These algorithms are assembled in Table 3.2. In order to understand how the computer vision approaches operate, the mathematic foundation of algorithms listed in the table is given in the next section.

Table 3.1: Components of the seat occupancy detection (SOD) system, user discrimination system (UDiS) and facial driver recognition (FDR) system

	SOD	UDiS	FDR
AS	AS$_1$: Support smart airbag to prevent the airbag deployment at unoccupied seats	AS$_1$: Support concurrent usage of new HMI concepts incl. dual-view touch screen and MMI knob	AS$_1$: Biometric memory function AS$_2$: Mandatory driver assistance AS$_3$: Biometric anti-theft protection
OS	OS$_1$: For each camera frame and each seat to decide whether the seat is occupied	OS$_1$: Frame-based determination of the action originator OS$_2$: Action-based determination of the action originator	OS$_1$: Single driver identification (SDI) supporting AS$_1$ and AS$_2$ (SDI-sec. with manual enrollment, SDI-conv. with semi-automatic enrollment) OS$_2$: Permanent driver verification (PDV) supporting AS$_3$ (PDV-sec. with manual enrollment, PDV-conv. with semi-automatic enrollment)
C	C$_1$: Omni-directional camera + DSP board	C$_1$: Omni-directional camera + DSP board C$_2$: Industrial camera facing center panel	C$_1$: Low-cost CCTV camera facing driver through steering wheel aperture + CarPC
IS	IS$_1$: Five NIR modules uniformly distributed in passenger compartment	IS$_1$: Five NIR modules uniformly distributed in passenger compartment IS$_2$: 940 nm LED-Lamp IS$_3$: 880 nm LED-Lamp + 940 nm LED-Lamp	IS$_1$: NIR LED ring around camera lens
CV	CV$_1$: Template matching CV$_2$: Face detection	CV$_1$: Motion-based approach (movement in driver/ passenger regions) CV$_2$: Shape-based approach (arm orientation) CV$_3$: Hand detection (hand position-/trajectory-based decision making)	CV$_1$: Appearance-based face recognition CV$_2$: Feature-based face recognition

Table 3.2: Image processing and pattern recognition algorithms utilized for seat occupancy detection (SOD), user discrimination (UDiS) and facial driver recognition (FDR)

	Pre-processing	Feature extraction	Feature selection	Matching
SOD_CV$_1$ (Template matching)	- Local normalization - Edge detection (Canny operator)	-	-	- Cross-correlation
SOD_CV$_2$ (Face detection)	-	- SMQT - Haar-like features	-	- SNoW classifier - AdaBoost - SVM
UDiS_CV$_1$	- Smart image differencing - Binarization - Morphological operations	-	-	- Rules-based classification
UDiS_CV$_2$	- Edge detection (Sobel operator) - Morphological operations	-	-	- Rules-based classification
UDiS_CV$_3$ (Hand localization)	-	- Haar-like features	-	- AdaBoost
FDR_CV$_1$	- Face localization - Best-fit plane subtraction - Histogram equalization	- Principal component analysis (eigenfaces)	- Information amount criterion	- NN - *k*-NN - Simplified Fuzzy ARTMAP
FDR_CV$_2$	Proprietary (no information)			

3.2 Exemplary image processing algorithms as building components of automotive applications

This section summarizes image processing algorithms which are used later as the construction elements for the three proposed automotive imaging applications: seat occupancy detection (SOD), user discrimination (UDiS), and facial driver recognition (FDR). All stages of the pattern recognition pipeline form pre-processing, to feature extraction and selection, to classification are addressed. In particular, the pre-processing subsection includes the descriptions of global and local approaches to image processing, thresholding, edge detection, local normalization, best-fit plane subtraction, histogram equalization and morphology. The feature extraction subsection includes descriptions of Haar-like features and successive mean quantization transform (SMQT) features. The feature selection subsection comprises the description of principal component analysis (PCA). The last subsection is devoted to classification approaches and includes the description of five classifiers: k-NN, ARTMAP, SNoW, SVM and AdaBoost.

3.2.1 Pre-processing

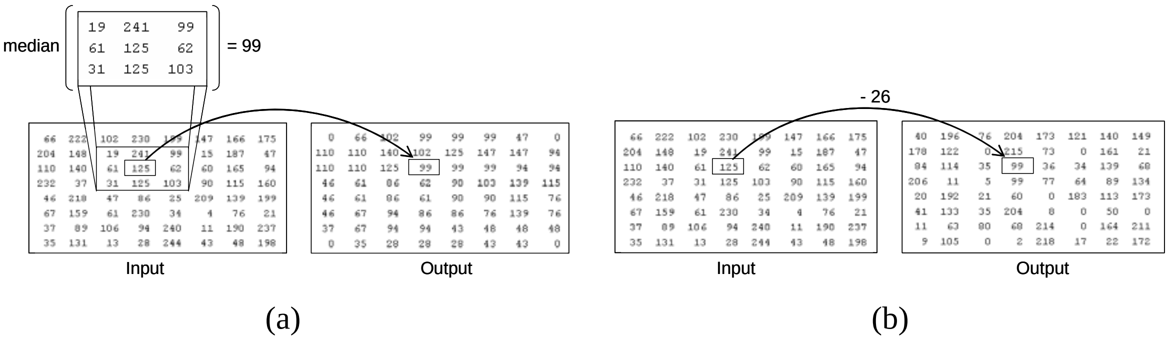
Pre-processing aims at enhancing the quality of an input image in the way that the object of interest depicted in the image is emphasized. An input as well as an output of the pre-processing module is an image. An image is a matrix of pixel intensities where the intensities take integer values usually from 0 to 255 (for a standard 8-bit format). In color RGB images a pixel is represented by three values: intensity of red, intensity of green and intensity of blue colors expanding an image to three matrices. However, the color information is often sacrificed for the sake of processing speed by converting colors to gray-values. This is a common practice especially for 24/7 surveillance applications where the same processing is applied to day and night images. Formally, an image I is defined by a two-dimensional function $I(x, y)$ where x and y are spatial coordinates. The amplitude of I at any pair (x, y) is an intensity value at this point. Assuming a color image is given by three functions $R(x, y)$, $G(x, y)$ and $B(x, y)$ then the standard transformation to gray-scale image $I(x, y)$ is given by (3.1).

$$I(x, y) = 0.299 \cdot R(x, y) + 0.587 \cdot G(x, y) + 0.114 \cdot B(x, y) \tag{3.1}$$

Hereafter, I shrink to gray-scale images describing image processing approaches.

Global vs. local processing

Local processing methods use a small neighborhood of a pixel in an input image to get a new intensity value in the output image. This operation is also called spatial filtering and applied on all pixels of the original image except for boundary pixels. In order to process boundary pixels the original image has to be enlarged to half of the neighborhood size from all four sides. The neighborhood is often selected as a rectangular shape with an odd number of rows and columns enabling the specification of the central pixel of the neighborhood, so-called origin. In filtering theory this rectangular region is pixel-wise multiplied to the convolution kernel of the same size and the sum of the resulting intensities is returned as the filtering outcome. In general, the neighborhood can have arbitrary shape. In mathematical morphology applied for image analysis this neighborhood is referred to as a structural element.



19	241	99
61	125	62
31	125	103

66	222	102	230	199	147	166	175
204	148	19	241	99	15	187	47
110	140	61	125	62	60	165	94
232	37	31	125	103	90	115	160
46	218	47	86	25	209	139	199
67	159	61	230	34	4	76	21
37	89	106	94	240	11	190	237
35	131	13	28	244	43	48	198

0	66	102	99	99	99	47	0
110	110	140	102	125	147	147	94
110	110	125	99	99	99	94	94
46	61	86	62	90	103	139	115
46	61	86	61	90	90	115	76
46	67	94	86	86	76	139	76
37	67	94	94	43	48	48	48
0	35	28	28	28	43	43	0

(a)

66	222	102	230	199	147	166	175
204	148	19	241	99	15	187	47
110	140	61	125	62	60	165	94
232	37	31	125	103	90	115	160
46	218	47	86	25	209	139	199
67	159	61	230	34	4	76	21
37	89	106	94	240	11	190	237
35	131	13	28	244	43	48	198

40	196	76	204	173	121	140	149
178	122	0	215	73	0	161	21
84	114	35	99	36	34	139	68
206	11	5	99	77	64	89	134
20	192	21	60	0	183	113	173
41	133	35	204	8	0	50	0
11	63	80	68	214	0	164	211
9	105	0	2	218	17	22	172

(b)

Figure 3.2: (a) median filtering is an example of local processing, (b) brightness reduction is an example of global processing

In global approaches, in contrast, modifications are provided to the whole image so that any input pixel can affect a large number of output pixels [247]. For instance, histogram processing is assigned to global processing because modifications applied to histograms are reflected in simultaneous changing of all pixels. Examples of local and global processing are shown in Figure 3.2.

Thresholding

It is much easier to find some formal descriptors of an object of interest in binary images, where object pixels have one value and background pixels another. The process of transformation of a gray-scale image to a binary image is referred to as thresholding or binarization. Pixels of a gray-scale image I are converted to pixels of a binary image I_{bin} based on the predefined threshold τ. All pixels with an intensity greater than or equal to the threshold are replaced with 1 (white) and all other pixels with 0 (black), as in (3.2).

$$I_{bin}(i,j) = \begin{cases} 1 & if \quad I(i,j) \geq \tau \\ 0 & if \quad I(i,j) < \tau \end{cases} \tag{3.2}$$

The selection of an appropriate threshold is the most challenging part of this process. There are two approaches to this end: using global threshold throughout the image and choosing adaptive threshold locally for each pixel with respect to the content of the pixel's neighborhood.

The most common algorithm for determining global threshold is proposed by Otsu [187]. The threshold is chosen to minimize the intra-class variance of black and white pixels. The algorithm works well then the histogram of an image has two well-defined peaks, one for foreground pixels and another for background pixels. Otherwise the significant amount of pixels can be misclassified.

The adaptive threshold relies on the assumption that small image regions are likely to be uniformly illuminated and pixels within local regions are therefore uniformly distributed. Hence, the mean value or the median can be chosen as a local threshold. Another standard technique for dynamic thresholding is introduced by Chow and Kaneko [40]. An image is divided into overlapping blocks and local histograms are investigated to determine locally optimal thresholds. The blocks with uniformly distributed pixels are ignored. Finally, the threshold at each pixel is the result of the interpolation of thresholds in blocks containing this pixel. The Chow-Kaneko algorithm has the drawback of being computationally expensive and is, therefore, hardly acceptable for real-time applications. In fact, there is a vast number of algorithms for selecting an optimal threshold operating globally as well as locally. A good overview of such algorithms is introduced in [89].

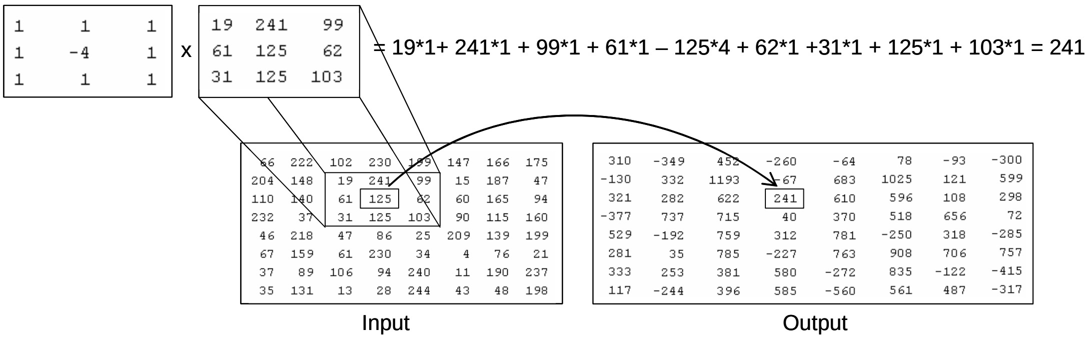
Edge detection

Formally, edge detection is another way of converting gray-scale images to binary images by determining edges of objects. Edge pixels are those pixels where the intensity changes abruptly. Intensity changes at a particular point can be calculated using partial derivatives at that point with respect to vertical and horizontal directions. The direction of the largest growth of the image intensity at some point is given by a gradient vector which is the vector of partial derivatives at that point. The steepness of intensity changing is described by the magnitude of the gradient while the edge direction is described by the angle of the gradient. In some cases using second-order partial derivatives is more advantageous. The magnitude of the second-order derivatives is given by the Laplace operator representing the divergence of the gradient. While for gradient's magnitude the maximum values indicate edge pixels, for the Laplace operator the zero-crossings indicate edge pixels.

In practice, partial derivatives are approximated by finite differences. Hence, the gradient is considered a local operator calculated in a local neighborhood of a pixel. This is why gradient operators can be expressed by a collection of convolution kernels. First-order derivatives can be approximated by convolution kernels of 2x2 pixels. For second-order derivatives, convolution kernels of at least 3x3 pixels are required. In image processing, the Laplace operator is approximated by the Laplace filter:

$$\Delta = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \text{ for 4-neighborhood and } \Delta = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -4 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ for 8-neighborhood.}$$

For detecting edges the Laplace operator is used in combination with Gaussian smoothing and referred to as Laplacian of Gaussian (LoG) [95]. An example of applying the Laplace filter is visualized in Figure 3.3.



1	1	1
1	-4	1
1	1	1

x

19	241	99
61	125	62
31	125	103

= 19*1+ 241*1 + 99*1 + 61*1 – 125*4 + 62*1 +31*1 + 125*1 + 103*1 = 241

66	222	102	230	199	147	166	175
204	148	19	241	99	15	187	47
110	140	61	125	62	60	165	94
232	37	31	125	103	90	115	160
46	218	47	86	25	209	139	199
67	159	61	230	34	4	76	21
37	89	106	94	240	11	190	237
35	131	13	28	244	43	48	198

Input

310	-349	452	-260	-64	78	-93	-300
-130	332	1193	-67	683	1025	121	599
321	282	622	241	610	596	108	298
-377	737	715	40	370	518	656	72
529	-192	759	312	781	-250	318	-285
281	35	785	-227	763	908	706	757
333	253	381	580	-272	835	-122	-415
117	-244	396	585	-560	561	487	-317

Output

Figure 3.3: Convolution of an image with the 8-neighborhood Laplace filter

The Sobel operator [218] is an example of an approximation of first-order derivatives with a smoothing. There are two convolution kernels utilized:

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \text{ and } K_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

The first one detects edges in the horizontal and the second one in the vertical direction. Convolution of an image I with kernels K_x and K_y results in two images G_x and G_y representing the derivatives in corresponding directions. The edge strength at a point (i, j) is reflected in a gradient's magnitude $G(i, j) = \sqrt{G_x(i, j)^2 + G_y(i, j)^2}$ and the edge direction is given by a gradient's angle $\Theta = arctan\,(G_y(i, j)/G_x(i, j))$.

The Canny algorithm [30] has been developed as the optimal edge detector. Three optimality criteria are addressed: optimal detection, optimal localization and minimal response.

- optimal detection - there are no spurious responses and all important edges are detected;
- optimal localization - the located edges are minimally deviated from the actual edges;
- minimal response - if multiple responses to a single edge occur, only one is considered a true response and the remaining ones are ignored.

Canny's edge detector consists of five stages:

1. *Image smoothing.* The original image is convolved with the two-dimensional Gaussian filter to smooth an image and reduce noise.

2. *Differentiation.* Two gradient images G_x and G_y are created by convolving the smoothed image with first-order derivatives of the one-dimensional Gaussian in x and y directions. The gradient magnitude G and gradient orientation Θ are calculated from gradient images G_x and G_y equally to that of the Sobel operator. The orientations are quantized so that they become one of four orientations: vertical, horizontal and the two diagonals.

3. *Non-maximum suppression.* Edge points are those, where the gradient magnitude G locally obtains maximum. In order to better localize edges, the maximum magnitudes along the edge direction are emphasized and the non-maximum magnitudes are suppressed. Therefore, for

each of four quantized directions the edge pixels are checked to reach the local maximum in the direction which is perpendicular to the considered one. For instance, the edge direction is considered vertical if the pixel's magnitude is greater than magnitudes of adjacent pixels from left and right.

4. *Edge thresholding.* In order to eliminate spurious responses, the threshold-based processing with hysteresis is applied. Two thresholds high and low are utilized to detect continuous curves. Firstly, the high threshold is applied to determine evident edges. The non-evident edges are detected by tracing strong edges and utilizing directional information. The directional magnitude needs to exceed the low threshold discovering faint sections of edges.

5. *Feature synthesis.* Final edges are generated by aggregating edges from synthesized images at multiple scales. Synthesized edge images result from modifying standard deviations of the Gaussian filter. In fact, there may be several scales of operators that give significant responses to edges. The best localization of the edge, however, is obtained using the operator with the smallest scale while larger scales tend to disclose shadow and shading edges, or edges between textured regions [30].

Edge detectors are compared in Figure 3.4.

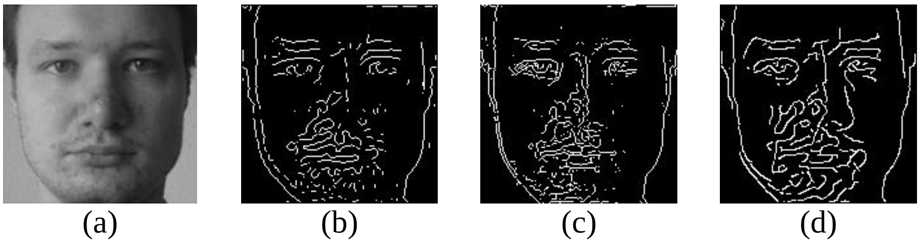
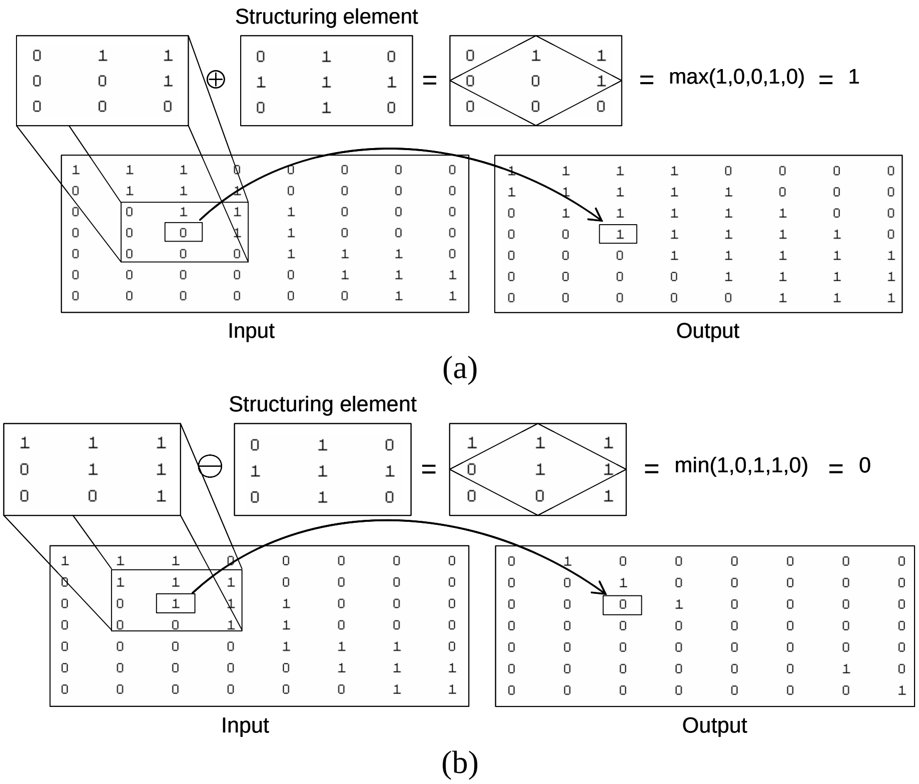


Figure 3.4: Edge detection: (a) original image; (b) after LoG; (c) after Sobel; (d) after Canny

Morphology

Binary silhouettes of objects derived from thresholding or edge detection include noise and therefore require post-processing. To this end, mathematical morphological operations are adopted for image processing to improve shapes in binary images. Morphology implies local image processing where each pixel of an image is modified with respect to its neighbors within the structuring element regarding the applied morphological operation. Size and shape of the structuring element can be arbitrarily selected and determines the sensitivity to specific shapes. Formally, a structuring element is a matrix containing 0 or 1 where the pixels with values of 1 define the neighborhood. The center pixel of the structuring element is called the origin. It identifies the pixel which is currently processed.

Dilation and erosion are two basic morphological operations. Dilation adds pixels to the boundaries of objects while erosion removes pixels on boundaries. The number of pixels added or removed depends on the size and shape of the structuring element (see Figure 3.5). Dilation and erosion are often used in combination to implement morphological opening or closing. Opening is erosion followed by dilation and closing is dilation followed by erosion. The same structuring element is used for both operations. Opening removes small objects from an image while preserving the shape and size of large objects. Closing fills gaps inside objects and connects closely located objects by smoothing their outer edges. The effect of these operations is demonstrated in Figure 3.6.



0	1	1
0	0	1
0	0	0

0	1	0
1	1	1
0	1	0

0	1	1
0	0	1
0	0	0

Input

1	1	1	0	0	0	0	0
0	1	1	1	0	0	0	0
0	0	1	1	1	0	0	0
0	0	0	1	1	0	0	0
0	0	0	0	1	1	1	0
0	0	0	0	0	1	1	1
0	0	0	0	0	0	1	1

Output

1	1	1	1	0	0	0	0
1	1	1	1	1	0	0	0
0	1	1	1	1	1	0	0
0	0	1	1	1	1	1	0
0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	0	1	1	1

(a)

1	1	1
0	1	1
0	0	1

0	1	0
1	1	1
0	1	0

1	1	1
0	1	1
0	0	1

Input

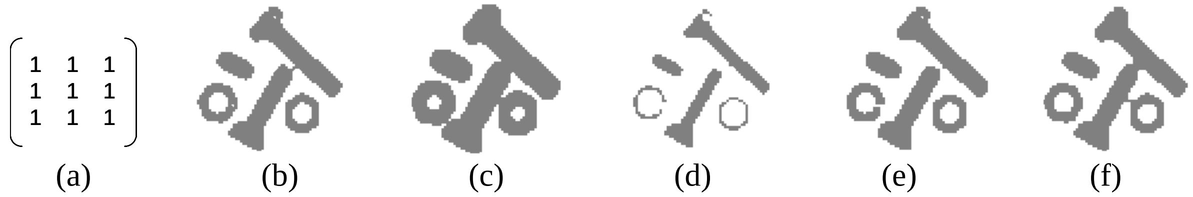
1	1	1	0	0	0	0	0
0	1	1	1	0	0	0	0
0	0	1	1	1	0	0	0
0	0	0	1	1	0	0	0
0	0	0	0	1	1	1	0
0	0	0	0	0	1	1	1
0	0	0	0	0	0	1	1

Output

0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1

(b)

Figure 3.5: Morphological dilation (a) and erosion (b) with diamond-shaped structuring element



$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

(a) (b) (c) (d) (e) (f)

Figure 3.6: (a) Structuring element, (b) original image, (c) after dilation, (d) after erosion, (e) after opening, (f) after closing

Informally, any operation on a binary silhouette can be considered a morphological operation e.g. filling holes, removing small shapes, connect shapes, shrinking and making skeletons from shapes. For more operations and their detailed descriptions read digital image processing book of Pratt [200] or Gonzalez [88].

Local normalization

Local normalization makes the local mean and variance of an image uniform [62]. This is especially useful for correcting non-uniform illumination or shading artifacts. The local normalization is computed for each image block using (3.3), where $I(x, y)$ is the original image block, $\mu_I(x, y)$ is an estimation of a local mean of $I(x, y)$, $\sigma_I(x, y)$ is an estimation of the local standard deviation, and $I_{norm}(x, y)$ is the output image block.

$$I_{norm}(x,y) = \frac{I(x,y) - \mu_I(x,y)}{\sigma_I(x,y)} \tag{3.3}$$

The estimation of the local mean and standard deviation is performed through spatial smoothing. The parameters of the algorithm are the sizes of the smoothing windows s_1 and s_2 which control the estimation of the local mean and local variance, respectively. Figure 3.7 demonstrates the effect of local normalization.

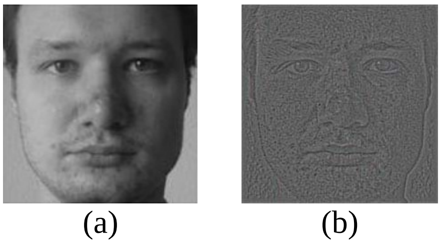


Figure 3.7: Local normalization: (a) original image, (b) after normalization

Best-fit plane subtraction

In case an image is illuminated from one side, the opposite side becomes dark due to shading. Generally, the digital illumination correction is a very complicated task requiring the knowledge of the object geometry, object reflection and positions of illumination sources. Possessing an image, theoretically, the 3D shape of an object can be reconstructed from shading [257, 128] and uniform lighting can be modeled. Practically, for accurate reconstruction the general 3D model of the object is required [21]. Subtraction of the best-fit plane (BFP) is a simplified illumination correction. The best-fit plane is a plane that optimally fits a set of N data points (x_i, y_i, z_i) in the sense of minimum least-squares of deviations. Image intensities represent a special case where measurements are made over a rectangular grid.

The standard equation of a plane is $Ax+By+Cz = D$ which can be written as $A/Cx+B/Cy+z = D/C$. Renaming constants A/C to a, B/C to b and D/C to c, the plane can be described as:

$$z(x,y;a,b,c) = c - ax - by \tag{3.4}$$

Minimizing the sum of the squares of the vertical distances between the data and the plane J with respect to the parameters a, b, and c the best-fit plain can be obtained.

$$J(x_i,y_i;a,b,c) = \sum_{i=1}^{N} (z_i - z(x_i,y_i;a,b,c))^2 = \sum_{i=1}^{N} (z_i - (c - ax_i - by_i))^2 \tag{3.5}$$

Minimum occurs at a point where all three partial derivatives yield zero: $\partial J/\partial a = 0, \partial J/\partial b = 0$and $\partial J/\partial c = 0$. The system of these equations has only one solution and can be solved analytically as shown in [54]. Assume that the underlying grid has m nodes in x-dimension and n nodes in the y-dimension then the function $z(x,y)$ has the size (m,n). Denoting x-points as $\hat{x}_i$, y-points as $\hat{y}_j$ and z-points as $\hat{z}_{ij}$, the parameters a, b and c are given by equations (3.6), (3.7) and (3.8).

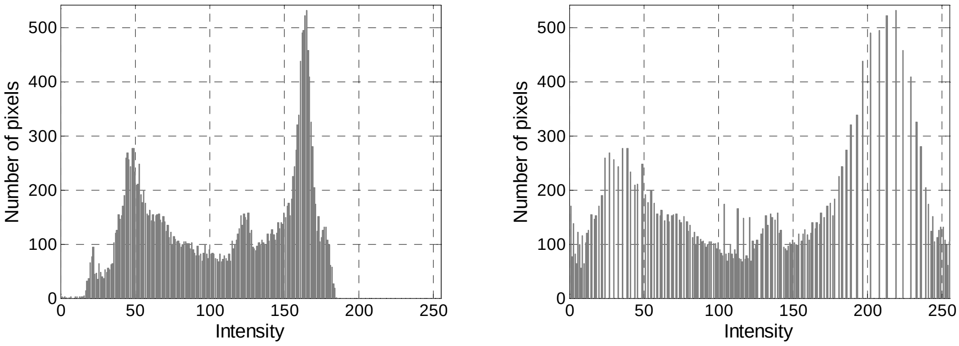
$$a = -\frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\hat{x}_i\hat{z}_{ij} - \frac{1}{m}\left[\sum_{i=0}^{m-1}\hat{x}_i\right]\left[\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\hat{z}_{ij}\right]}{n\sum_{i=0}^{m-1}\hat{x}_i^2 - \frac{m}{n}\left[\sum_{i=0}^{m-1}\hat{x}_i\right]} \tag{3.6}$$

$$b = -\frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\hat{y}_j\hat{z}_{ij} - \frac{1}{n}\left[\sum_{j=0}^{n-1}\hat{y}_j\right]\left[\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\hat{z}_{ij}\right]}{m\sum_{j=0}^{n-1}\hat{y}_j^2 - \frac{n}{m}\left[\sum_{j=0}^{n-1}\hat{y}_j\right]} \tag{3.7}$$

$$c = z_{avg} + ax_{avg} + by_{avg} \tag{3.8}$$

Histogram equalization

Another standard technique for normalizing illumination is spreading out the most frequent intensity values in the histogram (see Figure 3.8) referred to as histogram equalization [87].

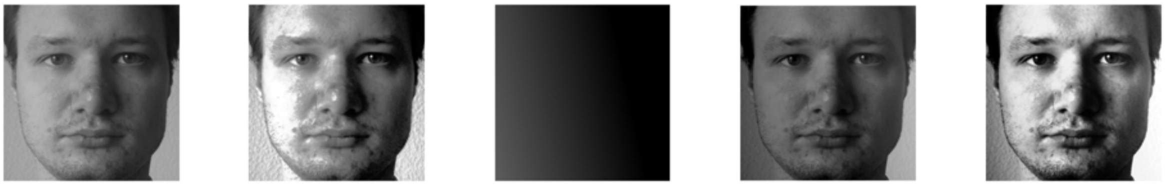


(a)

(b)

Figure 3.8: (a) histogram before and (b) after equalization; the corresponding images can be seen in Figures 3.9a and 3.9b

In order to equalize the histogram the pixel intensities are replaced by normalized y-values form the cumulative histogram located at x-position which corresponds to the pixel intensity. Applying histogram equalization on raw images often leads to derating of pixels with marginal intensities. Bright pixels become 255 and dark pixels become 0 destroying details in the image. Therefore, histogram equalization should follow the best-fit plane subtraction.



(a) (b) (c) (d) (e)

Figure 3.9: Effect of histogram equalization: (a) original image, (b) after histogram equalization, (c) best-fit plane (BFP), (d) after BFP subtraction, (e) after BFP subtraction and subsequent histogram equalization

3.2.2 Feature extraction

The computer vision approaches utilized in automotive applications proposed in this thesis include the feature extraction stage only as a part of detection algorithms namely for detection of faces and hands. Two different kinds of features (Haar-like features, SMQT) are applied in face detectors for seat occupancy detection. For user discrimination, Haar-like features are engaged for hand detection. Facial driver recognition includes face detection which makes use of Haar-like features.

Haar-like features

Haar-like feature is an invention of Viola and Jones [242], who proposed considering intensity differences between parts of local blocks as discriminative characteristics. A single Haar-feature is almost useless for distinguishing an object of interest from the background. In any case, the number of Haar-like features which can be extracted from an image is huge and the combination of features can reliably indicate an object. Features are named after the Haar wavelet [98, 41] and the name reflects well the intuition behind it. The Haar wavelet is the simplest possible wavelet comprised of a sequence of square-shaped functions of different scales establishing an orthonormal basis for the space of square-integrable functions on the unit interval [0, 1]. This means that any function can be approximated by a linear combination of Haar functions. Establishing Haar-like features is an attempt to adapt Haar functions for image processing. A Haar-like feature considers adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image.

Originally five feature groups were proposed for an image block:

- Difference between left and right parts (edge features)
- Difference between top and bottom parts (edge features)
- Difference between the sum of left and right parts and the central part (line features)
- Difference between the sum of top and bottom parts and the central part (line features)
- Difference between the central part and the circumjacent area (center-surround features)

Lienhart and Maydt [147] extended the feature set by 45-degree tilted features resulting in ten feature groups. Tilted features enable better detection of tilted edges leading to the significant improvement of original features. Figure 3.10 gives an impression of the compilation of features.

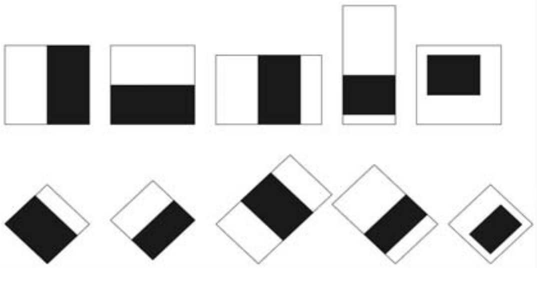


Figure 3.10: Original and tilted Haar-like features, the sum of pixel intensities within white rectangles are subtracted from the sum of pixel intensities within black rectangles

Features can vary in scale and proportion of left/right, top/bottom and center/circumjacent areas forming a huge feature set even for small image blocks. For instance, from a block of 24x24 pixels approximately 120,000 features can be extracted [147].

Haar-like features can be very efficiently computed from the integral image. The integral image (also referred to as summarized area table or SAT) is an image where each pixel presents the sum of intensities of the pixels located in the upper left position regarding the current pixel in the original image. For computation of edge features six SAT values are required, for line features eight SAT values and for center-surround features nine SAT values.

With I denoting original image, the integral image is given by $SAT(x,y) = \sum_{i \leq x, j \leq y} I(i,j)$. The intensity sum in a rectangular image block r of size (w,h) at the location (x,y) is then given by (3.9):

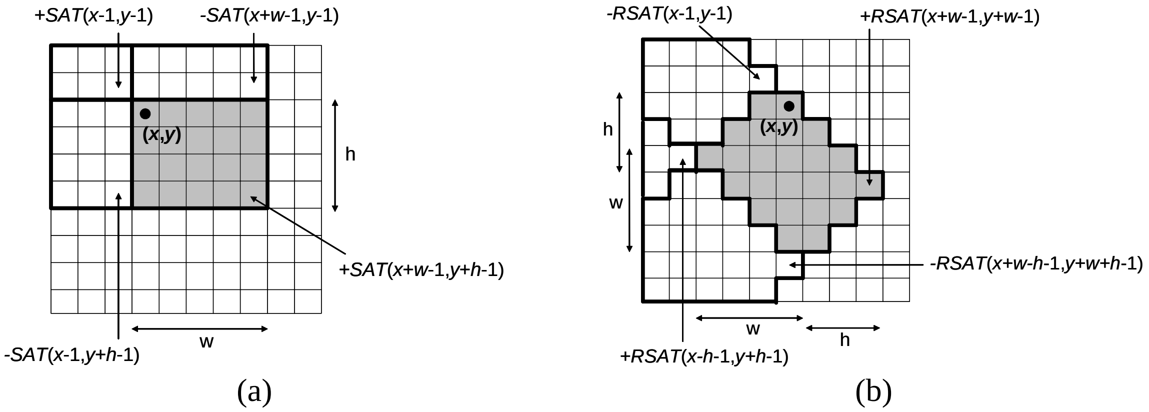
$$Sum(r) = SAT(x-1,y-1) + SAT(x+w-1,y+h-1) - SAT(x-1,y+h-1) - SAT(x+w-1,y-1) \tag{3.9}$$

with $SAT(-1,y) = SAT(x,-1) = 0$.

For computation of 45-degree tilted Haar-like features the definition of the integral image is extended by the rotated summarized area table $RSAT$ which gives the sum of intensities of the pixels located in the left top corner regarding the current pixel in the original image: $RSAT(x,y) = \sum_{i \leq x, |y-j| \leq x-i} I(i,j)$. $RSAT$ can be calculated by passing an image from left to right and from top to bottom calculating $RSAT(x,y) = RSAT(x-1,y-1) + RSAT(x-1,y) + I(x,y) - RSAT(x-2,y-1)$ with $RSAT(-1,y) = RSAT(-2,y) = RSAT(x,-1) = 0$ and afterwards by passing an image from right to left and from bottom to top calculating: $RSAT(x,y) = RSAT(x,y) + RSAT(x-1,y+1) - RSAT(x-2,y)$. Based on the resulting integral image the intensity sum in a 45-degree tilted rectangular block r of size (w,h) at the location (x,y) can be calculated using (3.10):

$$Sum(r) = RSAT(x+w,y+w) + RSAT(x-h,y+h) - RSAT(x,y) - RSAT(x+w-h,y+w+h) \tag{3.10}$$

Figure 3.11 visualizes the computation of rectangular sums in standard and rotated integral images.



(a)

(b)

Figure 3.11: Computation of rectangular sums in the (a) standard integral image and (b) 45-degree rotated integral image (modified from [147])

Since Haar-like features are differences of rectangular sums, from (3.9) and (3.10) it becomes obvious that all features are computed in a constant time independently of the block size. The set of Haar-like features is a very powerful instrument for describing image patterns which can be hardly formalized. The crucial advantage of Haar-like features lies not in high discriminative power of individual features, but in the fact that a huge set of such simple features can be very rapidly computed, ranked and combined to a cascade classifier. For the fast search of the object of interest in the image it is required that an “interesting” image region contains almost all highly ranked features which allows to immediately leave out of consideration large “non-interesting” image regions.

In application to face detections, it is observed that in most facial images the eyes region is darker than the cheeks region as well as that eye regions are darker than the region between eyes containing the nasal bridge [242]. The first case is well described by the edge feature representing two adjacent rectangles that cover eye and cheek regions while the second case is well described by the line feature with two sided rectangulars covering eyes and one central rectangular covering the nasal bridge (see Figure 3.12).

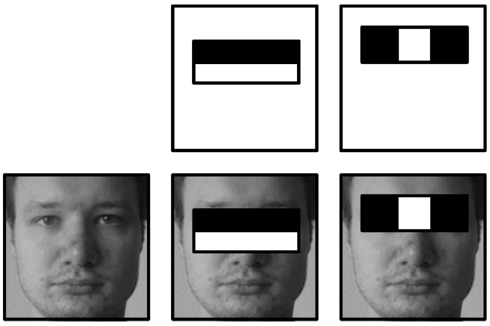


Figure 3.12: The first and second "best" Haar-like features. The first feature represents the difference in intensity between the eyes region and the region across the upper cheeks. The second feature represents the intensity difference between eye regions and the region across the nasal bridge

Successive Mean Quantization Transform

The Successive Mean Quantization Transform or shortly SMQT is an invention of Nilsson et al. [179]. This signal transformation aims at revealing the underlying structure of a digital signal by compensating disparities in sensor signals caused by gain and bias. In image processing, SMQT aims at representing an image insensitively to varying illumination. Two images with identical structures have the same SMQT representation. The same structure is understood in terms that one image can be generated from another by biasing or gaining it by a scalar value.

The SMQT features are extracted for each pixel of an image from a local area surrounding it, whereby the pixel intensities inside the area are replaced by the quantization values resulting from the successive comparison of pixel intensities with the mean intensity of the area. Theoretically, local areas can be of arbitrary shape. Traditionally, the sliding window approach is applied to split an image into square blocks of particular size. Alternatively, circular shapes can be used. According to Nilsson [179] the SMQT can be seen as a structural breakdown of intensity information or a progressive focus on the details in an image.

Formally, the $SMQT_L$ can be interpreted as a perfect binary tree of Mean Quantization Units (MQUs) disposed at L levels. Assuming x is a pixel and $D(x)$ is its surrounding local area containing N pixels. The intensity values in this local area $D(x)$ are denoted with $V(x)$. With μ denoting the mean intensity of the local area $D(x)$: $\mu = \mu(V(x)) = \frac{1}{N}\sum_{x' \in D(x)} V(x')$, the mean quantization unit Q of the local area $D(x)$ is defined as $Q : V(x) \rightarrow U(x)$ where $U(x_i) = 1$ if $V(x_i) > \mu$ and $U(x_i) = 0$ otherwise. Additionally, the MQU splits the local area $D(x)$ into two areas $D_1(x)$ and $D_0(x)$. The area $D_1(x)$ contains pixels whose intensities are higher than the mean intensity: $D_1(x) = \{x_i | V(x_i) > \mu, x_i \in D(x)\}$ and the area $D_0(x)$ contain pixels whose intensities are less than or equal to the mean intensity: $D_0(x) = \{x_i | V(x_i) \leq \mu, x_i \in D(x)\}$.

The MQUs are recursively applied on areas $D_0(x)$ and $D_1(x)$ generating binary units $U_{l,n}$ with $l, l = 1, 2, ..., L$ denoting the current level and $n, n = 1, 2, ..., 2^{l-1}$ denoting the running number of the binary units at the level l (see Figure 3.13).

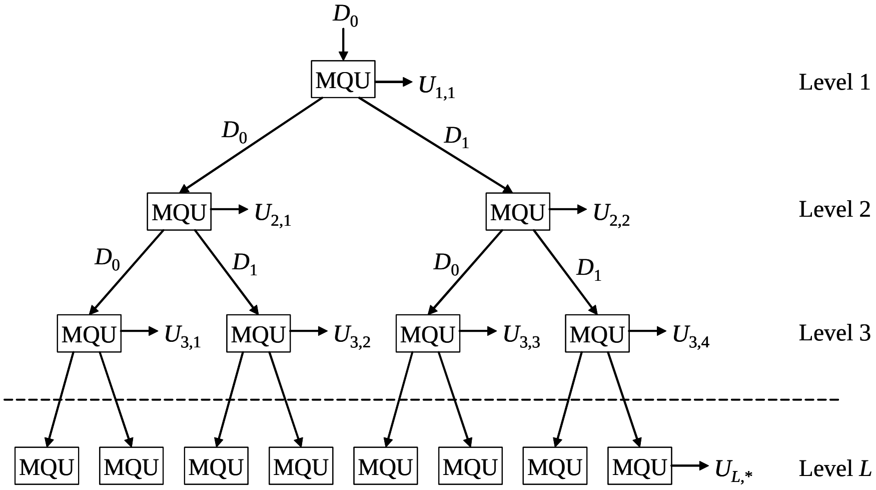


Figure 3.13: Visualization of $SMQT_L$ as a perfect binary tree of Mean Quantization Units (MQUs) disposed at L levels

Finally, the $SMQT_L$ is constructed as the superposition of binary units $U_{l,n}$ over all levels by assigning them the weight of 2^{L-l} at the level l. So, the output image of $SMQT_L$ is given by:

$$M(x) = \left\{ x | V(x) = \sum_{l=1}^{L} \sum_{n=1}^{2^{l-1}} U_{l,n} \cdot 2^{L-l} \right\}, \quad \forall x \in M \tag{3.11}$$

The tree level L also gives the number of bits in $SMQT_L$. Hence, $SMQT_1$ has a one bit representation - {0,1}, $SMQT_2$ has a two-bit representation - {0,1,2,3}, $SMQT_3$ a three-bit representation - 0,1,...,7 and so on. Therefore, the number of quantization values in $SMQT_L$ yields 2^L. The SQMT of an image with the level lower than the number of bits in the image can be seen as a compression of the dynamic range. For standard 8-bit images, the $SMQT_8$ provides an uncompressed image with enhanced details [179]. This is why the transform can be also considered a pre-processing or illumination enhancement technique.

Since the MQU is the main constructing element of the SMQT and the MQU is insensitive to gain and bias the complete SMQT is inductively reckoned to be insensitive to gain and bias [179] (see Figure 3.14).

The $SMQT_1$ is very similar to local binary patterns (LBP) with the only difference that the thresholding is provided based on the mean intensity and not on the intensity of the central pixel. For a local pattern (of e.g. 32x32 pixels) the feature vector can be constructed as a sequence of values generated by $SMQT_1$ applied to all possible subpatterns (of e.g. 4x4 pixels). Since the window of 32x32 pixels contains 29x29=841 subpatterns of 4x4 pixels, the resulting feature vector has 841 dimensions where the value of each dimension varies from 0 to 2^{4x4}. For $SMQT_L$ values vary from 0 to 2^{Lx4x4}.

3.2.3 Feature reduction

There are two fundamentally different approaches to feature reduction: (1) ranking features with the succeeding elimination of low-ranked ones and (2) the transformation of the feature set. While in the first approach the semantic of features remains unchanged, in the second approach the semantic is destroyed and the new features become non-trivial combinations of old ones.

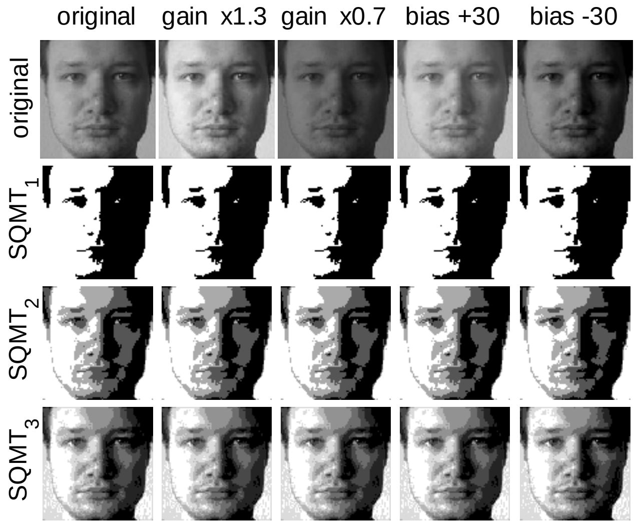


Figure 3.14: Three levels of SMQT applied on a face image of 50x50 pixels

For ranking features, the univariate analysis is usually implied meaning that features are evaluated individually and any correlation between them is disregarded. In order to take the correlation into account, the multivariate analysis is utilized. Here, one evaluates feature subsets and not individual features. Note that feature transformation approaches always signify multivariate analysis.

There are three strategies for ranking features. John and Kohavi [124] detached “filters” and “wrappers”. Filters evaluate features independently of a further applied classifier purely based on expert knowledge (so called heuristics).Wrappers evaluate features based on the outcome of the classifier. In recent studies e.g. in [214], filter and wrapper are complemented by the “embedded” strategy. Here, the feature selection is an inherent property of a classifier (e.g. decision trees or adaptive boosting).

The transformation of a feature set usually follows one of two following objectives: the transformed features do not correlate and are ranked by their variance (see principal component analysis [192]), or the transformed features do not correlate and are ranked by their discrimination power for particular groups of data (see linear discriminant analysis [72]). The second objective is very similar to the general objective of the classification task. In the following, only principal component analysis is described in detail, since it is the only feature reduction approach used in one of the proposed automotive applications namely in appearance-based face recognition for driver identification.

Principal component analysis

The main objective of principal component analysis (PCA) is reducing the data dimensionality without losing a significant amount of information. This is done by decorrelating dimensions through shifting and rotating basis axes. Note that the improved basis is generated based on the set of training samples. Examples of correlated dimensions are height and weight when describing a person, because tall persons are usually heavier than short persons. If semantics of dimensions are not important (e.g. for grouping persons regarding gender) the new dimension can represent an average of height and weight. Such a reduction from two to one dimension will not lead to significant loss of information representing the general body size of a person.

For a particular data set, PCA consecutively looks for linear combinations of original dimensions and selects one with the maximum variance of data points. Every next extracted dimension is orthogonal to all previous ones. The final number of extracted dimensions is least value of the number of data samples minus one and the number of original dimensions. Practically, the extraction of the new basis is done by solving the eigenvectors problem with the covariance matrix of original data shifted to the center of the coordinate system. The eigenvectors represent the new basis and the matrix with eigenvectors forming columns is the rotation matrix from the original to the desired coordinates. The importance of new dimensions is determined by the corresponding eigenvalues. The higher the eigenvalue is, the higher the variance of data points is and, therefore, the dimension encompasses the higher amount of information. The data transformation is provided by shifting data to the center of coordinate system and by rotating it through multiplication by the matrix of eigenvectors placed in columns. Starting from new coordinates, the dimensionality reduction is achieved by neglecting dimensions with low eigenvalues. PCA is also known as KLT transform named after Karhunen and Loeve. Note that in high-dimensional space vectors tend to scatter in few directions. PCA indicates these directions by principal eigenvectors. So, a few eigenvectors with highest eigenvalues preserve the vast majority of information and the remaining eigenvectors with eigenvalues approaching zero are not significant. Figure 3.15 shows PCA in action.

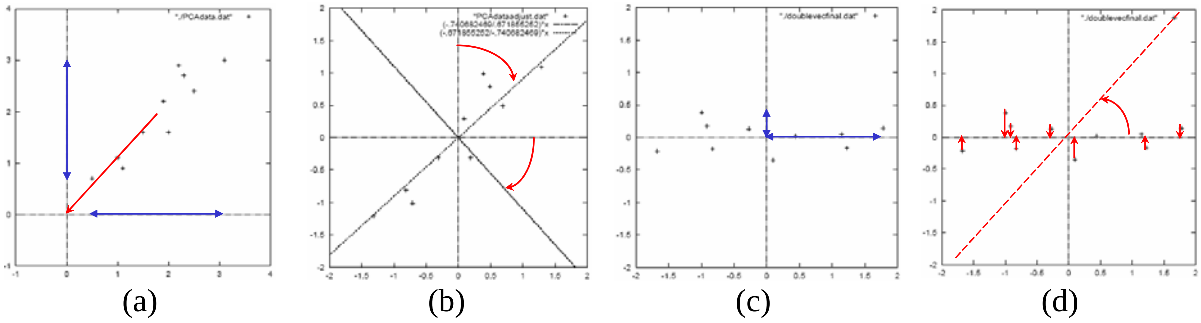


Figure 3.15: PCA: (a) shifts data points to the center of the coordinate system, (b) rotates the coordinate system to achieve maximal scatter in the principal axis, (c) evaluates significance of the resulting dimensions and (d) reduces insignificant dimensions (modified from [226])

In regard to appearance-based face recognition where an image is considered as a point in high-dimensional space and the set of all faces is assumed to form a low-dimensional manifold, the main idea is to analyze the distribution of face images in the whole image space and reduce the dimension of facial pattern by means of linear projection into more compact subspace. Applying PCA to this end has been established by Turk and Pentland [234] and called "eigenfaces". The formal description of PCA is exemplified by facial images.

Let us assume that faces in a training set are represented by gray-scale images of the size $w \times h$ pixels. Writing pixel intensities in a line and normalizing intensity values to the interval [0,1], a face is then represented by an N-dimensional vector ($N = w \times h$) of real values from 0 to 1. Let $\Gamma_1, \Gamma_2, \ldots, \Gamma_M$ be a collection of M training facial images written down as column vectors. The average face Ψ is then given by:

$$\Psi = \frac{1}{M} \sum_{i=1}^{M} \Gamma_i \tag{3.12}$$

In order to have the average face Ψ in the center of the coordinate system, let us introduce Φ_i be the difference between the face Γ_i and the average face Ψ and representing a face shifted to the center of the coordinate systems.

$$\Phi_i = \Gamma_i - \Psi \tag{3.13}$$

The covariance matrix Σ of shifted faces $\Phi_1, \Phi_2, \ldots, \Phi_M$ is defined as:

$$\Sigma = \frac{1}{M}\sum_{i=1}^{M} \Phi_i \Phi_i^T = AA^T, A = \ [\Phi_1, \Phi_2, ..., \Phi_M] \tag{3.14}$$

where A is a matrix of shifted faces located in columns. The covariance matrix is the square, symmetric, positive-semidefinite matrix. The eigenvalues λ_k and eigenvectors u_k of such matrix can be found from the equation: $\Sigma u_k = \lambda_k u_k$. Since the training set contains M images, the covariance matrix Σ can contain at most $M - 1$ meaningful eigenvectors. In a proper case the number of training images is significantly higher than the dimension of the image ($M \approx N^2$) so that the dimension of covariance matrix is $N \times N$ and it contains N meaningful eigenvectors. In practical applications, however, the opposite situation often arises, namely the number of training images is significantly lower than their dimension ($M < N$). In this case there is no need for solving the computationally intensive eigenvalues problem for $\Sigma = AA^T$ of the dimension $N \times N$, but the eigenvalues problem can be solved for the smaller matrix $\Sigma^* = A^T A$ of the dimension $M \times M$. The eigenvectors u_k of the matrix Σ then can be derived from the eigenvectrors v_k of the matrix Σ^* as $u_k = Av_k$, and the eigenvalues of Σ and Σ^* are equal. Let us demonstrate it in equations:

$$\Sigma^* v_k = \lambda_k v_k \Leftrightarrow A^T A v_k = \lambda_k v_k \tag{3.15}$$

After the left multiplication with A:

$$A(A^T A v_k) = A(\lambda_k v_k) \Leftrightarrow AA^T(Av_k) = \lambda_k(Av_k) \Leftrightarrow AA^T u_k = \lambda_k u_k \tag{3.16}$$

Returning to Φ_i, the u_k is given by:

$$u_k = \sum_{i=1}^{M} v_{ki} \Phi_i \tag{3.17}$$

where v_{ki} is the i^{th} component of the vector v_k. The eigenvectors u_k form a basis in the eigenspace. Turk and Pentland call them eigenfaces. The dimension of eigenspace and correspondingly the number of eigenfaces is the minimum value of N and $M - 1$. Eigenfaces with higher eigenvalues preserve rough structural information of a face contributing more significantly in the face appearance. In contrast, eigenfaces with lower eigenvalues contain only fine details and can be neglected without destroying general face appearance. Figure 3.16 shows the eigenfaces from 1 to 9 and from 42 to 50 for the training set of 3299 facial images as well as the diagram of eigenvalues.

With U denoting the matrix of eigenfaces in columns $U = [u_1, u_2, ..., u_k], k = 1, 2, ..., min(N, M - 1)$, the projection of a facial image Γ_i to the eigenspace is performed by multiplying the shifted version of it (Φ_i) by the transposed matrix of eigenfaces from the right. So, the resulting projection vector w_i is given by:

$$w_i = U^T \Phi_i = U^T(\Gamma_i - \Psi) \tag{3.18}$$

The dimension of the vector w_i is equal to the number of eigenfaces $w_i = [w_{i1}, w_{i2}, ..., w_{ik}]^T$. The back projection of a vector w_i to the original coordinate system is performed by multiplying it by the matrix of eigenfaces from the right. So, the shifted facial image Φ_i is represented by a linear combination of eigenfaces u_k and elements of the vector w_i:

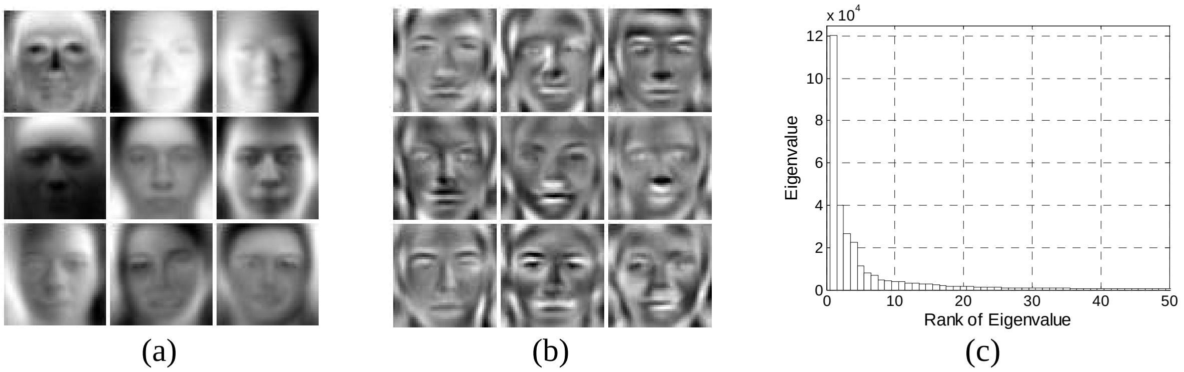


Figure 3.16: (a) First 9 eigenfaces $(u_1, ..., u_9)$, (b) eigenfaces from 42 to 50 $(u_{42}, ..., u_{50})$, (c) diagram of 50 highest eigenvalues $(\lambda_1, ..., \lambda_{50})$

$$\Gamma_i - \Psi = \Phi_i = Uw_i = \sum_{k=1}^{min(N,M-1)} w_{ik}u_k \tag{3.19}$$

This is why w_{ik} are often called coefficients of eigenfaces transformation. Figure 3.17 visualizes the eigenfaces representation for two facial images. The dimension reduction is performed by considering the limited set of transformation coefficients.

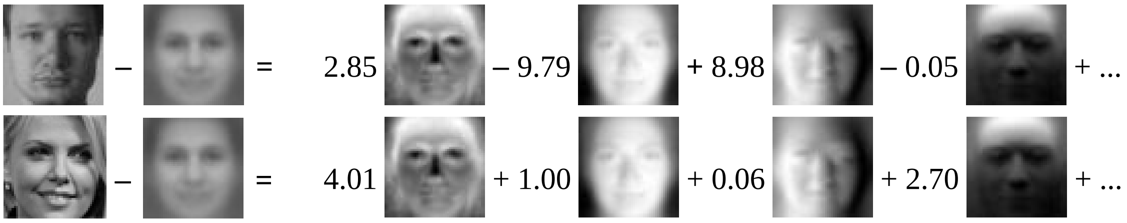


Figure 3.17: Two examples of eigenfaces decomposition of facial images[1]

3.3 Classification

The world of classification techniques is immense, so that the selection of one or another classifier for particular application is often a lottery. Some researchers are proud of one particular classification scheme and apply it everywhere. Other researchers compare all known classifiers for the addressed application under the constraint of the given dataset. So, they specify the "conditionally best" classifier and apply it to similar tasks. To the best of my knowledge, there is no standard benchmark for general grading of classification approaches. Duda et al. in their book [55] state that there is no universal classification technique, which is identically effective for all classification tasks and suggest deciding for one or another classifier only under the consideration of the particular application. Unfortunately, they do not give formal criteria for selection of classifier apart from common sense.

[1] The image in the second row is taken from: `http://en.wikipedia.org/wiki/File:Charlize_Theron_WonderCon_2012_(Straighten_Crop).jpg` and licensed under the Creative Commons Attribution-Share Alike 2.0 Generic license

Each of three automotive applications proposed in this thesis includes a matching stage. The word matching is intentionally used here instead of classification because often the word classification embraces the whole concept of statistical pattern recognition. Here, the classification means only the matching part of it. Formally, the classification is the matching of the test sample of an unknown object against reference models of known objects. In UDiS the classification of hands is based on rules manually derived from training samples. SOD makes use of template matching and three face detectors. Template matching utilizes a degraded version of classification, where one reference image (also referred to as template) is consecutively compared with image blocks of the same size. Each of the face detectors applies a sophisticated classifier. These are adaptive boosting (AdaBoost), sparse network of Winnows (SNoW) and support vector machine (SVM). FDR engages two classifiers: k-nearest neighbor (k-NN) algorithm and adaptive resonance theory mapping (ARTMAP) network. In the following theses five classification techniques are explained in detail.

3.3.1 Distance function

A distance function or metric is the central concept required for matching. This function reveals the degree of simmilarity/dissimilarity between two vectors. Formally, for a given set M a function $d : M \times M \to \mathbb{R}$ is a metric on the set M ($\mathbb{R}$ - the set of real numbers) if for all vectors a, b and $c \in M$ the following conditions are met:

1. $d(a, b) \geq 0$ (non-negativity),
2. $d(a, b) = 0$ if and only if $a = b$ (identity of indiscernibles),
3. $d(a, b) = d(b, a)$ (symmetry),
4. $d(a, b) + d(b, c) \geq d(a, c)$ (triangle inequality).

Note that non-negativity follows from the conditions 2-4: $2d(a, b) = d(a, b) + d(a, b) = d(a, b) + d(b, a) \geq d(a, a) = 0$. The pair (M, d) is referred to as a metric space.

The commonly used metrics are three Minkowski metrics (city-block distance, Euclidean distance and max-distance) [224], dot product [104], cosine similarity [225], Pearson correlation coefficient [71] and Mahalanobis distance [156].

In the n-dimensional metric space $M = \mathbb{R}^n$ with $a = (a_1, ..., a_n)$ and $b = (b_1, ..., b_n)$, the family of Minkowski metrics is given by (3.20). For city-block distance p=1, for Euclidean distance p=2 and for max-distance $p = \infty$.

$$d(a, b) = \sqrt[p]{\sum_{i=1}^{n} (a_i - b_i)^p} \tag{3.20}$$

Dot product expresses the angle between vectors a and b ignoring their lengths and is given by $a \cdot b = \sum_{i=1}^{n} a_i b_i$. Cosine similarity is the normalized version of the dot product which is given by $cos(a, b) = \frac{\sum_{i=1}^{n} a_i b_i}{||a||||b||}$ where $||\cdot||$ is the Euclidean norm (e.g. $||a||= \sqrt{\sum_{i=1}^{n} a_i^2}$).

In order to achieve translation, rotation and scale invariance, the classic distance function is substituted by a pseudo-, semi- and quasi-metric. In case of semi-metric, the triangle inequality is not required. Pseudo-metric does not require the identity of indiscernibles. For quasi-metric the symmetry is not required. Pseudo-, semi- and quasi-metrics can be arbitrarily combined. The best known semi-metric is the weighted Euclidean distance: $d(a, b) = \sqrt{(a - b)^T A(a - b)}$ where A is a quadratic positive-definite matrix. Note that here the dimensions have different contributions to the final score. If A is a diagonal matrix, then the diagonal elements define the importance of the corresponding dimensions. If A is the inverse covariance matrix derived from the set of training vectors, the metric is called Mahalanobis distance. If vector a is substituted by $a - \mu_a$ and vector b by $b - \mu_b$ where μ_a and μ_b are the mean vectors of the corresponding sets, and the cosine similarity is utilized, the resulting metric is called Pearson correlation coefficient.

3.3.2 k-Nearest neighbor algorithm

The k-nearest neighbor (k-NN) algorithm [44] is an example of a practically successful non-parametric classification technique, where the training degenerates to collecting of training samples also called prototypes. The decision to assign a test sample to one or another class is made by examining labels of k nearest prototypes and selecting the class containing the most number of prototypes. The simplest modification of k-NN is the nearest neighbor (1-NN) algorithm. Here, the nearest prototype determines the class of the test sample. The term "nearest" actually means the "most similar". The degree of similarity is given by a distance function as described above.

Since the simplest modification (1-NN) addresses only the best match, it is quite vulnerable against outliers. With the growing number of addressed neighbors, the k-NN becomes more robust against outliers. An important feature of the k-NN is that the outcome includes not only the class label of the test sample but also the matching score representing the distance/similarity between the test sample and the winner class. In case of 1-NN the matching score is the distance/similarity to the best match. In case of k-NN the matching score can be either the normalized number of matches within the winner class or the average distance/similarity to the matches within the winner class.

Using k-NN, all training patterns must be permanently stored. In large scale applications with a high number of training patterns for each object, the algorithm becomes expensive regarding memory storage. On the one hand, this feature is often considered the main drawback of k-NN. On the other hand, storing all training patterns enables independence from any assumptions about the general data distribution as well as individual class distributions. Moreover, the k-NN algorithm is flexible regarding the extension of class models meaning that class prototypes can be easily extended by additional patterns requiring no re-training and having no influence on the algorithm's stability.

1-NN has a nice geometrical interpretation, namely the decision space is formed by Voronoi cells [8]. Each cell is an attraction zone of the corresponding prototype (see Figure 3.18).

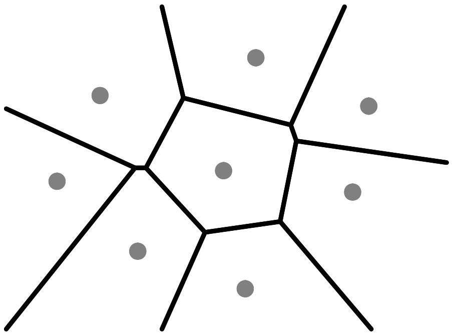


Figure 3.18: Voronoi cells

The nearest neighbor algorithm has very close relation to the Bayes classifier [51], which is considered optimal from the probabilistic point of view. Since Bayes is a parametric classifier, it requires knowledge about sample distributions within classes to extract the parameters of these distributions. Assuming, the true values of parameters are known, the Bayes classifier decides for a class having the maximal probability to contain the test sample. In this respect, the 1-NN is considered to be sub-optimal because its classification error is always higher than that of the Bayes classifier. However, as shown in [55], when the number of prototypes approaches infinity, the classification error of 1-NN is limited by twice the Bayesian error.

In k-NN, if the number of prototypes n approaches infinity and almost all prototypes contribute to decision making (or to be more precise k approaches infinity as well), k-NN becomes asymptotically very close to Bayes. This is formally shown in [55] and also proven that if the value n/k approaches infinity, the classification error of k-NN converges to the Bayesian error. Hence, being in possession of a high number of training samples well representing classes and using k-NN with relatively high k guarantees the low classification error which is close to the optimal one. Due to this reason and because of its simplicity and plausibility, k-NN is a favorable classifier in biometrics.

3.3.3 Neural networks and adaptive resonance theory mapping

Another technique, widely used for classification of biometric patterns, is a feed-forward neural network with the back-propagation learning e.g. multilayer perceptron (MLP) [103]. Eleyan et al. [61] show that MLP can outperform the nearest neighbor algorithm in the domain of face recognition. However, MLP has several limitations such as an absence of clear semantics of internal weights or the rule as to how many layers or neurons are required. The training of the network is done in the off-line mode implying that all training samples must be collected prior to the training. The cost function usually converges to a local minimum during back-propagation. Finally the network can be either in training or in classification mode (stability-plasticity dilemma). So, MLP does not have a mechanism to adaptively learn face appearances changing over time.

In an attempt to avoid the aforementioned limitations, Grossberg proposed in 1976 [91, 92] the adaptive resonance theory (ART) as a family of neuronal networks which possess an incrementally growing structure and provide stable on-line learning. The incrementally growing structure ensures that all patterns presented to the network will be learned and compactly stored. Stable on-line learning allows for adapting long time memory (LTM) weights of a network after each successful classification. These two characteristics lead to solving of stability-plasticity dilemma [32] and are very important for successful recognition of object patterns which are quite changeable over time. Since ART has been invented by biologists, the terminology applied for describing machine learning processes is very unusual and often abstruse. Sarle [215] gives interpretations of originally used terms to make the theoretical concept comprehensible for computer science practitioners. For instance, a cluster (or class) is addressed as “maximally compressed pattern recognition code” and LTM is nothing else than cluster seeds (or prototypes). Moore [175] explains that an ART network is in fact an iterative clustering consisting of two steps: being in possession of a training sample, first, to find the most similar cluster and, second, to update the cluster’s seed towards the training sample. In ART the first step is addressed as “attentional subsystem” and the second step as “orienting subsystem”.

The basic ART architecture is introduced for unsupervised learning (clustering). By connecting two ART networks using a mapping layer, the network gains the ability of unsupervised learning (classification) where inputs of one network are associated with inputs of another. The combined architecture is called ARTMAP for ART mapping. For pattern classification the simplified version of ARTMAP is more appropriate. Here, the second ART network does not provide clustering but directly receives class labels and delivers them to the mapping layer. The original ART1 network is restricted to binary data. Fuzzy ART [33] is an extension of ART1 to handle continuous data normalized to the interval [0,1]. The detailed description of the functionality is given for the simplified fuzzy ARTMAP (SFAM) network [125].

Formally, SFAM consists of two fully connected layers of nodes called F_1 and F_2, and a mapping layer F^{ab} connected to F_2 through learned associative links. F_1 is the input layer with $2M$ nodes where M is the dimension of the input vector. F_2 is the "competitive" layer with N nodes representing clusters. A cluster is given by a single prototype: $w_j = (w_{1j}, w_{2j}, ..., w_{2Mj})$. All F_1-to-F_2 connections are associated with real-valued weights $w_{ij} \in [0, 1] : i = 1, 2, ..., 2M$ and $j = 1, 2, ..., N$. The mapping layer F^{ab} has L nodes representing the number of classes in the output space. All F_2-to-F^{ab} connections are associated with binary weights $w^{ab}_{jk} \in \{0, 1\} : j = 1, 2, ..., N$ and $k = 1, 2, ..., L$. Fuzzy ARTMAP (FAM) networks also operates with L_1-norm $|a|= \sum_i a_i$ and fuzzy AND operator $\wedge$ defined as $(a \wedge b)_i = min(a_i, b_i)$. Three parameters exist: the choice parameter $\alpha > 0$, the learning rate $\beta \in [0, 1]$ and the vigilance parameter $\rho \in [0, 1]$. Figure 3.19 schematically demonstrates an SFAM network.

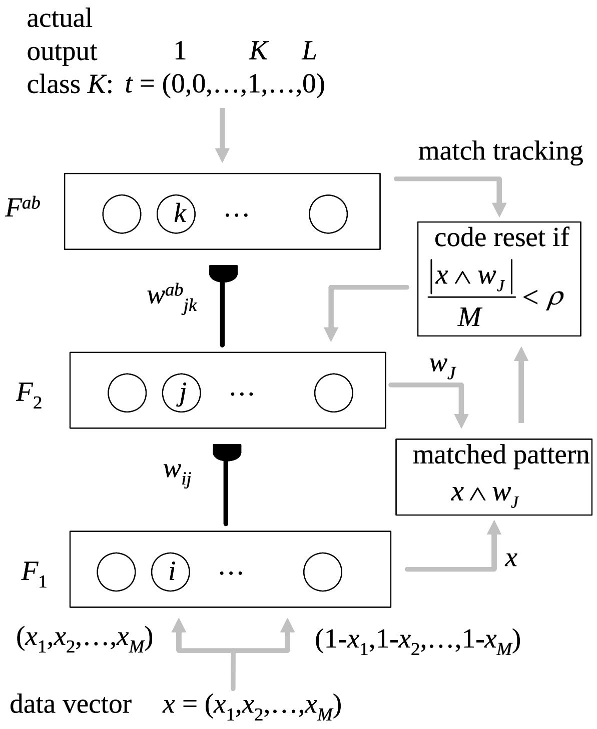

actual
output 1 K L
class K: t = (0,0,…,1,…,0)
match tracking
F^{ab}
k
…
code reset if
$\frac{|x \wedge w_J|}{M} < \rho$
w^{ab}_{jk}
F_2
j
w_J
matched pattern
$x \wedge w_J$
w_{ij}
F_1
i
x
$(x_1,x_2,\ldots,x_M)$
$(1\text{-}x_1,1\text{-}x_2,\ldots,1\text{-}x_M)$
data vector $x = (x_1,x_2,\ldots,x_M)$

Figure 3.19: Simplified fuzzy ARTMAP network

Having training pairs (x, t) where x is a training data vector $x = (x_1, x_2, ..., x_M)$ and t is a vector of output labels $t = (t_1, t_2, ..., t_L)$, where only one element t_K that corresponds to the target class K becomes 1 and the remaining elements are 0. In the learning mode, training vectors are sequentially presented to the network via input layer F_1 and the corresponding class labels via mapping layer F^{ab}. SFAM operates as follows:

1. *Initialization.* No one of F_2 nodes is committed, weights $w_{ij} = 1$ and $w^{ab}_{jk} = 0$. Parameters α, β and ρ_{max} are set by user and $\rho' = 0$, $\epsilon = 0^+$.
2. *Complement coding.* A training sample $x = (x_1, x_2, ..., x_M)$ of the dimension M is expanded to the vector $x = (x_1, x_2, ..., x_M, 1-x_1, 1-x_2, ..., 1-x_M)$ of the dimension $2M$. The vigilance ρ' is reset to its initial value.
3. *Prototype selection.* Input x activates neurons in F_1 and is propagated through weighted connections w_{ij} to F_2. Activation of node j in F_2 is determined by the *choice function* given

by (3.21). The winner-takes-all rule is applied. Actually, vector x is successively compared with vectors w_j and the node j with the highest response (denoted as J) is activated.

$$J = argmax_j \left(\frac{|x \wedge w_j|}{\alpha + |w_j|} \right) \tag{3.21}$$

The prototype vector w_J of the node J is propagated back onto F_1 to perform the *vigilance test* (given by Equation (3.22)). The test determines the degree of similarity between w_J and x, and compares the result with the vigilance parameter ρ'.

$$\frac{|x \wedge w_j|}{|x|} = \frac{|x \wedge w_j|}{M} \geq \rho' \tag{3.22}$$

If the test is passed, then the node J is considered to "resonate" with an input. Otherwise, the network resets the active F_2 node and searches for another node J that eventually passes the vigilance test. In case all F_2 nodes do not satisfy the match criterion, an uncommitted F_2 node becomes active and the network enters the learning stage (Step 5).

4. *Class prediction.* The mapping layer F^{ab}, on the one hand, directly receives the class label t_K and, on the other hand, the active pattern form F_2 is propagated to F^{ab} via associative connections w^{ab} by means of the *prediction function* given by (3.23).

$$S_k^{ab} = \sum_{j=1}^{N} \delta_j w_{jk}^{ab}, \quad \delta_j = 1 \Leftrightarrow j = J \tag{3.23}$$

where δ_i determine states of F_2 nodes. The most active F^{ab} node $K = k(J)$ yields the class prediction. If node K yields an incorrect class prediction, a *match tracking* signal slightly raises vigilance ($\rho' = |x \wedge w_J|/M + \epsilon$) to continue searching among F_2 nodes (Step 3). The search stops either if the node J that has previously learned the correct class prediction K becomes active or if the vigilance reaches the limit ρ_{max} and an uncommitted F_2 node becomes active with the succeeding learning (Step 5).

5. *Learning.* Learning input vector x implies updating prototype vector w_J if the winner node J corresponds to the committed F_2 node (see Equation (3.24)), or creating an associative link to F^{ab} if the winner node J corresponds to the uncommitted F_2 node. A new association between the J node in F_2 and the K node in F^{ab} is provided by setting $w_{jK}^{ab} = 1$ where K is the target class label for x.

$$w_J = \beta(x \wedge w_J) + (1 - \beta) w_J \tag{3.24}$$

The training stops after weights w_{ij} and w_{jk}^{ab} have converged. If fast learning is chosen (β =1), training patterns need to be propagated through the network only once. Otherwise, several training epochs are required. The class prediction for a given test vector x is done by performing Steps 2, 3 and 4 and associating x with the class $K = k(J)$ that corresponds to the activated node J.

Note the complement coding is required to normalize vectors and more importantly to solve the "cluster proliferation" problem. As demonstrated by Moore [175] the noisy data invokes continuous creation of new clusters, because cluster seeds are given by the minimum values that become less and less during a training finally approaching zero. With complement coding, prototypes automatically contain low and high margins of the cluster and therefore can be imagined as hypercubes or boxes [215]. In a two-dimensional case a hypercube corresponds to a rectangle. The vigilance parameter ρ is a kind of similarity threshold determining the spread of samples within a cluster or rather placing an upper limit on the size of each category box [215], the less ρ the greater boxes. The geometrical interpretation of the class forming process is shown in Figure 3.20 and compared to that of k-NN.

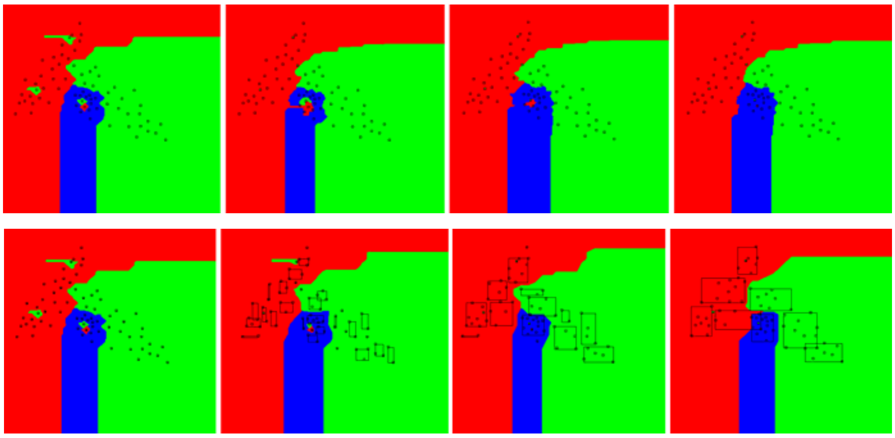


Figure 3.20: Two-dimensional case for comparison of classification results of k-NN with city-block distance in the top row (k=1, k=2, k=3, k=4) and fuzzy ARTMAP in the bottom row (ρ=1.0, ρ=0.95, ρ=0.9, ρ=0.85)

The critical point of ART is that the class boundaries depend on the order of training patterns generating scepticism about the statistical “consistency” of the approach [215]. Despite the apparent inconsistency, FAM has been successfully applied for a number of applications in the domain of pattern recognition. The most notable examples are 3-D visual object recognition [24] or texture segmentation [18], as well as face recognition [9, 94, 231, 47]. However, FAM should not be applied to noisy data because of the possibility of category proliferation [252].

3.3.4 Sparse network of Winnows

The sparse network of Winnows (SNoW) is a further architecture for incremental supervised learning invented by Roth [31]. It can be seen as a multi-class classifier that is specifically designed for high-dimensional sparse vectors.

The version of SNoW described here is a two-class classifier applied by Nilsson [180] for face detection. Although SNoW is applied to binary input vectors it does not shrink the generality of the approach, because any integer vector (addressed in image processing) can be easily transformed to a binary vector using sparse binary representations of decimal numbers. An example of such transformation is given in Figure 3.21. After transforming a K-dimensional L-bit vector to a binary vector, the resulting vector becomes $M = K \cdot 2^L$ dimensions.

For the formal description of the SNoW classifier let us introduce the training set of M-dimensional binary feature vectors $x_k \in \mathbb{R}^M$ as $x_k = (x_{k,1}, x_{k,2}, ..., x_{k,M})^T$ with the corresponding class labels $y_k : y_k \in \{-1, +1\}, k = 1, 2, ..., N$. So the input of the classifier is a pair (x_k, y_k).

The SNoW classifier makes use of two lookup tables denoted as h^+ and h^-. Lookup tables are the real-value vectors of the same dimension as the feature vector x. They can be generally understood as class models or class prototypes. The heuristic used to estimate the probability $Pr(y = +1|x) \sim Pr(x|y = +1)$ that the vector x belongs to the class ‘+1’ is $\sum_{i=1}^{M} h_i^+ x_i$ and the

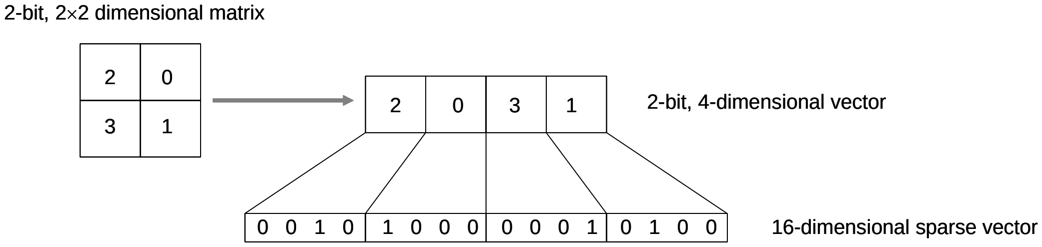

2-bit, 2×2 dimensional matrix
2 0
3 1
2 0 3 1
2-bit, 4-dimensional vector
0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 0
16-dimensional sparse vector

Figure 3.21: 4-dimensional 2-bit vector is transformed to a sparse 16-dimensional binary vector

heuristic for the probability $Pr(y = -1|x) \sim Pr(x|y = -1)$ is $\sum_{i=1}^{M} h_i^- x_i$. The decision to assign a test feature vector to one or another class is made by comparing these heuristics. The higher sum wins. Since x is a binary vector, the computation of the sum can be interpreted as the search in the lookup table for features that are presented in the vector x. At the same time, the real values from lookup tables determine the importance of features be assigning them weights.

Formally, the training of the SNoW classifier is the learning of the lookup tables h^+ and h^- from the given training patterns. The learning is based on the Winnow update rule [150] and consists of three steps: initialization, promotion/demotion and termination. Since tables h^+ and h^- are learned simultaneously utilizing the same learning procedure, the learning steps are described only for h^+ without losing generality.

1. *Initialization.* Initially h^+ is initialized with zeros: $h^+ = (0, 0, \ldots, 0)$. The auxiliary vector h^+_{active} is introduced to indicate the currently committed dimensions (features). This binary vector has the same dimension as h^+ and is also initialized with zeros: $h^+_{active} = (0, 0, \ldots, 0)$. There are three training parameters that are set by a user: the threshold γ, the promotion coefficient $\alpha > 1$ and the demotion coefficient $0 < \beta < 1$. Set the iteration $t = 1$ and begin with Step 2.

2. *Promotion/demotion.* The principal idea here is that the table components providing a correct decision are promoted and the table components providing a wrong decision are demoted. The training pairs $(x_k, y_k), k = 1, 2, ..., N$ are consecutively presented to the classifier.

   a) Choose the first training vector $x_k, k = 1$ and set the number of classification errors at the current iteration t to zero: $e(t) = 0$.

   b) Determine the uncommitted active features $i : x_{k,i} = 1$ and $h^+_{active,i} = 0$. For these features set h_i^+ and $h^+_{active,i}$ to 1.

   c) If the dot product $(h^+ \cdot x_k) < \gamma$ and $y_k = +1$ then the active table components are *promoted*: $h_i^+ = \alpha h_i^+$ for indices i that correspond to active features in $x_k (x_{k,i} = 1)$ and the error is incremented: $e(t) = e(t) + 1$.

   d) If the dot product $(h^+ \cdot x_k) \geq \gamma$ and $y_k = -1$ then the active table components are *demoted*: $h_i^+ = \beta h_i^+$ for indices i that correspond to active features in $x_k (x_{k,i} = 1)$ and the error is incremented: $e(t) = e(t) + 1$.

   e) If $k = N$ then go to termination (Step 3) otherwise take the next training vector $x_k, k = k + 1$ and go to Step 2b.

3. *Termination.* If $e(t) = 0$ or the number of iterations t has reached the predefined limit then terminate the learning of the table h^+. Otherwise the iteration is incremented $t = t + 1$ the learning is continued with the Step 2.

If both lookup tables h^+ and h^- operate in the same domain, or to be more precise, the same training vectors are used to learn both tables, then they can be replaced by the single lookup table $h : h = h^+ - h^-$, and the corresponding heuristic for the class selection of the test vector x becomes $\sum_{i=1}^{M} h_i x_i > \theta$ where θ is a decision threshold. The decision function $f(x)$ of the SNoW classifier is given by Equation (3.25).

$$f(x) = \begin{cases} +1 & if \quad \sum_{i=1}^{M} \left(h_i^+ - h_i^-\right) x_i = \sum_{i=1}^{M} h_i x_i > \theta \\ -1 & otherwise \end{cases} \tag{3.25}$$

SNoW seems to operate similarly to the nearest centroid classifier [83]. In fact, the class prototypes given by lookup tables are consecutively updated towards new training vectors if they belong to the same class and away from training vectors from another class. Properly defined promotion and demotion coefficients guarantee that class prototypes converge to class means. However, incorrect selection of these parameters can lead to the inconsistency of the classifier. It is interesting to note that the same machine learning approach applied to classification of documents using $tf * idf$ vectors is referred to as the Rocchio classifier [163].

3.3.5 Support vector machine

While FAM and SNoW are inferential or rather constructive approaches suffering from the deficiency of a strong theoretical foundation and often criticized for statistical inconsistency, the support vector machine (SVM) is based on mathematical models that are both theoretically well-founded and geometrically intuitive [15]. Moreover, the approach is systematic, reproducible and properly motivated by statistical learning theory. The theoretical background of SVMs is so-called Vapnik-Chervonenkis (VC) theory which has been developed by Vapnik and Chervonenkis since the 1960s and involves concepts such as: consistency of a learning process, convergence a of learning process, generalization ability of a learning process, and constructing of learning machines (for more detail see books of Vapnik [239, 240]). The first practical approach, however, is proposed in 1995 in the paper of Cortes and Vapnik [43] and addresses binary pattern classification.

Generally, SVM relies on the principle of structural risk minimization, meaning that the optimal separating hyper-plane is obtained as a trade-off between the empirical risk to misclassify a sample and the complexity of the generated class models. This guarantees a high level of generalization. Therefore, SVM is even effective for classification of high-dimensional data being in possession of a small set of training samples. Furthermore, SVMs can solve linearly non-separable problems using the kernel trick [2].

In order to formally describe SVM let us introduce the training set of M-dimensional feature vectors $x_i \in \mathbb{R}^M$ as $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,M})^T$ with the corresponding class labels $y_i : y_i \in \{-1, 1\}, i = 1, 2, ..., N$ where N is the number of training vectors. The training of the classifier is carried out based on pairs (x_i, y_i). After training, the class of the test vector is determined by a linear decision function $f(x) = sign((w \cdot x) - b)$ where the vector w determines the orientation of a separating plane and the scalar value b determines the offset of the plane from the origin. Three concepts form the basis of SVM: margins, duality and kernels.

Margins For two linearly separable classes the infinite number of hyper-planes can be found that correctly separate class vectors (see an example for the two-dimensional case in Figure 3.22a. The principal question is: what separating hyper-plane should be considered the "best" one?

In the SVM approach, the separating hyper-plane for two classes is determined exclusively by the vectors lying on the boundaries (so-called support vectors) and the "bests" separating hyper-plane

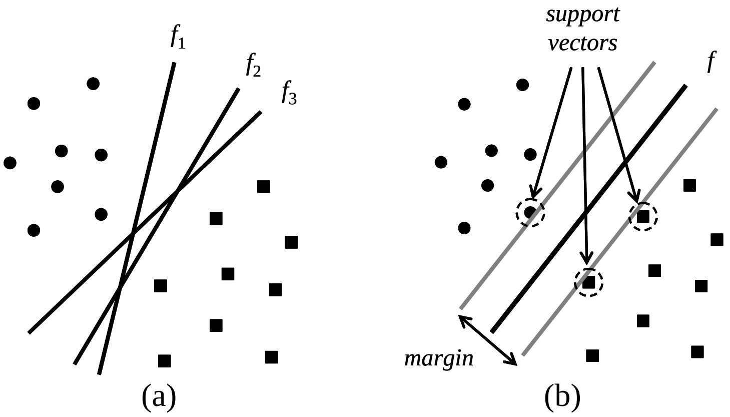


Figure 3.22: In a two-dimensional case a separating hyper-plane is represented by a line: (a) three arbitrary selected separating lines; (b) separating line selected by SVM

is that one, which is maximally distant from support vectors of both classes. In other words the separating plane maximizes the margin between classes and is located directly in the middle of the margin (see Figure 3.22b).

Duality The problem of discovering the optimal separating hyper-plane is formulated as a quadratic optimization problem with constraints that can be solved by well-known techniques [183]. There are two mathematically identical ways to define the hyper-plane maximizing the between-class margin. The first strategy is considering convex hulls of the classes and obtaining the separating plane as a bisector of the line connecting the closest points of the convex hulls (see Figure 3.23).

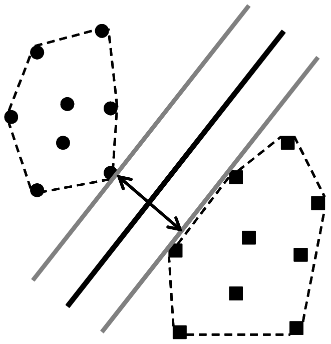


Figure 3.23: Convex hulls and the optimal separating line

The closest points in two convex hulls can be found by solving the following quadratic optimization problem:

$$\frac{1}{2}\left\|\sum_{i:y_i=1}\alpha_i x_i-\sum_{i:y_i=-1}\alpha_i x_i\right\|^2 \underset{\alpha}{\rightarrow} \min \quad s.t. \quad \sum_{i:y_i=1}\alpha_i=\sum_{i:y_i=-1}\alpha_i=1, \quad \alpha_i\geq 0, i=1,2,...,N \tag{3.26}$$

An alternative strategy is the straightforward maximization of the margin between two parallel supporting planes. A plane is considered to support a class if all points in the class are located on one side of that plane. Geometrically, this procedure can be interpreted as pushing apart the supporting planes until they bump into the support vectors [15]. Formally, vectors labeled with '+1' ($y_i = 1$) can be described by the inequality $(w \cdot x_i) - b \geq 1$ as well as vectors labeled with '-1' ($y_i = -1$) can be described by the inequality $(w \cdot x_i) - b \leq -1$. The supporting planes are given by equations $(w \cdot x_i) - b = 1$ and $(w \cdot x_i) - b = -1$ correspondingly. The margin between the supporting planes yields $2/(w \cdot w)$ (see [239] for details). So, the problem of margin maximization is equivalent to the minimization of the scalar product $(w \cdot w)$ with respect to the general constraint $y_i((w \cdot x_i) - b) \geq 1$:

$$\frac{1}{2}(w \cdot w) \xrightarrow[w,b]{} \min \quad s.t. \quad y_i((w \cdot x_i) - b) \geq 1 \tag{3.27}$$

Applying Lagrange multipliers α_i and taking into account the Kühn-Tucker conditions, the optimization problem (3.27) can be modified to the following optimization problem [239] that is in fact dual to (3.27) and equivalent to (3.26):

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \xrightarrow[\alpha]{} \max \quad s.t. \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \alpha_i \geq 0, i = 1, 2, ..., N \tag{3.28}$$

Since (3.26) and (3.28) rely on the same support vectors, the solution of (3.28) is equal to the solution of (3.26) and yields the normal vector to the plane $w = \sum_{i=1}^{N} y_i \alpha_i x_i$. The threshold b is determined by the support vectors [15].

If two classes are linearly non-separable, then the data points located on the "wrong" side of the separating plane are added as penalties in (3.27):

$$\frac{1}{2}(w \cdot w) + C\sum_{i=1}^{N} \xi_i \xrightarrow[w,b,\xi_i]{} \min \quad s.t. \quad y_i((w \cdot x_i) - b) + \xi_i \geq 1, \xi_i \geq 0, i = 1, 2, ..., N \tag{3.29}$$

and included as an additional constraint to the optimization problem (3.28):

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \xrightarrow[\alpha]{} \max \quad s.t. \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C, i = 1, 2, ..., N \tag{3.30}$$

Hence, the optimal separating hyper-plane not only maximizes the margin but also minimizes error. Coming back to the convex hulls, there is a very intuitive interpretation of the error minimization. By adding the aforementioned constraint ($0 \leq \alpha_i \leq C$) to the dual optimization problem (3.26), the convex hulls are reduced in size. Selecting an appropriate threshold C, the initially overlapped convex hulls become non-overlapped. This can be clearly seen in Figure 3.24. This concept is addressed as constructing soft-margin separating hyper-planes [43].

Kernels However, there are situations when two classes cannot be sufficiently well separated by a linear function. The most common example is illustrated in Figure 3.25a. Here, two-dimensional space vectors of one class encompass vectors of another class. The classes cannot be separated by a line. However, a quadratic function (e.g. circle) can perfectly separate these classes. Applying higher-dimensional separating hyper-planes is equivalent to projecting data points into a higher-dimensional space (e.g. sphere) and using a higher-dimensional separating hyper-plane (e.g. plane instead of the line) as shown in Figure 3.25b. The general purpose is that in new higher-dimensional

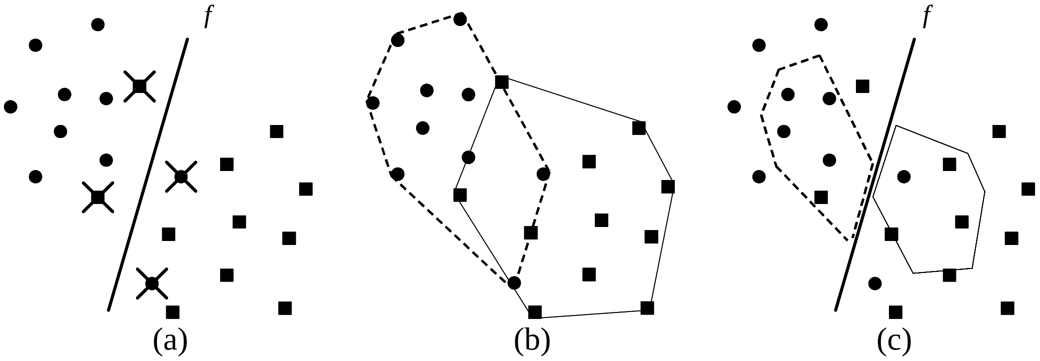


Figure 3.24: (a) An example of linearly non-separable classes (misclassified vectors are marked by crosses); (b) overlapped convex hulls; (c) reduced convex hulls after including the constraint $0 \leq \alpha_i \leq C$

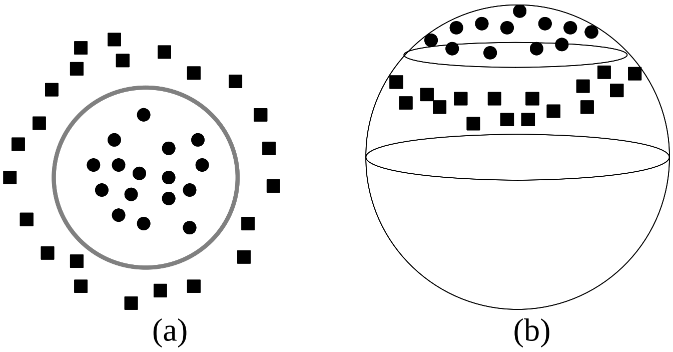


Figure 3.25: (a) An example where the linear discriminant function fails; (b) the same data points projected onto the sphere

space the data points either become linearly separable or the number of wrongly assigned points becomes low.

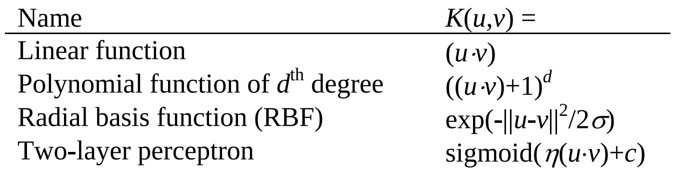
Formally, the projection into a higher-dimensional space is done by substituting the dot product $(x_i \cdot x_j)$ in (3.30) by the generalized dot product in the Hilbert space also referred to as the kernel function $K(x_i, x_j)$. According to the Hilbert-Schmidt theory the kernel function is any symmetric function satisfying Mercer's theorem [239]. This simple trick is possible because the data points are included into the optimization problem (3.29) exclusively as the dot product. Utilizing some nonlinear function of data points $\theta(x_i)$ in the place of data points x_i implies the substitution of the dot product $(x_i \cdot x_j)$ by the dot product $(\theta(x_i) \cdot \theta(x_i))$ that is identical to the kernel function $K(x_i, x_j)$. In the literature [2], the substitution of the dot product by its generalized version is referred to as the kernel trick. The optimization problem is then modified to:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \underset{\alpha}{\rightarrow} \max \quad s.t. \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C, i = 1, 2, ..., N \tag{3.31}$$

and the decision function is constructed as:

$$f(x) = sign\left(\sum_{i \in S} y_i \alpha_i K(x, x_i) - b\right) \tag{3.32}$$

where S is the set of support vectors.
The most commonly used kernels:



Name	$K(u,v) =$
Linear function	$(u \cdot v)$
Polynomial function of d^{th} degree	$((u \cdot v)+1)^d$
Radial basis function (RBF)	$\exp(-\|u\text{-}v\|^2/2\sigma)$
Two-layer perceptron	$\text{sigmoid}(\eta(u \cdot v)+c)$

Concluding the description of SVM the procedure of learning and applying the classifier can be formalized as follows:

1. Initialize the classifier by selecting two parameters: the constraint C in (3.31) as the trade-off between maximizing the between-class margin and minimizing error obtained with the training set and the kernel function with appropriate parameters.

2. Solve the optimization problem (3.31) discovering support vectors and optimal separating plane.

3. Classify a test vector x using (3.32).

The most important characteristic of the statistical learning process is the predictability of the misclassification rate arising with unknown sets of data points. Formally, it is called bounds of the generalization error that is obtained as a function of the classification error with the given training data. As shown by Vapnik [240], linear functions maximizing the between-class margin minimize bounds of the generalization error. In other words, the wider margin guarantees better generalization. In VC-theory it is called structural risk minimization. Moreover, the upper bound can be calculated as a function of the number of training vectors, number of support vectors, radius of the sphere containing the training vectors and in some sense by the dimensionality of the vector space (see [240] for details). However, the size of the margin is not directly dependent on the dimensionality. Hence, the high-dimensional data will not necessary cause overfitting.

Margin maximization can also be expressed in terms of the complexity reduction of the decision function. According to Occam's razor principle (see Subchapter 2.1.3) more simple decision functions are preferred because complex decision functions have the capacity to very precisely fit the training data and are therefore likely to provide poor generalization. Hence, it can intuitively be asserted that the margin maximization automatically reduces complexity of the decision function because a "skinny" separation margin can take many possible orientations and still strictly separate the training data and a "fat" separation margin has a limited flexibility to separate the training data [15].

3.3.6 Adaptive boosting

Boosting is a term describing the process of creating an ensemble of classifiers from resampled data [197]. The basic idea is to use subsets of training data to train a bunch of subclassifiers (also referred to as weak classifiers) and then combine them to a cascade (also referred to as a strong classifier).

Formally, boosting is an iterative process creating three weak classifiers on each iteration. The first classifier C_1 is trained from a randomly chosen subset of training data. The second classifier C_2

is trained from the data representing the most informative subset defined by C_1. It is required that only half of the samples used to train C_2 are correctly classified by C_1 (another half is misclassified). The third classifier C_3 is trained from samples on which C_1 and C_2 disagree. The decisions of C_1, C_2 and C_3 are combined through three-way majority voting [197].

In other words, the resampling of training data is done to permanently focus on the most informative training samples playing the major role in designing the decision function. This is why the foregoing classifiers throw off the high number of correctly classified training samples ("easy" ones) and the succeeding classifiers focus on misclassified samples ("hard" ones). This very intuitive way of constructing separating hyper-planes is illustrated in Figure 3.26. The final strong classifier is created by the recursive application of boosting. Schapire [216] proved that boosting can be seen as the strict probably approximately correct (PAC) learning introduced by Valiant in [237].

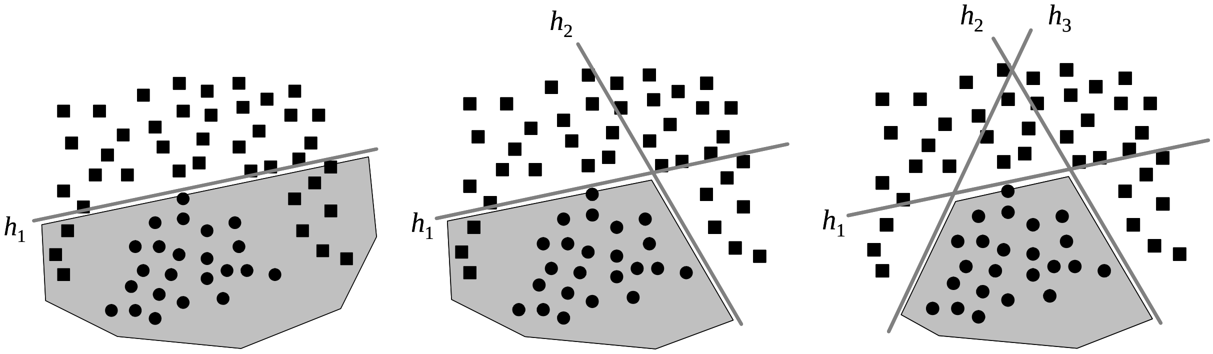


Figure 3.26: 1st, 2nd and 3rd iterations of boosting

Adaptive boosting (AdaBoost) is an algorithm introduced by Freund and Schapire in 1995 [76] which practically implements boosting and solves many previously unsolved technical complications. Adaptive boosting is initially designed as a binary classifier to solve two-class problems, but in their later publications Freund and Schapire [77, 78] present modifications of AdaBoost for multi-class and regression problems.

The main idea of the AdaBoost algorithm is to assign a dynamic weight to each training sample so that the samples are ranked according to their importance on the current iteration of the learning process. Let us denote the weight of the i^{th} training sample on t^{th} iteration as $D_t(i)$. In the beginning, the weights are initialized with equal values forming a uniform distribution. On each iteration of learning, the weights of misclassified samples are increased so that the "hard" samples can be distinguished on the subsequent iterations and the succeeding weak classifiers put emphasis on "hard" samples. The weak classifier h_t can be any decision function associating the input samples with one or another class $h_t : x \to \{-1, +1\}$ with respect to the distribution D_t. The most common example of the weak classifier is a one-level decision tree also referred to as decision stump. Weak classifiers are ranked based on their classification error which is the sum of weights of the misclassified samples. The parameter measuring the importance of the weak classifier h_t is denoted as α_t. The value of α_t gets larger as the classification error gets smaller. The final step is the update of the distribution of weights D_t. The weight of training samples misclassified by h_t is increased while the weight of samples correctly classified by h_t is decreased. This means that the often misclassified samples gradually accumulate weights and the weights of correctly classified samples approaches zero. The final hypothesis H is a weighted majority vote of the T weak hypotheses where α_t is the weight assigned to h_t.

Let us give a formal description of AdaBoost. Given the training set of N vectors $x_i, i = 1, 2, ..., N$

with the corresponding class labels $y_i : y_i \in \{-1, 1\}$ the training process can be formalized as follows:

1. Initialize the sample weights by equal values: $D_1(i) = \frac{1}{N}, i = 1, 2, \ldots, N$
2. For $t = 1, 2, \ldots, T$ do

   a) Train a weak classifier h_t with respect to the distribution D_t,

   b) Calculate the training error ϵ_t of the weak classifier $h_t : \epsilon_t = Pr(h_t(x_i) \neq y_i) = \sum_{i:h_t(x_i)\neq y_i} D_t(i)$. If $\epsilon_t = 0$ or $\epsilon_t > 0.5$, set T to $t-1$ and break.

   c) Assign the weight $\alpha_t = \frac{1}{2} log \frac{1-\epsilon_t}{\epsilon_t}$ to the weak classifier h_t,

   d) For each sample $i = 1, 2, \ldots, N$ decrease the weight of the sample that has been correctly classified and increase the weight of the sample that has been misclassified:

   $$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & if \quad h_t(x_i) = y_i \\ e^{\alpha_t} & if \quad h_t(x_i) \neq y_i \end{cases} = \frac{D_t(i) \cdot exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}$$

   where Z_t is a normalization constant, such that $\sum_{i=1}^{N} D_{t+1}(i) = 1$.

   The resulting strong classifier is given by $H(x) = \sum_{t=1}^{T} \frac{\alpha_t}{\sum_{t=1}^{T} \alpha_t \cdot h_t(x)}$.

An important characteristic of AdaBoost is its immunity to overfitting. Intuitively, the high number of iterations of boosting T leads to too precise fitting of training data and therefore to high generalization error. However, as shown by Schapire et al. [217] generalization error is entirely independent of T.

As shown in [79], AdaBoost described in terms of maximization of the between-class margin has a very close relation to SVM. In AdaBoost the margin for the sample (x, y) is defined by:

$$y\left(\sum_t \alpha_t h_t(x)\right) / \sum_t \alpha_t \tag{3.33}$$

where $h(x)$ denotes the vector of weak classifiers $(h_1(x), h_2(x), \ldots, h_N(x))$ which generally represents the instance vector. The weight vector is represented by the vector of coefficients $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$. In these terms the problem of margin maximization can be generally formulated as:

$$\max_{\alpha} \min_{i} \frac{(\alpha \cdot h(x_i))y_i}{||\alpha|| \cdot ||h(x_i)||} \tag{3.34}$$

So, both approaches maximize the minimum margin on training samples and the only difference is that SVM uses the l_2 norms for both the instance vector and the weight vector, just as AdaBoost uses the l_∞ norm for the instance vector ($||h(x_i)||_\infty = 1$) and l_1 norm for the weight vector. However, as explained in [79] the significant difference between SVM and AdaBoost is reflected by the following aspects: (1) utilizing different norms can result in very different margins, (2) computational requirements are different, and (3) a different approach is used to search efficiently in high-dimensional space.

Chapter 4

4 Design

This chapter addresses the design and implementation of the three proposed applications: seat occupancy detection, discrimination of driver and passenger hands, and facial driver recognition. This includes the introduction of application and operational scenarios, choice of type and position of camera and active illumination sources, and the development of software components under application constraints.

4.1 Vision-based seat occupancy detection

This section is comprised of the following subsections:

- Introduction
- Application scenario
- Operational scenario
- Acquisition system
- Illumination
- Computer vision approaches
- Template matching for empty seat detection
- Face as an identifier of an occupied seat
- Combination of the empty seat detection and face detection
- Evaluation objectives

4.1.1 Introduction

As mentioned in Chapter 1, the vision-based approach to seat occupancy sensing is the most promising one among available sensing approaches regarding cost and versatility of use. The development of an optical seat occupancy detection system is one of the key challenges of the VisMes project, the joint project of the AMSL and Volkswagen Group Research. In the framework of the project the Audi A6 Avant is retrofit with an omni-directional camera, five near-infrared lighting modules, and a video capturing system. This hardware setup is used in this thesis as the basis for the empirical performance evaluation of the proposed computer vision system.

According to the methodology proposed in Chapter 3, a vision-based automotive application consists of five components: application scenario, operational scenario, acquisition device, illumination modules and computer vision algorithms. These components are schematically illustrated in Figure 4.1.

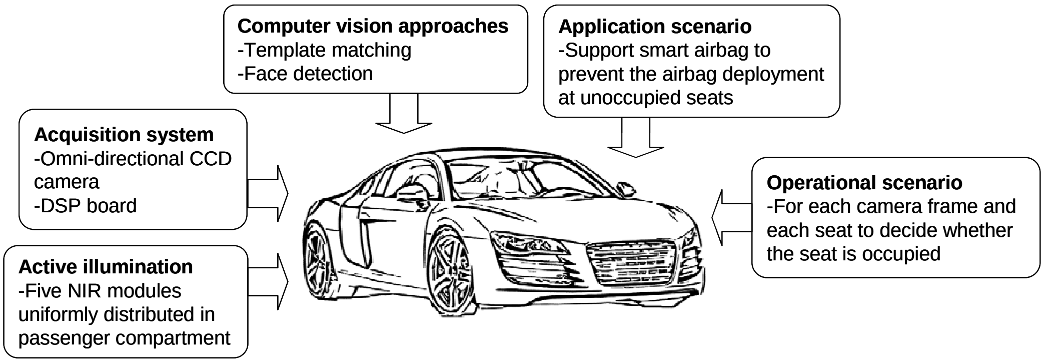

Computer vision approaches
-Template matching
-Face detection
Application scenario
-Support smart airbag to prevent the airbag deployment at unoccupied seats
Acquisition system
-Omni-directional CCD camera
-DSP board
Operational scenario
-For each camera frame and each seat to decide whether the seat is occupied
Active illumination
-Five NIR modules uniformly distributed in passenger compartment

Figure 4.1: Components of the vision-based seat occupancy detection system

4.1.2 Application scenario

Seat occupancy detection is designed to support conditional airbag deployment (so-called smart airbag). In case of an accident the airbag deployment must be prevented at unoccupied seats, seats occupied by child restraint systems or children less than 66 lbs, and in cases in which a passenger is out-of-position (OOP).

4.1.3 Operational scenario

The system begins to operate after the motor is started. The decision (occupied/not occupied) has to be provided for all five seats and each camera frame. The system has to be the real time system meaning that at each moment there is information about the seat occupancy. If the system is unable to decide for a particular frame, the decision has to be adopted from the previous frame. The system has to be able to work at day and night regardless of weather conditions.

4.1.4 Acquisition system

An omni-directional camera is mounted on the ceiling next to the windshield to simultaneously observe all seats in the car passenger compartment. The applied camera is Sony RPU-C3522 that has a CCD sensor and a 360-degree annular lens. The IR-filter is removed to enable capturing of the reflected NIR wavelengths up to 1.0 μm. This is possible due to the photoelectric properties of silicon which is the basic element of the imaging sensor. The frames are captured at a rate of 7.5 FPS with a resolution of 720 x 576 pixels. The camera frame contains two half pictures of a passenger compartment located in top and bottom parts. Joined together these two parts represent the complete 360-degree view of the passenger compartment (See Figure 4.2).

4.1.5 Illumination

Four NIR lighting modules are installed in the top of the door frames and one module in the center of the ceiling. The exact positions can be seen in Figure 4.3. The wavelength of NIR LEDs is approximately 850 nm. The combination of NIR modules and a NIR sensitive camera allows operating at night.

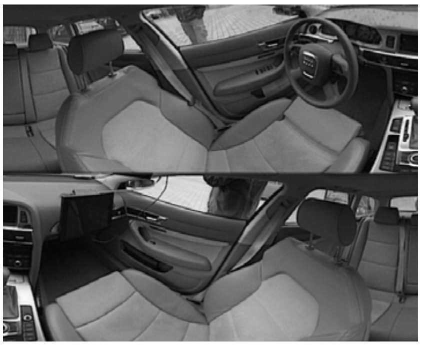


Figure 4.2: 360-degree view of the Audi A6 passenger compartment from the VisMes project

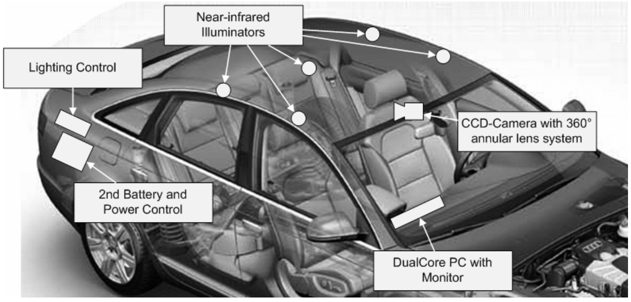

Near-infrared Illuminators
Lighting Control
CCD-Camera with 360° annular lens system
2nd Battery and Power Control
DualCore PC with Monitor

Figure 4.3: Experimental vehicle equipped with a 360-degree camera, five near-infrared lighting modules and digital signal processing system (reprinted from [160])

4.1.6 Computer vision approaches

There are two fundamentally different concepts for seat occupancy detection addressed here. The first one checks whether the seat is empty or not, and the second one looks for a face in the image region assigned to the seat. Both approaches can only indirectly reveal the current seat state. Empty seat detection provides no information about the occupancy type if the seat is occupied whereas face detection cannot help to determine whether the seat is empty or occupied by an object if no face is detected. So, the approaches complement each other detecting empty seats and distinguishing between subjects and objects. The empty seat detection is based on advanced template matching whereby the characteristic parts (templates) of an empty seat are searched for in the corresponding seat areas [160]. In order to perform face detection three well-known algorithms are challenged: the Viola-Jones algorithm [242], the algorithm of Kienzle et al. [127] and the algorithm of Nilsson et al. [180].

4.1.7 Template matching for empty seat detection

Template matching can be considered a simplified case of statistical pattern recognition applied to the small parts of an image in the framework of sliding window approach. The reference database consists of one image referred to as template which is considered a simplified statistical model of an object of interest. Classification degrades to the comparison of image blocks and the template. Images are represented by pixel intensities meaning that feature extraction is omitted. Hence, the success of template matching drastically depends on pre-processing, which is used to flatten out all dramatic influences of non-uniform illumination to the appearance of objects.

As mentioned above, template matching utilizes sliding window approach. The window of particular size (say $N \times M$ pixels) slides across the image I. The resulting local blocks R_{xy} ($x = 1, 2, \ldots, W$; $y = 1, 2, \ldots, H$) are compared to the template T (of the same size $N \times M$ pixels) using normalized cross-correlation coefficients (NCCC) denoted as r_{xy}. NCCC form correlation matrix r. The lateral shift of the sliding window is usually chosen to be less than the linear size of the window leading to the block overlap. Assume, the size of the original image I is $w \times h$ pixels and the step in both directions is selected to be s pixels, the linear dimensions of the correlation matrix r become $W = w/s - N$ and $H = h/s - M$. Equation 4.1 provides the computation of NCCC:

$$r_{xy} = \frac{\sum_{j=-M/2}^{M/2} \sum_{i=-N/2}^{N/2} \left[\left(I(x+i, y+j) - \mu_R \right) \cdot \left(T(i+\frac{N}{2}, j+\frac{M}{2}) - \mu_T \right) \right]}{\sqrt{\sum_{j=-M/2}^{M/2} \sum_{i=-N/2}^{N/2} \left(I(x+i, y+j) - \mu_R \right)^2} \sqrt{\sum_{j=-M/2}^{M/2} \sum_{i=-N/2}^{N/2} \left(T(i+\frac{N}{2}, j+\frac{M}{2}) - \mu_T \right)^2}} \tag{4.1}$$

where μ_T is the mean value of pixels from T and μ_R is the mean value of pixels from R_{xy}. Values r_{xy} are within the interval [-1,1]. Absolute values of r_{xy} approaching 0 indicate no correlation (dissimilarity of T and R_{xy}). In contracts, absolute values of r_{xy} approaching 1 indicate high correlation (similarity of T and R_{xy}).

The correlation matrix r can be interpreted as a two-dimensional surface over the original image. The maximal peak on this surface indicates the location of the image block which is the most similar to the template. The result of template matching is therefore the maximum value of the correlation coefficient $r_{max} = max(r)$ and its position (x,y) in the matrix r. Turning back to seat occupancy detection, an empty seat is indicated if r_{max} exceeds the predefined threshold τ which is set up manually or identified during the training of the system. Otherwise ($r_{max} \leq \tau$) the seat is indicated to be occupied. In the following, the maximal correlation coefficient is referred to as the matching score and denoted as S_{max}.

Template matching is computationally expensive because NCCC need to be computed in many points of the image namely $W \times H$ times. The computation of every single NCCC takes $N \times M$ operations. For practical reasons the selection of small templates is preferable.

The practical implementation of the algorithm begins with the selection of a reference frame of an empty car passenger compartment. In this reference image, seat regions are marked: front left (FL), front right (FR), rear left (RL), rear center (RC) and rear right (RR). Templates are selected inside of corresponding regions, cut and permanently stored. Having coordinates of seat regions and stored templates, the processing of video frames can be started. Based on the first frame from the video stream the frame stabilization is done in order to reveal the displacement between the reference frame and frames from the current video stream. Thanks to stabilization, the coordinates of seat regions are adjusted to actual camera images. The search for an appropriate template is provided only inside of the corresponding seat regions. The general workflow of the system is illustrated in Figure 4.4. If the template is found, the seat is classified as empty, otherwise it is classified as occupied without specifying the type of occupancy. In order to increase the robustness of the system, in case all five seats are recognized to be empty, the adaptation of the reference frame is provided and templates are refreshed.

Furthermore, three different template selection strategies are proposed: large template (A_1), small template (A_2) and several small templates (A_3). In particular, two small templates are taken as an example. The template selection strategies are combined with two pre-processing techniques: local normalization (B_1) and edge detection (B_2) complemented by the case of no image enhancement (B_0). In addition to combining pre-processing techniques and template selection

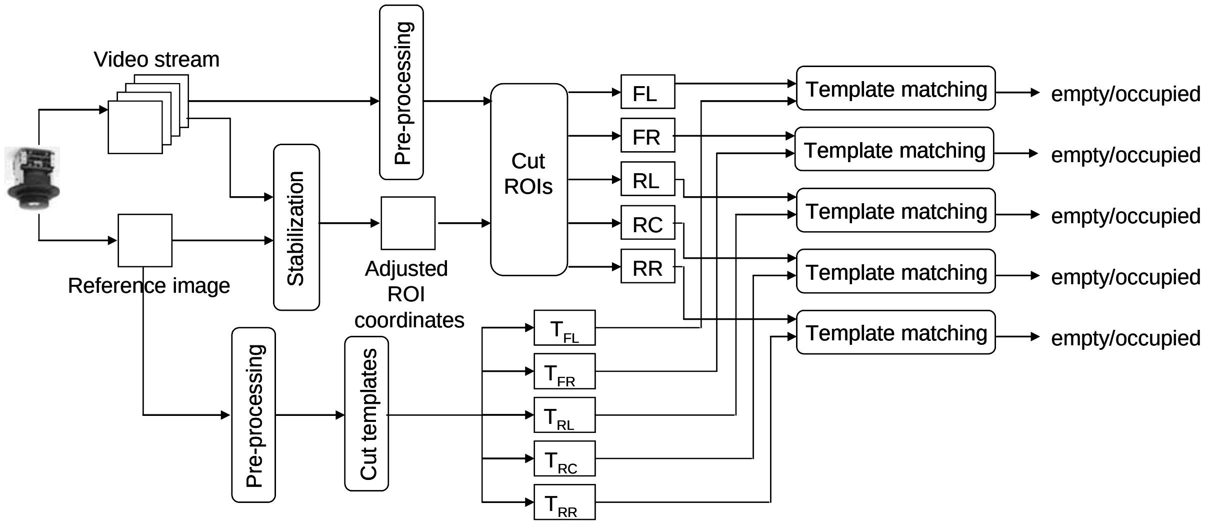


Figure 4.4: General workflow of the template matching approach

strategies, two approaches of matching score fusion are proposed: multi-algorithmic fusion and temporal fusion. Generally, each seat can be processed individually meaning that different pattern selection strategies, image enhancements and fusion methods can be used for different seats.

Frame stabilization

The stabilization has an objective to adjust the coordinates of the seat regions extracted from the reference frame so that the seat regions can be properly located in the current video stream. It is performed on the first frame from the video stream by aligning the center of the current video frame to the center of the reference frame. It is assumed that video acquisition starts after unlocking a door so that the first frame shows an empty car. The stabilization is based on template matching as well. The steering wheel pattern is taken as a template. The selection of the steering wheel is motivated by its stability, which guarantees that this pattern can be easily found even if it is shadowed or partially covered. In modern cars the position of the steering column can be adjusted by the driver. Therefore, the steering wheel pattern may have different relative positions in the reference and the current frame providing false coordinates of the seat regions. This issue is beyond the scope of this study whereas the test setup assumes the constant position of the steering column. Figure 4.5 shows the steering wheel pattern and the stabilization results in the case of the rear left seat. The coordinate shift is revealed from the matching of the steering wheel pattern and used to adjust the coordinates of the seat regions.

Template selection

Basically, there are no strict rules for template selection. Due to the lack of expert knowledge, there is no possibility to state in advance which size the template should have and where it should be located. This is why our study has solely an experimental nature and proposes three exemplary selected strategies to choose templates. It is only intended to reveal general tendencies regarding the template selection strategy but not to find out the optimal size and position of templates.

The first strategy A_1 is to select a large template having roughly the same size as the seat regions as shown in Figure 4.6. This choice helps to avoid problems connected with inaccurate stabilization. Even if the seat region is slightly shifted in the current video stream the major part of this region

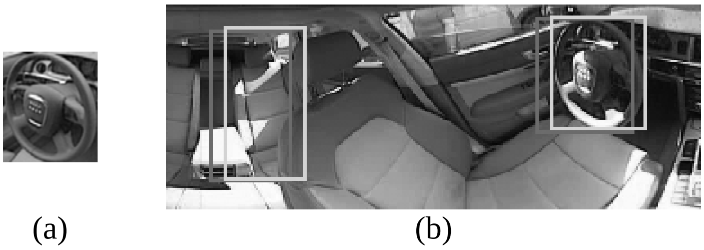


Figure 4.5: (a) Steering wheel pattern; (b) The dark squares denote the default location of the steering wheel and the rear left seat in the reference image while the bright squares denote the adjusted regions (reprinted from [160])

will match the template, surely under the condition that the seat is not occupied. In contrary, if the seat is occupied by a person or by a large object, the major part of the seat region will not match the template. However, this template selection strategy has a disadvantage of being highly sensitive to illumination changes and lighting artifacts caused non-homogeneous illumination e.g. shadows. Here, the lighting normalization during pre-processing is essential.

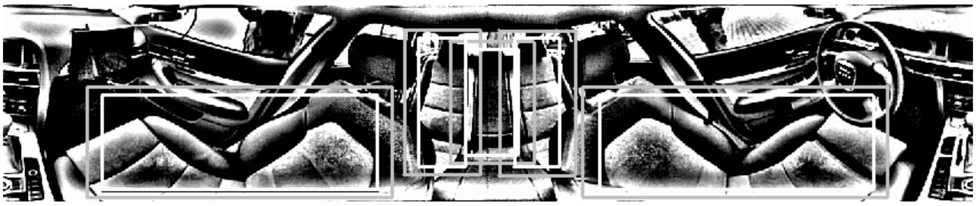


Figure 4.6: Local normalization B_1 and large template strategy A_1; seat regions in gray, templates in white (reprinted from [160])

The next strategy A_2 is to carefully select templates that are considerably smaller than the seat regions as shown in Figure 4.7. The most significant advantage of this choice is that small templates can be cut from the regions that are only slightly affected by the sun light coming from windows. If the templates are selected to be small enough, shadows will likely cover the whole template area having no harmful influence on the matching process. A disadvantage is that the template can be theoretically found on the wrong position but having a very high matching score. In order to avoid this kind of misdetection, two values generated by the template matching process are incorporated into the decision making process: the matching score S_{max} and the difference between the coordinates of the discovered pattern and the initial coordinates of the template. The difference is calculated in terms of the Manhattan distance. The mutual matching score is an average of S_{max} and the inversed distance value D.

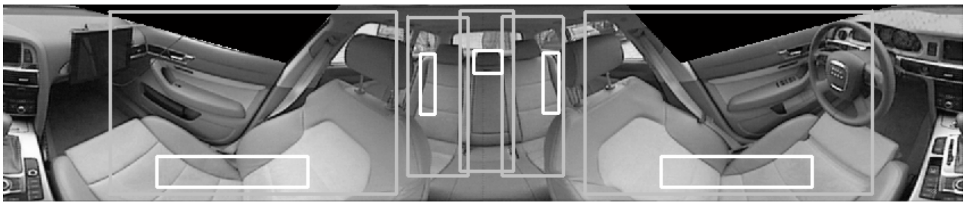


Figure 4.7: B_0 and small template strategy A_2; seat regions in gray, templates in white (reprinted from [160])

The third strategy A_3 is an extension of A_2 using two templates per seat region as shown in Figure 4.8. In this case the search process is done twice resulting in two matching scores for the corresponding templates. The mutual matching score is the maximum value of both.

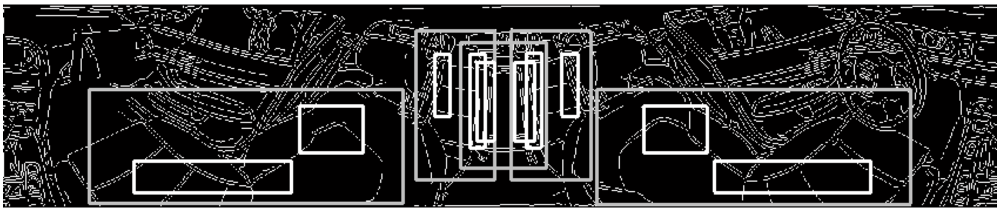


Figure 4.8: Edge detection B_2 and two small templates strategy A_3: seat regions in gray, templates in white (reprinted from [160])

Pre-processing

Pre-processing has the goal to flatten out the artifacts caused by non-uniform illumination e.g. shadows or flickering light. The proposed techniques are local normalization and edge detection. Pre-processing is applied to both templates and seat regions. The absence of pre-processing is referred to as B_0 and used as a reference. Local normalization is denoted as B_1. Figure 4.6 shows an example of applying local normalization to the image of the car passenger compartment. The formal description of local normalization and the rule to calculate the locally normalized image can be found in Section 3.2.1. The alternative image enhancement approach B_2 transforms image to edges using the Canny edge detector [30]. The algorithm is described in detail in Section 3.2.1. Figure 4.8 shows the result of applying Canny edge detector to the image of the car passenger compartment.

Matching score fusion

The fusion of matching scores is proposed to improve the robustness of template matching performed on one single frame or by one particular modification of the algorithm.

The multi-algorithmic fusion combines pre-processing techniques. In particular, for the fixed template selection strategy matching scores are produced by template matching using different pre-processing techniques. The mutual matching score is an average of the resulting matching scores. The multi-algorithmic fusion is visualized in Figure 4.9 regarding one particular seat and the current video frame i with no consideration of the temporal fusion. There are two combinations considered: B_{1+2} and B_{0+1+2}. The fusion strategy B_{1+2} combines matching scores of local normalization and edge detection subsystems by creating an average matching score. The fusion strategy B_{0+1+2} also incorporates the matching score of the reference approach with no pre-processing by also creating an average.

The temporal matching score fusion operates with an average matching score over a particular period of time. In our case eight successive frames, which correspond to the time period of approximately one second, are taken as an example. The intuition behind the temporal fusion is that it is very improbable that the algorithm makes incorrect decisions for several sequential frames. In contrary, single false decisions are possible because of overexposed frames or interlacing artifacts.

Let $S(i)$ be the matching score for the frame i of the particular seat. The fusion score $S_f(i)$ produced by the temporal fusion is given by Equation 4.2. Here, matching scores that are computed by the system for the last nf frames are replaced by their average. The frame with an index $i = 0$

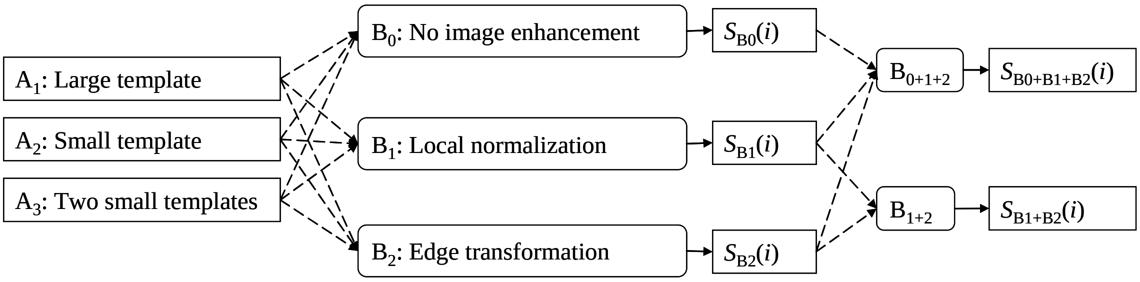


Figure 4.9: Multi-algorithmic matching score fusion

is the first frame in the sequence.

$$S_f(i) = \frac{1}{min(i+1, nf)} \sum_{k=0}^{min(i+1,nf)-1} S(i-k) \tag{4.2}$$

The score $S_f(i)$ indicates the seat occupancy state at the last frame of the sequence. The experimental video streams have the frame rate of 7.5 FPS. The default value of nf is seven which corresponds to eight frames and roughly one second.

The temporal fusion, applied to the locally normalized and edge frames, is denoted as B_1^T and B_2^T respectively. The multi-algorithm temporal fusion of both pre-processing approaches B_1 and B_2 is denoted as B_{1+2}^T.

4.1.8 Face as an identifier of an occupied seat

The detection of occupant faces is probably the most comprehensive way to determine if the vehicle seat is occupied by a person. Moreover, the location of a face can indirectly provide further information about the occupant, for instance body size or an eventual out-of-position situation. The facial image also provides information about the gender and the age of a person.

The face detection system in a car is designed similarly to the template matching for empty seat detection. After the frame stabilization, the seat regions (FL, FR, RL, RC and RR) are cut and the face detection is accomplished in each region independently. The seat regions for face detection are shown in Figure 4.10.

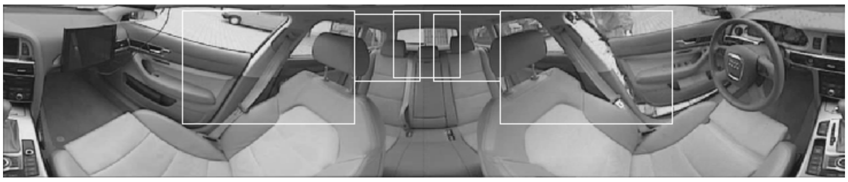


Figure 4.10: Seat regions for face detection (reprinted from [158])

Prior to face detection the general statistical model needs to be trained from face and non-face patterns. The general face detection workflow is shown in Figure 4.11. The face detection module returns an array of detected faces which are given by rectangles containing x and y coordinates of the left-top corner, width and height. A seat is considered to be occupied if at least one face is detected in the corresponding region.

The face detection algorithm can be freely chosen. The core objective of the proposed framework is to challenge well-known and easily accessible face detection algorithms for their feasibility applied

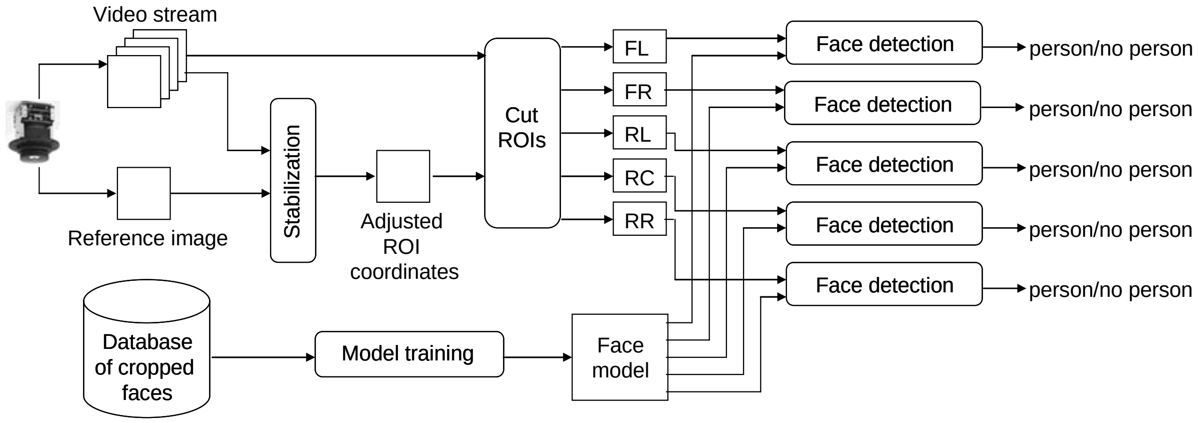


Figure 4.11: General workflow of the face detection approach

to our specific in-car images. A further objective is to find out whether face detection alone is sufficient for reliable seat occupancy detection.

Viola-Jones object detector

The object detector proposed by Viola and Jones [242] is based on three basic concepts, which makes the detection extremely rapid and robust:

- utilization of Haar-like features,
- selection of a limited set of critical features by means of boosting,
- combining critical features to the cascade classifier.

Haar-like features are named after Haar functions [98] which establish an orthonormal basis and therefore any function can be approximated by their linear combination. Haar-like features can be rapidly computed from the integral image. The integral image is an image where each pixel presents the sum of the gray values of all pixels located in the upper left position regarding the current pixel in the original image. The computation of the integral image is a very quick procedure requiring only a few operations per pixel. Each Haar-like feature can be computed quickly at any scale and location using a fixed number of operations.

Due to the fact that any sub-window of the image contains significantly more Haar-like features than the number of pixels in this sub-window [242], feature selection is necessary. Therefore, the second basic concept is the application of the adaptive boosting (AdaBoost) algorithm to select a limited set of critical Haar-like features from a very large set of potentially useful Haar-like features. For each single feature the weak classifier is trained. This classifier can be a simple threshold-based decision (stump) or a degraded decision tree (CART). In each stage of the boosting process, the best weak classifier is selected. The selection of a weak classifier implies the feature selection.

The third concept is the combining of critical features to the cascade classifier. After the learning process, the major number of the available Haar-like features is excluded and only a very limited set of important features is coupled to the classification cascade, which is called the strong classifier. Combining features into a cascade structure allows focusing exclusively on promising regions of the image rapidly excluding all irrelevant areas. In particular, all regions with a high miss-rate regarding the important features are withdrawn from consideration in the very first stages of the cascade. In contrary, the face-similar regions containing most of the critical features are passed on to the further stages of the cascade.

Notable practical work on analysis and implementation of the Viola-Jones algorithm has been done by Lienhart et al. [146, 147]. The set of Haar-like features is extended by the set of 45-degree rotated features, which can be rapidly computed in time comparable to the standard Haar-like features. In addition, the empirical analysis of adaptive boosting methods shows that a Gentle AdaBoost algorithm outperforms Discrete and Real AdaBoost algorithms in terms of detection accuracy and computational complexity. It is also shown that small decision trees used as weak classifiers, from which second and/or third order dependencies can be derived, may improve the detection performance at a detection speed comparable to that of stumps used as weak classifiers.

The Viola-Jones object detector is implemented as a part of the Open Computer Vision Library [115]. The library provides several cascades of Haar-like features for the detection of frontal and profile faces. The cascades are trained based on 20x20 pixel images. This size is stated to be optimal for face detection [146]. There are two cascades from the OpenCV package are used as a face model for the search for frontal (haarcascade_frontalface_alt.xml) and profile (haarcascade_profileface.xml) faces.

Face detection algorithm of Kienzle et al.

As already outlined in the description of the Viola-Jones algorithm, rapid and accurate object detection is performed in three basic steps. These steps include a quick extraction of meaningful features, a quick and accurate classification and an optimized image scanning method.

It is firmly established in [127] that by using SVM, the plain gray values outperform Haar wavelets and gradients in the domain of face detection, which motivates the use of gray values as features for subsequent classification. For the classification, Kienzle et al. utilize the SVM which is posed as the most feasible and even the most accurate solver of two-class problems. However, SVM is computationally expensive. Since the decision boundary is defined by a set of boundary vectors (support vectors) from both classes, the computational complexity primarily depends on the number of support vectors. Burges [28] introduces a reduced set of support vectors that approximates the complete set and therefore reduces the complexity of the decision function. This reduction accelerates the classification by the factor of 10 to 30 if applied to the domain of image classification whereby classification accuracy remains unchanged. In earlier works [28], the reduced set of support vectors is found by unconstrained optimization. Kienzle et al. impose a structural constraint on the synthetic points so that the resulting approximations can be evaluated via separable filters. This enhancement significantly decreases the computational complexity while scanning large images, which accelerates the rank deficient approximation of support vectors by the order of 4 to 6 as compared to the unconstrained computation of reduced set vectors.

The frontal face model has been trained by Kienzle et al. from 13,331 manually collected face images and 35,827 non-face images automatically cut from 206 background scenes. The size of training images is 19×19 pixels. The 1-norm soft margin SVM with a Gaussian kernel is used for training.

Face detection algorithm of Nilsson et al.

The Successive Mean Quantization Transform (SMQT) proposed by Nilsson et al. [180] for features extraction is an attempt to represent facial images insensitive to sensor variations and varying illumination. The SMQT features are extracted from local areas of an image whereby the intensities inside the area are replaced by the quantization value that is the closest to the mean intensity of the area. It is asserted that after performing SMQT, features will be identical if initial local regions have the identical structure. Images have identical structures if one image is biased by the scalar value or is gained from an original image. The number of quantization levels and the size of local regions are variable parameters that can be fine-tuned for better detection performance. The detailed description of SMQT is given in [180].

For the classification of local SQMT features, the extended Sparse Network of Winnows (SNoW) is utilized. The SNoW architecture represents a sparse network of linear units over a common pre-defined or incrementally learned feature space [256]. The learning policy is online and mistake-driven. In order to classify the test samples, the network creates lookup-tables of face and non-face patches which are updated in accordance with the Winnow update rule [150]. In order to accelerate the detection process, the SNoW classifier is split up into sub-classifiers also called weak classifiers. The weak classifiers are connected to the cascade, whereby the features included in the previous weak classifier are automatically included in the next weak classifier. In this manner, the final classifier includes all features and provides the lowest false detection rate. Depending on the required detection accuracy, the number of weak classifiers in the cascade can be fine-tuned.

Nilsson et al. have trained the face model from the set of facial images collected using a web camera. Right eye, left eye and the center point of the outer edge of the upper lip are manually marked to extract 32 × 32 pixel face patches. Approximately one million face patches are used for the training process. The training set of non-face patches involves approximately one million items which are randomly cut from a video collection containing no faces.

4.1.9 Combination of the empty seat detection and face detection

The combination of both approaches: empty seat and face detection allows not only to suppress airbag on empty seats but also to suppress it for forward-facing child seats (FFCS) and rear-facing infant seats (RFIS) and moreover to switch to low-risk deployment in case of small occupants.

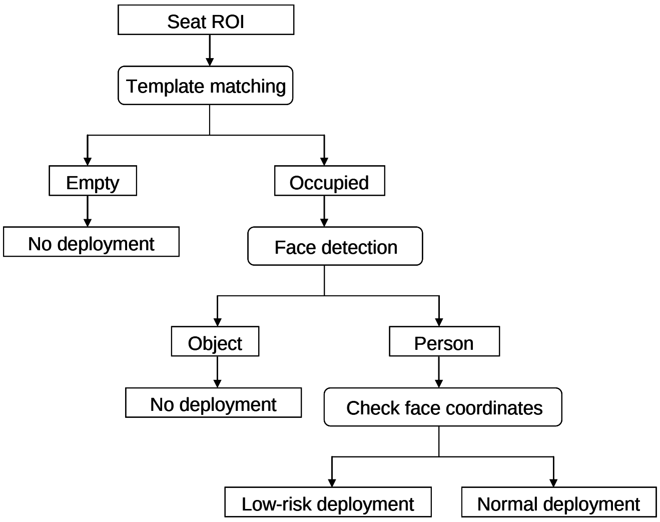


Figure 4.12: Combined approach for seat occupancy detection

The proposed approach firstly checks whether the seat is occupied using template matching for empty seat detection. If the seat is recognized to be empty, airbag deployment is prevented, otherwise the seat is classified as occupied and face detection is performed. If no face is found the seat is considered to be occupied by an object and airbag deployment is also prevented. In case a face is found, the body size of an occupant is estimated based on the face location. This process is shown in the form of an activity diagram in Figure 4.12.

4.1.10 Evaluation objectives

The experimental evaluation addresses the simplified version of seat occupancy detection identifying whether a seat is empty or not. It is endeavored to reveal which pre-processing applied to template matching provides better detection rates under uniform illumination and when non-uniform illumination occurs. Another test objective is to reveal whether face detection can substitute template matching claiming the face absence to be an indicator of the empty seat. For this purpose three well-known face detection algorithms are challenged with same test samples as template matching. Combining of template matching and face detection remains the issue of future studies.

4.2 Visual distinguishing between driver and front-passenger hands

This section is comprised of the following subsections:

1. Introduction
2. Application scenario
3. Operational scenario
4. Acquisition system
5. Illumination
6. Computer vision approaches
7. Image stabilization
8. Motion-based hand segmentation
9. Texture-based hand segmentation
10. Hand detection
11. Evaluation objectives

4.2.1 Introduction

The idea to introduce a visual user discrimination system is derived from the VisMes project as a meaningful addition to seat occupancy detection showing the versatility of an omni-directional camera for the observation of a car passenger compartment [108]. Here it is shown that the same camera can be used to maintain both systems simultaneously. However, the real car used in the VisMes project is utilized outdoors implying uncontrolled lighting conditions. In order to simulate particular lighting conditions and study the reaction of the system to these, the user discrimination system has been completely reproduced in the car simulator (further referred to as AMSLator). AMSLator includes a wooden frame with an integrated genuine dashboard, front seats, rear bench seat, touch screen mounted to the center console, a monochrome camera and two near-infrared lamps.

The development of the visual user discrimination system is done in accordance with the methodology proposed in Chapter 3 and summarized in Figure 4.13. A designing of the system is started with the introduction of application and operational scenarios. Then the appropriate hardware components such as acquisition device and illumination modules are selected. Based on the system in place, computer vision algorithms are chosen.

4.2.2 Application scenario

The user discrimination system is designed to distinguish between driver and front seat passenger hands during the interaction with the dual-view touch screen. The experimental car of the VisMes

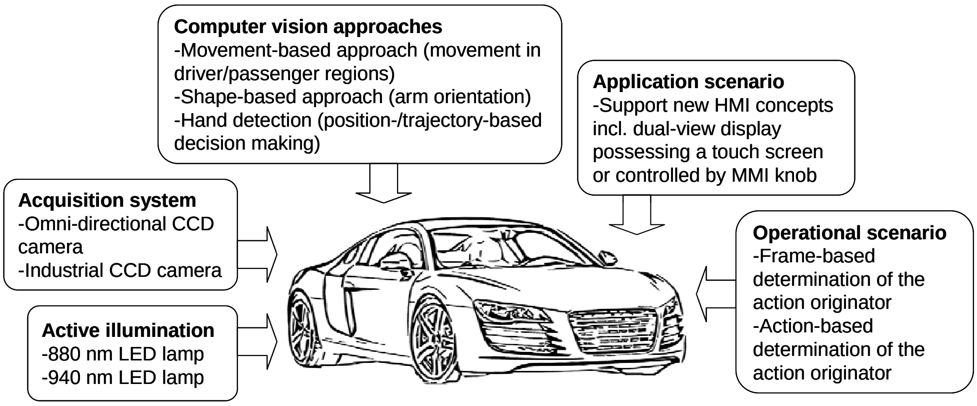

Computer vision approaches
-Movement-based approach (movement in driver/passenger regions)
-Shape-based approach (arm orientation)
-Hand detection (position-/trajectory-based decision making)
Application scenario
-Support new HMI concepts incl. dual-view display possessing a touch screen or controlled by MMI knob
Acquisition system
-Omni-directional CCD camera
-Industrial CCD camera
Operational scenario
-Frame-based determination of the action originator
-Action-based determination of the action originator
Active illumination
-880 nm LED lamp
-940 nm LED lamp

Figure 4.13: Components of the visual user discrimination system

project has no dual-view touch screen. Hence, the initial motivation of user discrimination was slightly different. The focus was on center console buttons and a MMI knob. Several buttons (e.g. air conditioning controls) are duplicated for a driver and passenger. The new ability to distinguish a between driver and passenger permits reduction of the number of buttons and thus lowering the manufacturing cost. User discrimination becomes an indispensable feature for the MMI controller which is, similarly to the dual-view touch screen, designed for cooperative usage.

4.2.3 Operational scenario

The camera has to permanently observe the center console area and the area between a driver and passenger to indicate motion toward the ICP display. The decision whether a driver or passenger is interacting with a touch screen (or MMI knob) must be made during the contact. The user discrimination system may receive a signal indicating the contact from a controller area network (CAN) bus. The decision is actually made at a particular moment of time for the current frame, but it can be based on the sequence of several previous frames. In fact, the system needs to start analyzing images before the contact occurs in order to track the moving hand. If the system is unable to choose between a driver and passenger, the previous selection has to be taken. The user discrimination is obsolete if a car is occupied only by a driver, so the system needs to work only if both front seats are occupied. It is assumed that a driver and passenger do not interact with the ICP simultaneously. In fact, dual interaction with the touch screen is possible, but the expected reaction of the system in this case is ambiguous. Moreover, the analysis of overlapping hands is beyond the scope of this work. The system has to be able to work day and night. This is why the color-based skin segmentation is not suitable for automotive applications. Furthermore, user discrimination should not be harmfully influenced by various pieces of clothes and gloves.

4.2.4 Acquisition system

The acquisition device of the first version of the user discrimination system is the same as used for seat occupancy detection, namely a Sony RPU-C3522 camera with a CCD sensor and a 360-degree annular lens, mounted on the ceiling next to the windshield of an Audi A6 Avant. The IR-filter is removed to enable capturing of NIR images. Frames are captured at a rate of 7.5 FPS with a resolution of 720 × 576 pixels. A camera frame consists of two parts: the bottom part shows a 360-degree view of the complete passenger compartment and the top part shows an enlarged section of the center console (See Figure 4.14a). The bottom part is a zoomed out version of those used for occupancy detection.

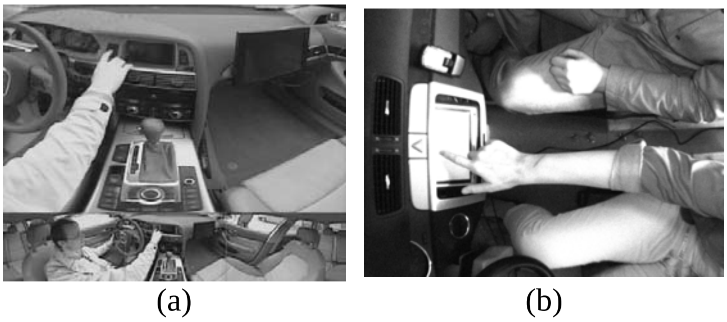


Figure 4.14: (a) Audi A6 Avant from the VisMes project; (b) AMSL car simulator

The second version of the user discrimination system makes use of an industrial camera (Imaging Source DMK 31BU03.H) mounted in the same position but facing the area between a driver and passenger. The position and the viewing angle are specified to capture the center console together with the arms of a driver and passenger. An exemplary camera image can be obtained from Figure 4.14b. The camera has a monochrome CCD sensor and a standard 8mm lens. The camera has no IR-filter, thus the whole spectral range from 400 nm to 1 μm can be acquired. The sensibility diagram is shown in Figure 4.15. The camera operates at a rate of 30 FPS with an image resolution of 1024 × 768 pixels.

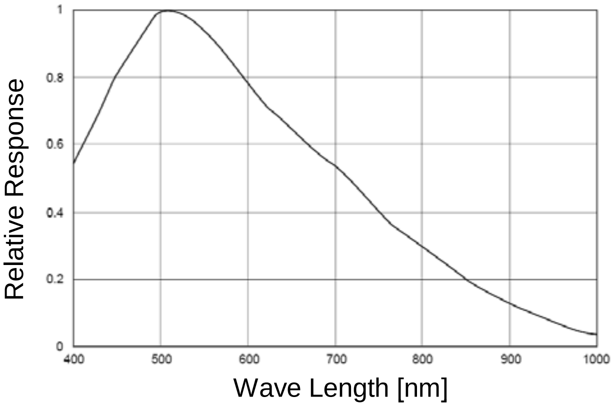
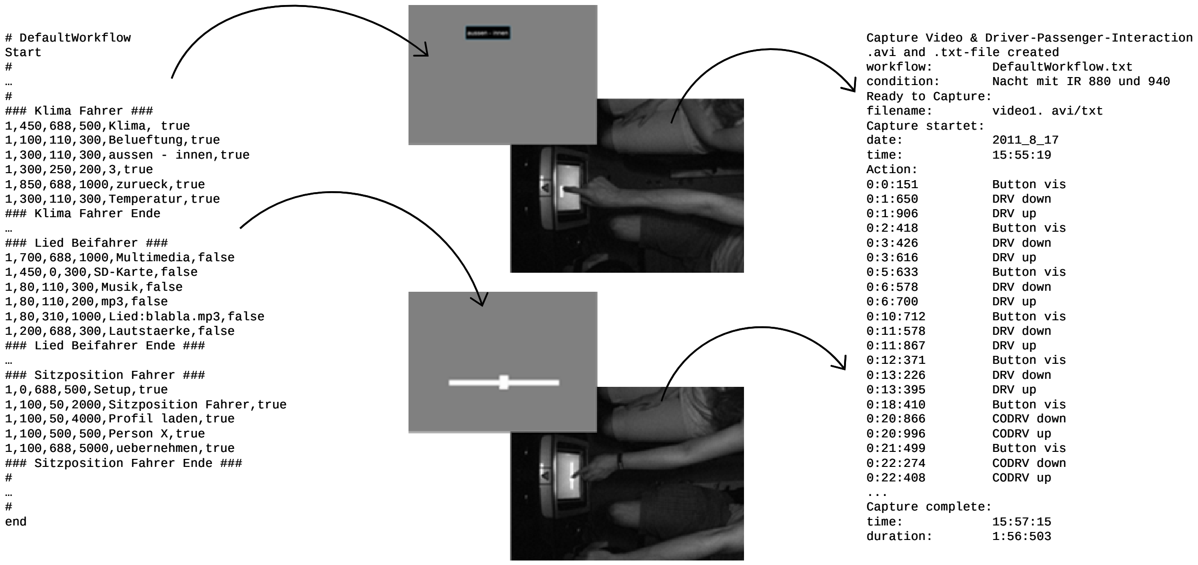


Figure 4.15: The spectral sensibility of the CCD sensor installed in an Imaging Source DMK 31BU03.H camera

An objective pursued with the car simulator is the development of the credible evaluation concept. For this purpose the touch screen has been integrated into the center console and synchronized with the acquisition system. Thus, each interaction with the touch screen is automatically stored and used as the ground truth for the identification of interactions.

The acquisition is organized as follows. After the start, interaction items are consecutively shown on the touch screen in accordance with predefined protocol. The event sequences correspond to real interactions (e.g. with a radio or navigation). At the beginning a driver and front seat passenger align their seat positions. Then a driver adjusts the air conditioning and both driver and passenger adjust the seat heaters. Afterwards a driver assigns a route in the navigation system. Finally, a passenger selects a song in the infotainment system and adjusts the volume. Two types of interaction items are used: buttons and slide bars. In order to let the system know who is currently interacting with the touch screen drivers have been instructed to touch only green items and passengers only red items. All touch screen interactions are synchronized with the video stream and stored in the log file along with the video. Events are documented with the timestamp regarding the recording start. Figure 4.16 illustrates the acquisition workflow and shows examples of interaction items as well as cuts from the events protocol and the log file.



```
# DefaultWorkflow
Start
#
…
#
### Klima Fahrer ###
1,450,688,500,Klima, true
1,100,110,300,Belueftung,true
1,300,110,300,aussen - innen,true
1,300,250,200,3,true
1,850,688,1000,zurueck,true
1,300,110,300,Temperatur,true
### Klima Fahrer Ende
…
### Lied Beifahrer ###
1,700,688,1000,Multimedia,false
1,450,0,300,SD-Karte,false
1,80,110,300,Musik,false
1,80,110,200,mp3,false
1,80,310,1000,Lied:blabla.mp3,false
1,200,688,300,Lautstaerke,false
### Lied Beifahrer Ende ###
…
### Sitzposition Fahrer ###
1,0,688,500,Setup,true
1,100,50,2000,Sitzposition Fahrer,true
1,100,50,4000,Profil laden,true
1,100,500,500,Person X,true
1,100,688,5000,uebernehmen,true
### Sitzposition Fahrer Ende ###
#
…
#
end
```

```
Capture Video & Driver-Passenger-Interaction
.avi and .txt-file created
workflow:         DefaultWorkflow.txt
condition:        Nacht mit IR 880 und 940
Ready to Capture:
filename:         video1. avi/txt
Capture startet:
date:             2011_8_17
time:             15:55:19
Action:
0:0:151           Button vis
0:1:650           DRV down
0:1:906           DRV up
0:2:418           Button vis
0:3:426           DRV down
0:3:616           DRV up
0:5:633           Button vis
0:6:578           DRV down
0:6:700           DRV up
0:10:712          Button vis
0:11:578          DRV down
0:11:867          DRV up
0:12:371          Button vis
0:13:226          DRV down
0:13:395          DRV up
0:18:410          Button vis
0:20:866          CODRV down
0:20:996          CODRV up
0:21:499          Button vis
0:22:274          CODRV down
0:22:408          CODRV up
...
Capture complete:
time:             15:57:15
duration:         1:56:503
```

Figure 4.16: The acquisition workflow in the AMSLator: the display shows the predefined sequence of interaction items, users interact with the touch screen and the resulted events are stored in the log file along with the video stream (reprinted from [107])

4.2.5 Illumination

The illumination system of the first version of the user discrimination system was discussed in Section 4.1.4. It includes four NIR modules in the top section of the door frames and one in the center of the ceiling. The second version of the system implemented in AMSLator includes two NIR lamps mounted on the ceiling very close to the camera. The lamps emit the wavelengths of 880 nm and 940 nm correspondingly.

4.2.6 Computer vision approaches

For the first version of the user discrimination system the straightforward method is suggested, namely motion estimation in particular image regions. One region is assigned to a driver and another one to a front seat passenger. If the amount of motion in one or another region exceeds a predefined threshold the corresponding occupant is considered to interact with a control element. The shape of a moving arm is derived from difference images. Such an image is the result of subtracting successive frames in a video stream. If the difference between images is not significant the preceding frame from the stream is applied to the subtraction. The difference images are converted to the binary form using an adaptive threshold. The binary images undergo morphological operations to improve the arm silhouette. The number of pixels belonging to the arm located in the driver or passenger region indicates the amount of motion and consequently the active person.

In the second version of the system the motion-based approach is complemented by hand

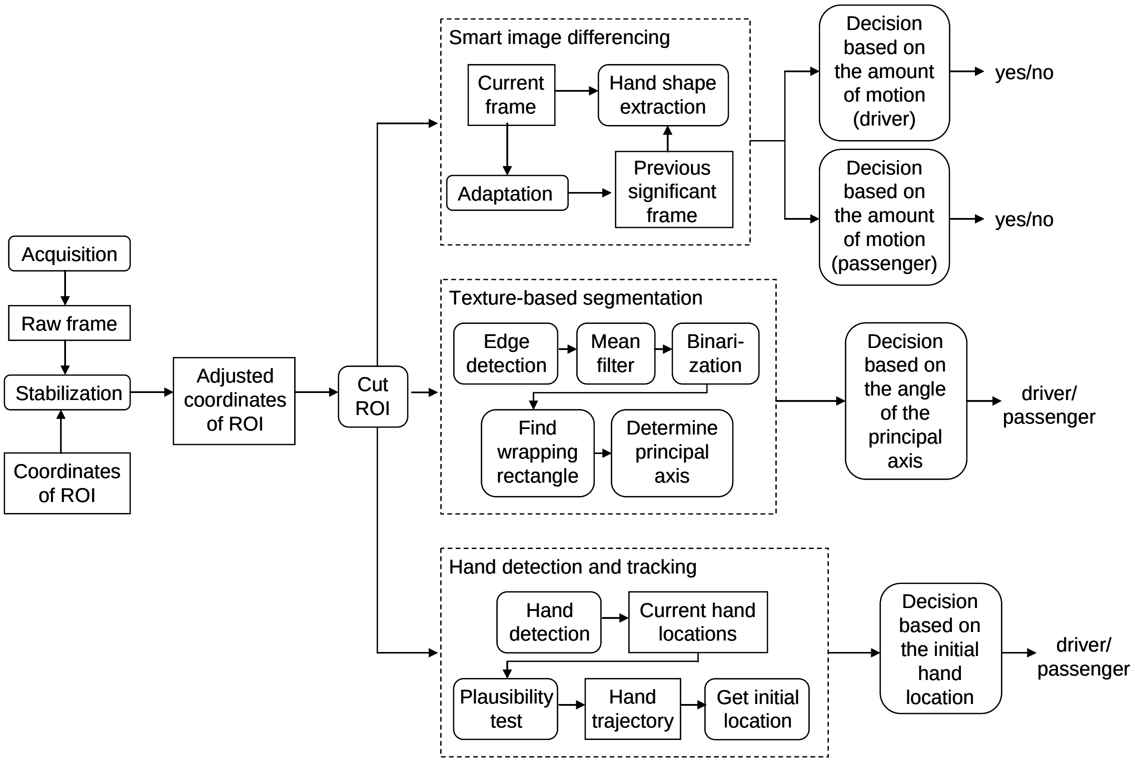


Figure 4.17: Image processing pipeline for motion-based user discrimination during interaction with center console

segmentation based on texture information. A hand is assumed to be brighter than the background under NIR illumination due to the reflection properties of the skin and the fact that an arm is closer to the camera. Camera images are transformed to edge images using the Sobel operator (see Subsection 3.2.1). The mean filter [86] is applied to suppress noise. The resulting images undergo binarization with a fixed threshold. The principal direction of the binary silhouette of the hand points out to the active occupant.

Further development of the user discrimination system is based on hand detection. The localization and tracking of driver and passenger hands is done thanks to the Viola-Jones object detector. The initial location of the active hand which is derived from the hand trajectory clearly indicates the actor. Figure 4.17 illustrates the general workflow of the user discrimination system including all three approaches.

4.2.7 Image stabilization

The stabilization of frames in the video stream is performed to determine the shift between the current frame and the reference frame and to readjust the coordinates of the mask taken from the reference image so that the mask matches the position of the center console area in the current frame. This ensures that the masking process cuts exactly the required region of interest (ROI) for subsequent processing. In videos from the VisMes project this ROI is the zoomed-in center console region as shown in Figure 4.18.

The offset in ROI coordinates is derived from the template matching as it done in Section 4.1.6. For the Audi A6 passenger compartment the template is the image of the display (see Figure 4.19a) and for the AMSLator the template is the image of the emergency button together with the wheel

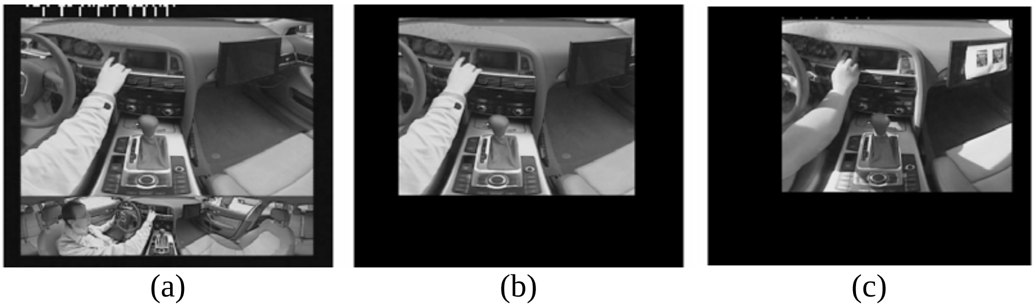


Figure 4.18: (a) Raw camera frame; (b) Masking of irrelevant area; (c) Stabilization effort during masking (reprinted from [108])

regulator to open or close the ventilator (see Figure 4.19c). The usage of these patterns is motivated by their stable appearance under various illuminations. Hence, the templates can be easily found even when the areas are shadowed. Moreover, these areas are very seldom covered.

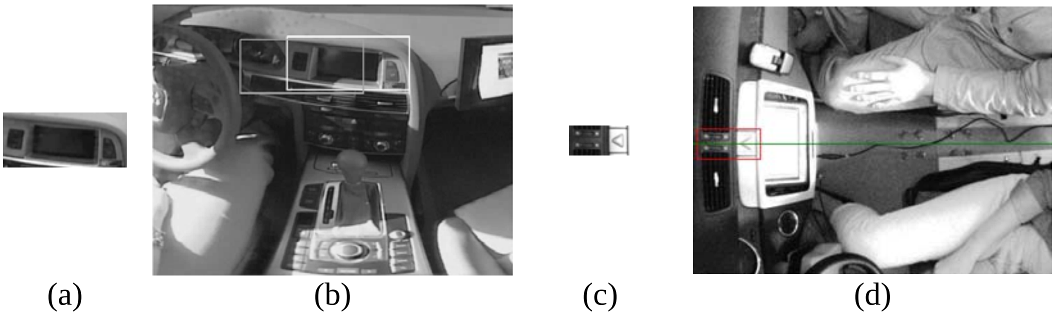


Figure 4.19: Template matching for the frame stabilization in both modifications of the user discrimination system: (a) stabilization template for the Audi A6; (b) zoomed-in center console region after stabilization, the gray rectangle would be in the position of the white rectangle if no stabilization is performed; (a) stabilization template for the AMSLator; (d) ROI after stabilization

While the stabilization result of the first setup (Audi A6) is the zoomed-in center console region with the display in the middle of the image regarding the x-coordinate, the stabilization result of the second setup (AMSLator) is the original camera frame that is slightly cropped so that the driver and passenger regions have exactly the same size, whereby the division line is defined by the template as shown in Figure 4.19d.

4.2.8 Motion-based hand segmentation

Motion-based hand segmentation is a simple and straightforward way to estimate the activity of a driver and a passenger and then to select the more active one. However, the motion-based approach relies on a premise that a hand is the only moving object within a center console region. Hence, such an algorithm can be easily deceived by dynamic shadows often occurring in moving vehicles. Nonetheless, in the first modification of the user discrimination system all experiments are provided in the standing car so that dynamic shadows can be caused only by moving instances inside of the vehicle. Note that in the motion-based system a moving shadow is the main source of false decisions.

Motion-driven hand shape extraction is based on the calculation of difference images. The difference image is the absolute difference between the current frame and the reference frame which is initialized by the first frame from the video stream. In order to address the varying illumination the reference frame needs to be permanently adapted by its replacement to some fresh image from the stream. Since a difference image is expected to represent silhouettes of moving objects, the reference image needs to have significant difference to the current one. So, in case fast movement occurs, the previous frame from the stream can be used as the reference. Contrary to this, in case of relatively slow motion the previous frame exhibits very low difference to the current one. This is why the frame buffer of eight sequential frames is collected and permanently updated by withdrawing the least recent frame and adding the current frame. So, if the last frame in the buffer exhibits significant difference to the current one, the last frame is taken as the reference. Otherwise, the less recent frame from the buffer is checked for a significant difference and if it matches, this frame is used as the reference. If no frame from the buffer fits this condition, the absence of motion during the time period of one second is indicated. In this case the reference frame is reinitialized by the first frame.

The difference images are converted into binary representations based on the adaptive threshold. The adaptive threshold is calculated based on the Otsu algorithm [187] and has a drawback if the difference image is flat, meaning that the smoothed histogram does not have well-defined peaks but only a couple of local minima. In this case the threshold is often erroneously selected and the whole image becomes constant. However, by ensuring that the significant difference is presented, this situation becomes very improbable.

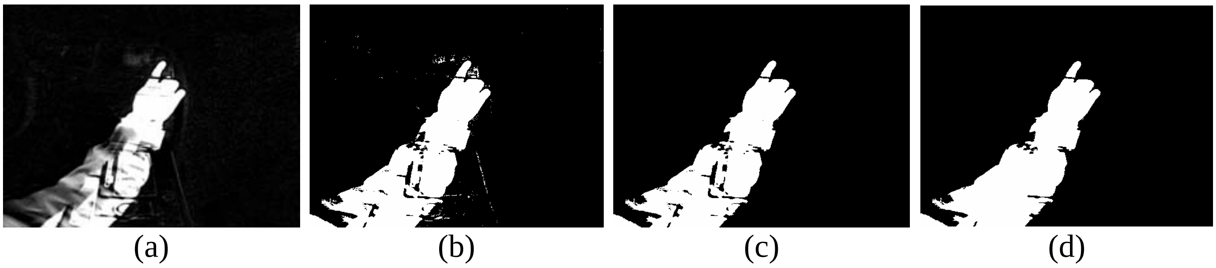


Figure 4.20: Motion-based hand segmentation: (a) difference image, (b) image after binarization, (c) after area opening, (d) after filling the gaps (reprinted from [108])

In order to eliminate noise and fill gaps in the hand silhouette represented on the binary image, two morphological operations are applied: ‘area open’ and ‘fill’ [86]. The area open operation removes all regions smaller than 400 pixels. The fill operation finds and removes isolated gaps inside of the silhouette. Figure 4.20 illustrates the hand-shape extraction process.

Based on the experimental data, two regions are determined. The first region is assigned to a driver and the second to a passenger. If a driver or a passenger interacts with a console element, the hand silhouette immediately appears in the corresponding region. Due to the region selection, it is very unlikely that the passenger’s hand will appear in the driver’s region and vise versa. Motion in a region is indicated if at least one active pixel appears. Figure 4.21 shows the regions for driver/passenger discrimination together with some examples of segmented hand silhouettes.

Note that the system does not directly distinguish between driver and passenger hands, but only detects the motion on driver and passenger sides. This motion is then interpreted as an interaction with the center console. Frame-by-frame decision making is performed separately for driver and passenger, thus facilitating a differentiation between four cases of interactions: driver interacts, passenger interacts, both interact, no one interacts. A system that only distinguishes between driver

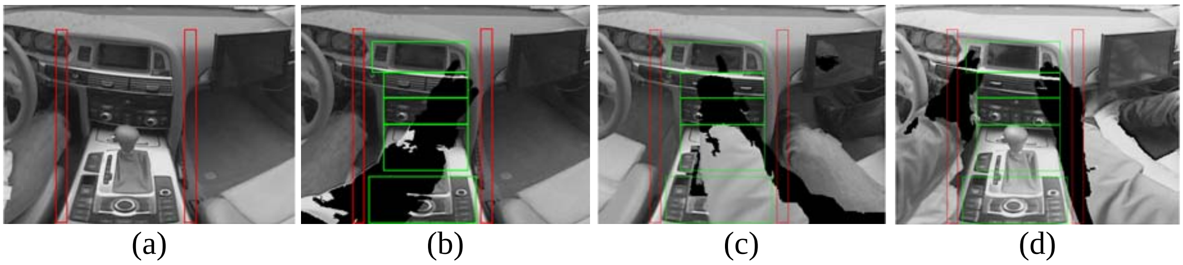


Figure 4.21: (a) Regions to determine driver/passenger interaction; (b) Driver touches a console button; (c) Passenger handles navigation touch screen; (d) Driver and passenger touch center console buttons (reprinted from [108])

and passenger is not able to address third and fourth cases. In addition, the proposed algorithm is able to detect the region in which the handled console element is located. These regions are shown in green in Figure 4.21.

4.2.9 Texture-based hand segmentation

Further development of the user discrimination system is carried out in the AMSLator. In order to simplify the user discrimination task and to avoid any ambiguities, it is assumed that a driver and passenger do not interact with the center console simultaneously. Moreover, it is switched from frame-based to action-based decision making which makes third and fourth cases from the previous paragraph obsolete.

The discrimination algorithm combines the motion-based approach with the texture-based hand segmentation. Each of the subsystems evaluates a current action and yields a matching score. A matching score is a value within the interval [-1, 1]. The score -1 means that an action is initiated by a driver. The score 1 indicates an action by a passenger. The values within the interval show a tendency of an algorithm to decide for a driver or passenger. Zero indicates the situation in which the algorithm is unable to make a decision.

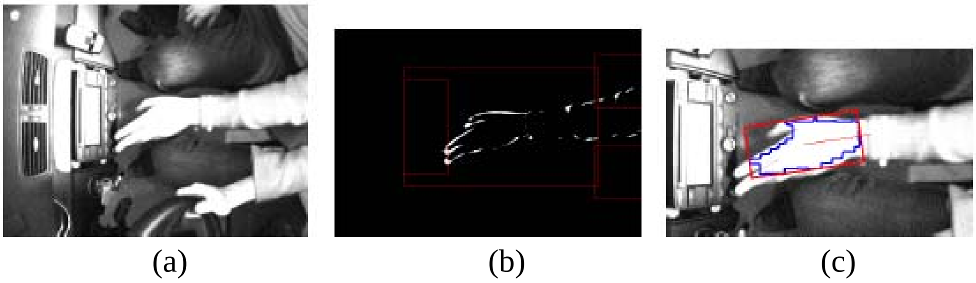


Figure 4.22: AMSLator screenshots: (a) original camera frame; (b) difference image; (c) edge and direction analysis (reprinted from [159])

The motion detector calculates the number of active pixels in driver and passenger regions which are denoted as H_d and H_p respectively. The motion identifier s_1 is given by Equation 4.3. An example of the difference image is shown in Figure 4.22b.

$$s_1 = 1 - \frac{2H_d}{H_d + H_p} \tag{4.3}$$

The second algorithm exploits the hand reflection under NIR leading to the fact that the hand exhibits strong brightness difference to background when it is stretched toward the center console. The segmentation begins with the edge detections using the Sobel operator. Afterwards, the mean filter is applied to reduce the intensity of gray values in small regions which are supposed to be not a part of a hand but rather elements of noise. Then the image is converted to the binary form using a fixed threshold. The threshold is revealed from test samples. The resulting binary image represents a series of disconnected blobs. In order to connect the blobs and to obtain the silhouette of the hand, the morphological opening is performed. Then the 'find contours' operation is applied to obtain a closed polygon of the detected object. If several polygons are detected, the filtering of small polygons is performed until only the biggest polygon remains. The principal direction of the minimal bounding box around the obtained polygon points to the occupant interacting with the center console. The result of the described processing is visualized in Figure 4.22c. The direction identifier s_2 describes the angle between the principal axis of the bounding box d and the horizontal line h and is given by Equation 4.4. The deviation of 45 degrees from h is considered to be maximal so that the angle rises in the interval $[-\pi/4, \pi/4]$ and the s_2 become values from -1 to 1.

$$s_2 = \tan \angle(d, h) \tag{4.4}$$

Finally, the identifiers $s_i, i = 1, 2$ are fused to the mutual score by means of forming an average. The mutual score is then compared with the threshold τ. Exceeding τ designates that the system chooses a passenger, otherwise a driver is chosen.

4.2.10 Hand detection

The motion-based approach and the angle-based descriptor do not lead to satisfactory user discrimination. The third alternative proposed here is the direct search of hand patterns in gray-scale camera images.

The most simple and plausible way for a pattern search is the template matching. Firstly, the reference pattern is defined. Then, this pattern is searched for using the sliding window approach. Namely, the window of a particular size slides across the image and the window content is compared with the reference. The correlation coefficient is usually used as a similarity measure. If the correlation value exceeds a predefined threshold, the pattern is considered to be found. However, this technique has serious limitations in case of highly variable appearance of the searched pattern caused by changing illumination. In this case, several templates have to be used meaning that template matching evolves to the machine learning.

The Viola-Jones algorithm [242] is an example of an effective machine learning approach for object detection. As already mentioned three concepts form the basis of this algorithm: Haar-like features rapidly computed from integral image representation, feature selection by means of adaptive boosting, and combining features into a cascade structure to focus on promising image regions and to rapidly exclude irrelevant regions.

In order to train a cascade of Haar-like features for hand detection, the set of 1645 positive samples (hands) and the set of 3049 negative samples (non-hands) are collected from training videos. Figure 4.23 shows several examples of positive and negative samples. Following the recommendations of Lienhart et al. [146] the size of positive images is fixed to 20x20 pixels and the Gentle AdaBoost algorithm is utilized for the training of the tree-based classification cascade with 15 stages. The resulting classifier is able to detect hands stretched toward the center console.

As mentioned in Section 4.2.2 the user discrimination system must operate in real time meaning that the decision whether a driver or a front seat passenger is currently touching the display has to be made during the contact. In a real car the notification that the contact has happened and the

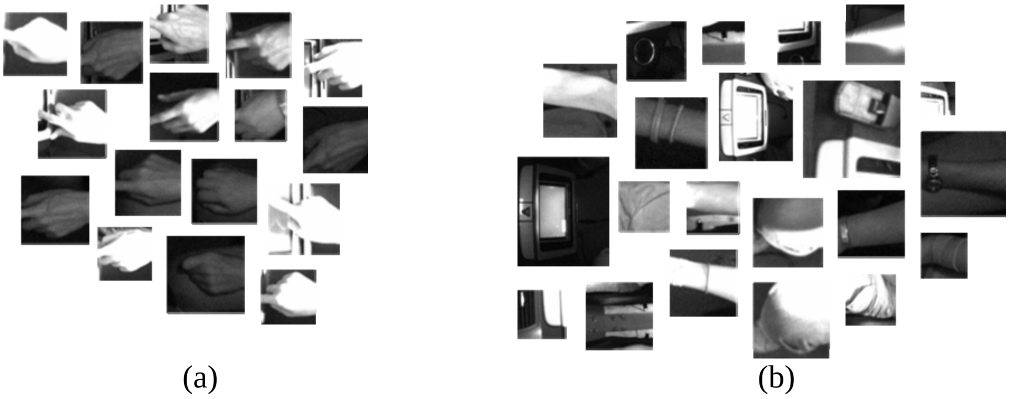


Figure 4.23: Training samples for the Viola-Jones hand detector: (a) positive, (b) negative (reprinted from [107])

judgment of the discrimination system is required comes from the CAN bus. In the car simulator this information is provided by logs collected during data acquisition. Logs include timestamps of three events: display element becomes visible, hand down and hand up. Thanks to logs the action-based processing and evaluation of interactions is possible.

Hand detection starts when a display element (button or slider) appears on the touch screen and finishes during a 'hand down' event. The decision has to be provided before a 'hand up' event. So the trajectory of the hand detected in the frame sequence between two events is obtained. The positions of the first reliably detected hand and the last detected hand determine the directional vector pointing to the active occupant. In fact, the y-coordinate of the first detected hand regarding the central horizontal line is sufficient to identify the actor. The formal description of the direction identifier s_3 is given similarly to s_2 as a tangent of the angle between the vertical line and the line stretched from the center of the first hand region to the center of the last hand region. If both hands are detected the distances between the center of the current hand region and the center of the touch screen is calculated. The occupant with the shortest distance is determined as the interacting one. Figure 4.24 visualizes this idea.

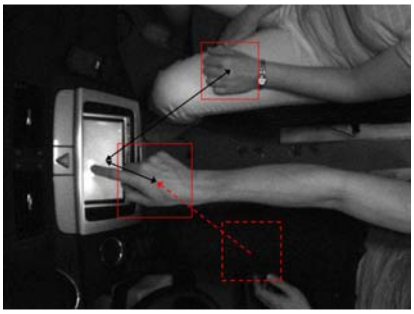


Figure 4.24: Distance calculation between the touch screen and the driver/front seat passenger hand (reprinted from [107])

In order to reduce the number of misclassifications, caused by the regions falsely detected as a hand, the motion-based approach calculating the amount of motion in driver and passenger regions is applied prior to position-based decision making. In fact, false detections are usually brought about by stable objects with an appearance similar to a hand. Therefore, these regions are found in the same location and there is no additional movement committed to these detections. Contrary to this, hands moving toward the touch screen cause additional movement in the detected regions and their surroundings. Thus, all stable hands are filtered as probable misdetections. In case no decision can be reliably made based on the hand position, the amount of motion is used as a principal factor for user discrimination. If the amount of motion in driver and passenger regions is similar and no hand has been detected the system takes the decision made for the previous event. The decision making process is illustrated in Figure 4.25. Note that three different motion thresholds τ_1, τ_2 and τ_3 are used depending on the result of hand detection.

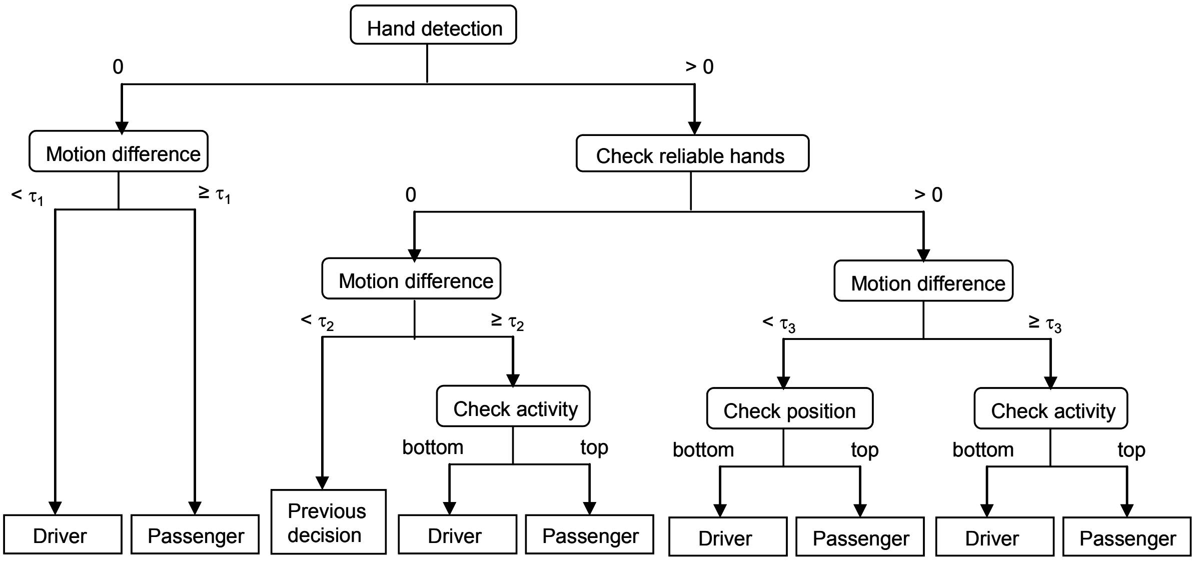


Figure 4.25: User discrimination based on motion and hand detection

4.2.11 Evaluation objectives

Since the dynamic non-uniform illumination is the main factor drastically influencing the recognition performance of all computer vision systems, evaluation concentrates on varying illumination and identifies changes in error rates of the proposed algorithms (motion-based hand segmentation, texture-based hand segmentation and hand detection). Three tests are suggested. The first test examines the necessity for active NIR illumination during daytime. The second test provides an answer to the question, whether the algorithms have the same recognition performance at different times of day (day, twilight, night). The third test covers strong lateral illumination evoking overexposure and dynamic shadows.

4.3 Facial driver recognition

This section is comprised of the following subsections:

1. Introduction
2. Application scenarios
3. Operational scenarios
4. Acquisition system
5. Illumination

6. Computer vision approaches (appearance-based vs. feature-based)
7. FaceART: face recognition system
8. Evaluation objectives

4.3.1 Introduction

Biometric driver recognition is an undisputedly valuable addition to the functionality of an automobile. It can be used in various application scenarios which demand the individual adjustment of automotive systems. Three examples of such applications are presented in Figure 4.26 together with other components of a facial recognition system within a vehicle. The proposed components reflect the designing methodology introduced in Chapter 3 and also include operational scenarios, an acquisition device, illumination modules, and face recognition algorithms.

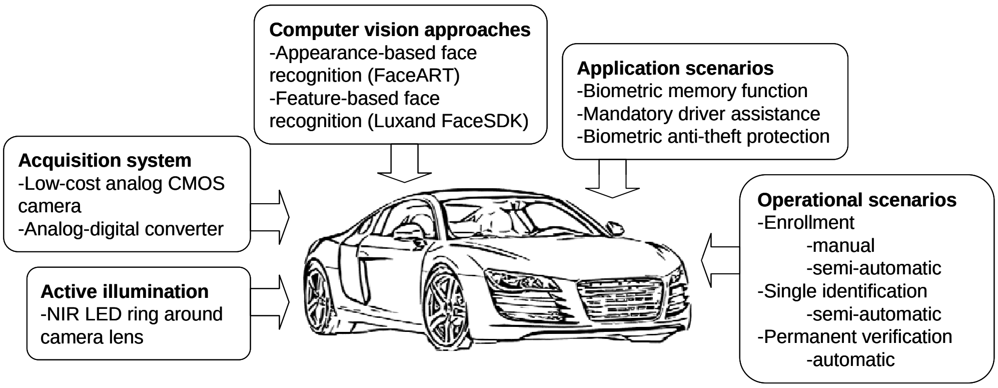

Computer vision approaches
-Appearance-based face recognition (FaceART)
-Feature-based face recognition (Luxand FaceSDK)
Application scenarios
-Biometric memory function
-Mandatory driver assistance
-Biometric anti-theft protection
Acquisition system
-Low-cost analog CMOS camera
-Analog-digital converter
Operational scenarios
-Enrollment
-manual
-semi-automatic
-Single identification
-semi-automatic
-Permanent verification
-automatic
Active illumination
-NIR LED ring around camera lens

Figure 4.26: Components of the facial driver recognition system

4.3.2 Application scenarios

Three application scenarios are proposed to improve the comfort, safety and security of a vehicle. The comfort improvement is achieved due to a biometric memory function implying that after driver identification, the seat, steering wheel, rear mirrors, air conditioning and car radio are automatically adjusted to a driver. Safety is gained through the mandatory driver assistance implying that after the driver is identified, the engine power is restricted and driver assistance systems are compulsively switched on for young and inexperienced drivers. Security increases thanks to biometric anti-theft protection implying that the car ignition is blocked until a registered driver is identified.

Regarding driver authentication, the proposed applications differ only regarding the moment when the authentication happens. The biometric memory function and the mandatory driver assistance require single driver identification directly after getting into a car. The biometric anti-theft protection (also referred to as biometric immobilizer) requires not only driver identification directly after getting into a car but also periodic driver verifications when the car stops and any door has been opened and closed again to prohibit driver substitution by an intruder. In case the verification fails, the security system has to warn an adversary driver, force him to stop the car and then block the engine.

A personal car is usually driven by family members and has, therefore, in most cases a very limited number of drivers. In terms of biometrics it is called a low-scale application. Defining the maximum number of drivers to be six, we keep space for a large family. Generally, there are three major factors reducing the performance of face recognition systems: varying face rotation, non-uniform illumination and short-term face alteration. The first factor can be controlled by an acquisition scenario while the next two factors have to be compensated by a face recognition algorithm.

4.3.3 Operational scenarios

Each biometric application consists of two principal stages: enrollment and authentication. During enrollment the reference data of potential users is collected and stored in a database (further referred to as reference database). During authentication an unknown biometric sample is matched against reference data aiming at determining the originator of this sample.

Regarding enrollment, several important questions arise:

- When and where is the registration of a new driver allowed?
- Should the registration be supervised and if yes who is admitted as supervisor?
- How many and what camera frames should be captured to reliably form a face model and can a single camera frame be sufficient?

Regarding authentication, those questions are:

- When does a driver have to be authenticated?
- Only once or repetitively?
- In fully automatic or in interactive mode?
- Which recognition performance can be considered sufficient?

The practical implementation of a driver recognition system implies answering all these questions.

Secure vs. convenient enrollment

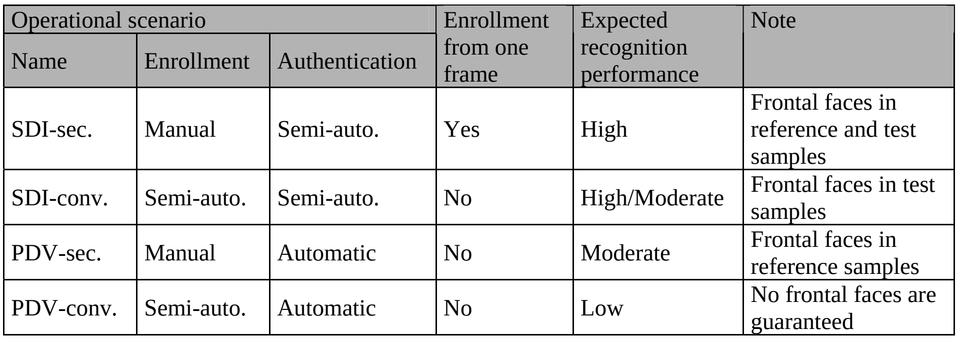
In order to answer the first and second enrollment questions, two security levels of driver enrollment are proposed: high-level security and low-level security. The *high-level security* permits the enrollment exclusively at a trustworthy location for instance an authorized sales office or car repair shop supported by a qualified/certified salesman or repair shop mechanic. A supervisor instructs a new driver to take a required position for reference shots and manually controls their quality. Since the driver registration precedes the first ride, at least one driver (e.g. car owner) needs to be registered directly after the car purchase in a sales office evoking no limitation regarding the convenience. However, the enrollment of further drivers requires visiting an auto repair shop significantly reducing the convenience. Nonetheless, this restriction guarantees the high quality of reference data and, therefore, high reliability of subsequent driver authentication. The *low-level security* implies sacrificing security to increase convenience, namely permitting the registration of new drivers any time and everywhere. The car owner or even any registered driver gets admission to start and manage the enrollment. In this case the level of security is different depending on how many persons are admitted to start the enrollment. The supervisor must be aware of security risks committed to the penetration of wrong or low quality images into a biometric template and must, therefore, manually filter the reference data. Another security risk is that an admitted driver can be forced to start the enrollment of an adversary.

The enrollment can be provided in manual or semi-automatic modes. While the manual registration requires that a new driver be instructed by a qualified supervisor to provide required actions (e. g. head rotations or head inclinations) for camera shots and manually controls the quality of captured images, the semi-automatic or interactive registration implies that the car instructs a new driver to take an appropriate position for a shot and automatically controls image quality. The automatic mode implies that the registration is done transparently so that drivers do not even get noticed. In fact, there is no need for automatic enrollment because this process has to be initiated by a third party (car owner, admitted driver). However, the update of biometric data in the reference database can be done automatically to avoid the periodical reenrollment. The third enrollment question will be answered by experiments comparing recognition performances after manual/semi-automatic enrollment/authentication modes are combined, as shown in Table 4.1.

Single driver identification vs. permanent driver verification

Referring back to driver authentication, two operational scenarios are proposed: single driver identification (SDI) and permanent driver verification (PDV). Single driver identification takes place directly after getting into a car and prior to engine ignition. Since it happens only once, there is no need for transparent automatic identification. The system can ask the driver to look at the camera and therefore can guarantee capturing frontal faces. In contrast, permanent driver verification implies the regular capturing and matching of a driver face either after a certain period of time or after each door closing. Permanently asking a driver to look at the camera can be very annoying, thus only automatic verification without notifying a driver is suitable. So in my consideration, the SDI is equal to the semi-automatic authentication (supports biometric memory function and mandatory driver assistance) and the PDV is equal to the automatic authentication (supports biometric immobilizer).

Table 4.1: Evaluation of operational scenarios for biometric driver recognition



Operational scenario			Enrollment from one frame	Expected recognition performance	Note
Name	Enrollment	Authentication			
SDI-sec.	Manual	Semi-auto.	Yes	High	Frontal faces in reference and test samples
SDI-conv.	Semi-auto.	Semi-auto.	No	High/Moderate	Frontal faces in test samples
PDV-sec.	Manual	Automatic	No	Moderate	Frontal faces in reference samples
PDV-conv.	Semi-auto.	Automatic	No	Low	No frontal faces are guaranteed

The essential difference between manual and semi-automatic enrollment is the human driven control of the image quality and the possibility of the prompt reenrollment in case the face model seems to be inappropriate. So, the manual enrollment suffers convenience in aid of security ('-sec.' addition to the name of the operational scenario). Semi-automatic enrollment, in contrast, invokes risks connected with the possibility of incorrectly generated face models. Here, security is suffered in aid of convenience ('-conv.' addition to the name of the operational scenario).

The semi-automatic authentication makes the system more reliable but less comfortable compared to an automatic one because of the need for driver interaction. Admittedly, the semi-automatic authentication can be equally well provided using fingerprints.

One technically interesting question is whether the enrollment can be carried out from only one accurately selected frame or whether it requires several frames with different face appearances. It is asserted that the single frame can be used only in case the reference frame is accurately selected during manual registration and the premise that a driver correctly poses his face with a neutral expression during the semi-automatic authentication.

4.3.4 Acquisition system

For the purpose of face monitoring and facial driver identification an experimental system has been implemented in a conventional vehicle Opel Vectra B. The low-cost color CCTV camera with 380 TV lines is mounted on the dashboard facing the driver's face through the steering wheel aperture. An analog-digital video converter shrinks the frame resolution to 480x360 pixels and delivers a digital video stream to a CarPC at the frame rate of 25 FPS. The camera has a CMOS imaging sensor and a standard 3.6 mm lens. Thanks to a CMOS sensor the imaging system does not suffer from the blooming effect. The ultra wide-angle lens enables acquisition of a wide area and guarantees that the driver's face is unlikely to get out of the view range. The camera includes no IR-filter and therefore enables image acquisition at night having an active IR illumination. Figure 4.27 shows randomly chosen frames. The resolution is intentionally selected so that a localized face rectangle usually does not exceed 100x100 pixels. Formally speaking, these images cannot be considered as biometric data and are not appropriate for large-scale applications. Hence, the data leakage from the reference database will not lead to identity theft.

In contrast to two formerly proposed applications (SOD, UDIS) where the camera position is undisputed, the position of a camera for facial driver recognition can vary. In fact, since the prospects for the camera integration go widely beyond the facial driver recognition the camera should simultaneously serve as many automotive applications as possible. However, the most versatile omni-directional camera, which is proposed for the SOD and enables simultaneous observation of all passengers, shows very low face detection rates and cannot be suggested for facial driver recognition. In fact, the capturing of frontal faces is of crucial importance for reliable face detection and recognition. Several authors install a camera on the side of the steering wheel resulting in sideway facial views [229, 253] or above the steering wheel [82, 116], which eventually distracts a driver. Locating the camera in the dashboard behind a steering wheel (or on the steering column) is probably the best solution. In this case the frontal view of a face can be observed most of the time through the steering wheel aperture [142]. The seldom occurring obstructions caused by the rotated steering wheel do not diminish recognition performance because these moments are irrelevant from safety and security points of view and can be ignored. The wide-angle lens is an indispensable component regarding the short distance between the camera and a face. Theoretically, this camera setup may support the eye closure estimation for drowsiness recognition. Practically, the much higher resolution and another lens are required to provide high-quality images of the eyes.

4.3.5 Illumination

Additional illumination modules are necessary for operating at night. Due to the fact that the artificial light in a visible spectrum can distract a driver, light sources must operate in an invisible spectrum. The only imperceptible harmless waves that can be captured by an imaging sensor are NIR waves within the span of 800 to 1000 nm. There are two ways of positioning illumination sources. Several diffusive lighting modules can be mounted on the ceiling of the passenger compartment, or the directional light source can be settled behind the camera facing the driver's face. The second choice is preferable because of the uniformly illuminated face and the effect of bright pupils, which is very useful for robust eye detection [122]. In particular, the camera body includes a ring of 30 NIR light-emitting diodes (LED) around the camera lens (see Figure 4.28). The wavelength emitted by the diodes is approximately 850 nm. The NIR lamp operates not only at night but also during daytime to flatten out strong lateral illumination.

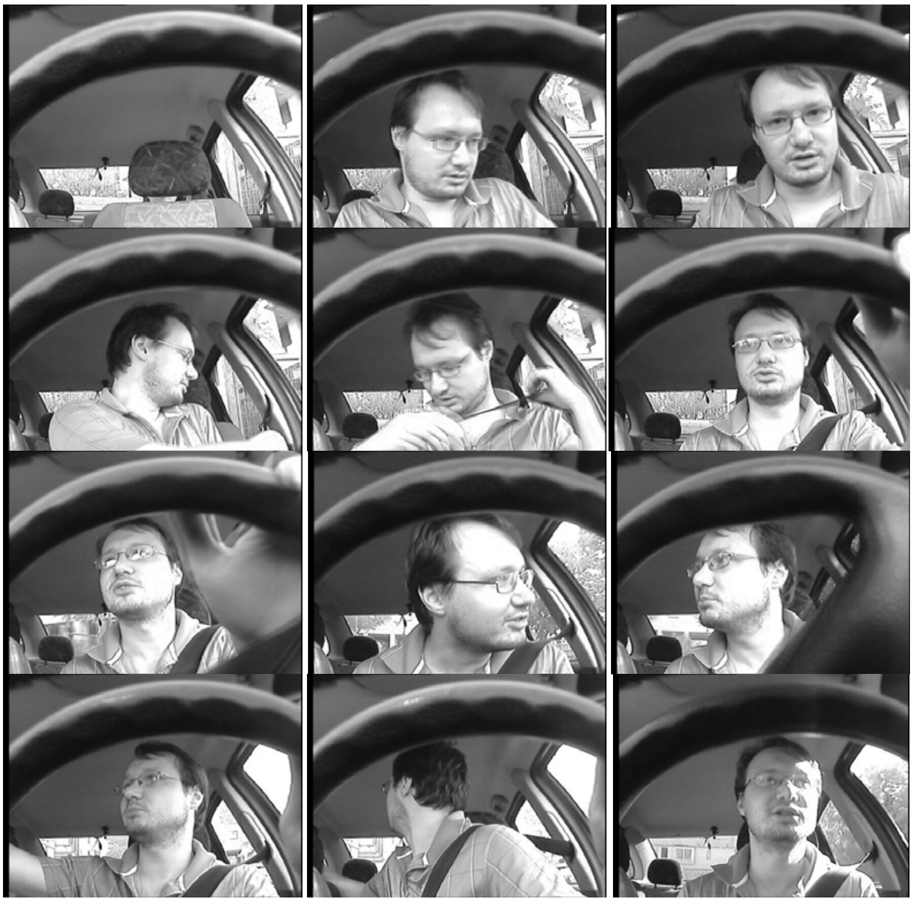


Figure 4.27: Randomly chosen frames from the experimental car Opel Vectra B

4.3.6 Computer vision approaches

Automatic face recognition (AFR) has a long research history of more than 40 years. The AFR approaches can be considered mature for practical usage however not for large-scale applications. There are commercial AFR systems integrated into login tools of operational systems [148], social networks [167], casino cheater recognition software [27] etc. It is very hard to compete with such systems developed and optimized by dozens or hundreds of engineers. Modern academic AFR systems have recently reached a very high level of recognition performance as well [236]. This can be seen in results of competitions between academic systems e.g. Face Verification Competition carried out on the XM2VTS database [165, 171, 169]. The competitions for commercial systems are organized by the National Institute of Standards and Technology (NIST) and called Face Vendor Recognition Test (FRVT) [81]. Tests are carried out on the FERET database [193]. The algorithms are usually tuned to perform well with standard databases leading to very impressive classification rates during the competition. Although the organizers of competitions endeavor to provide large datasets and plausible test protocols in order to reach statistically significant recognition performance, the real-life performance may drastically differ. The recent comprehensive overviews of standard face recognition technologies are provided in [118, 39]. Despite the broad variety of face recognition approaches grouped according to the input data (2D, 3D), feature extraction algorithm (holistic, analytic, hybrid), classification algorithm etc, the basic categorization of AFR methods transforms into the contradistinction of appearance-based (holistic) and feature-based (analytic)

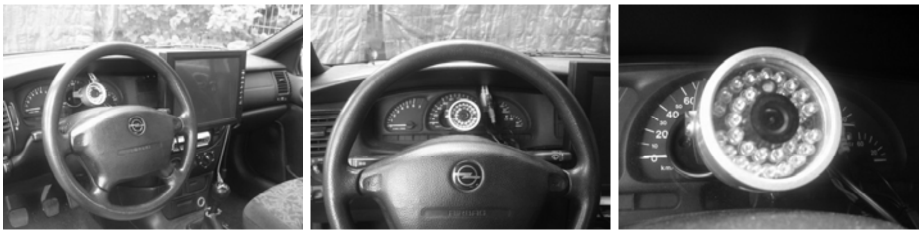


Figure 4.28: NIR camera mounted to the dashboard of the experimental car

approaches [153] compared in Table 4.2.

Table 4.2: Comparison of appearance-based and feature-based approaches for face recognition

Important criteria	Appearance-based approaches	Feature-based approaches
Demand for high-quality images	No	Yes
Demand for supervised enrollment	No	Yes
Learn from one image	No	Yes
Rotation invariant	No	Partially
Illumination invariant	No	Partially
Commonly used algorithms	Eigenfaces (PCA), Linear discriminant analysis (LDA)	Elastic bunch graph matching (EBGM)

Clearly, feature-based methods are more reliable in terms of recognition performance, but have serious demands on data quality and are computationally more expensive. Talking about the enrollment from a single frame, an employment of a feature-based method is the only feasible way to achieve the convincing recognition performance. Nonetheless, the demands on the acquisition process should be similar to taking biometric photos for travel passports. The demands are listed in the standards ANSI INCITS 385-2004 and ISO/IEC 19794-5-2005, and imply strongly supervised process (incl. a high-end camera, a high-resolution image, uniform illumination and an exactly frontal face). This is hardly implementable in a car. Without the fulfillment of the mentioned demands the facial landmarks cannot be reliably found, drastically reducing the recognition performance. Appearance-based approaches are more flexible regarding the quality of captured data. Aiming at the implementation of a low-cost system, we have at our disposal only low-price cameras and, therefore, theoretically appearance-based face recognition seems to be preferable.

I started research with the concept that the classification approach is the most important part of the face recognition system. If appropriate pre-processing is done, there is no need for feature extraction because a face is perfectly presented by pixel intensities. In fact, if the face size is not large (say 50x50 pixels) and a low-scale application is addressed (say less than 10 users) then indoor face recognition works well even using the nearest neighbor classifier and the more sophisticated classifiers excluding outliers provide comfortable recognition performance. If the face size becomes relatively large (say 200x200 pixels), then a dimension reduction is required e.g. by means of PCA. In case the number of users increases, a sophisticated classifier becomes indispensable. In a first concept I have proposed a Fuzzy ARTMAP neural network as an almighty tool for classification of facial patterns which has an adaptive structure permanently modifying the network weights to better classify the altering patterns. I assert that Fuzzy ARTMAP theoretically outperforms MLP and the nearest neighbor classifier. I have implemented all these concepts in FaceART face recognition software. However, tests on large-scale databases e.g. XM2VTS (see Section 5.3) have shown that FaceART cannot compete even with the modern academic face recognition systems [161]. This is why I was disappointed in the first concept, finished the development of FaceART, and switched attention to a second concept proclaiming features to be the most important.

Declaring features to be more important than a classifier has a rather theoretical basis. In fact, regarding Duda et al. [55], the nearest neighbor rule does not deliver the classification error more than twice that of the Bayesian classifier which is the optimal classifier from the probabilistic point of view. Moreover, the classification error of the k-NN classifier approaches those of the Bayesian classifier when the number of training samples is significantly higher than the number of classes. So, in case the expected classification error is for example under 1%, it does not matter which classifier is applied. The error variance of different classifiers will be marginal, which has also been proven in tests. So using e.g. linear SVM as a classification approach minimizing the structural risk should be sufficient to achieve high recognition performance assuming the features are well selected and the relatively high number of uniformly distributed training samples is provided.

Surely, feature selection can significantly improve appearance-based approaches selecting for instance the more important eigenfaces. Be advised that eigenfaces are sorted based on the information amount and not based on their discrimination power regarding the presented classes. However, the feature selection helps here only under the assumption that the pixel intensities represent faces well. Moreover, sophisticated classifiers include feature selection. They generate as many features as possible and consecutively sort them removing all insignificant ones. In fact, pixel intensities do not represent faces well, so that the expert knowledge is required to determine important face components manually, transmit this information to an algorithm, and provide the logic how face components are collocated. And exactly this task is done by feature-based approaches.

The implementation of a feature-based face recognition system is beyond the scope of this thesis. It was not necessary because of accessibility of well implemented and quite successful commercial systems. In fact, the development of a new face recognition system was never the objective of this thesis. The objective is to find out if automatic face recognition can be successfully integrated into a car and to determine the implementation constraints and general limitations. So, in order to test both appearance-based and feature-based face recognition concepts, one academic AFR system based on PCA and one commercial AFR system based on fiducial points are examined in the car passenger compartment and compare recognition performances.

As a proponent of appearance-based approaches, the FaceART face recognition system, which I developed in 2006-2007 at the University of Magdeburg [161], is addressed. In the most sophisticated operating mode, the localized and normalized facial images are represented by coefficients of PCA transform and classified by a fuzzy ARTMAP neural network. This system will be addressed later in detail. As a proponent of feature-based approaches, I have selected the Luxand FaceSDK [155]. This AFR system is available on the market at a reasonable cost and has a very respectable recognition performance with common face databases. The developers have tested it with the FERET database [193] and my personal tests have been carried out with the XM2VTS database [170] (see Section 5.3). The software contains face detection and face recognition functionality. In the examined version 2.0, 40 fiducal points (eyes, eyebrows, mouth, nose, face contour) are localized on a face and used for forming the biometric template (see Figure 4.29). Head rotations cause affine transformation of the geometric face model and therefore elastic modifications in the point structure rendered on the visible plane. The system is declared to be robust against ±30 degrees in-plane rotation and ±10 degrees out-of-plane rotation.

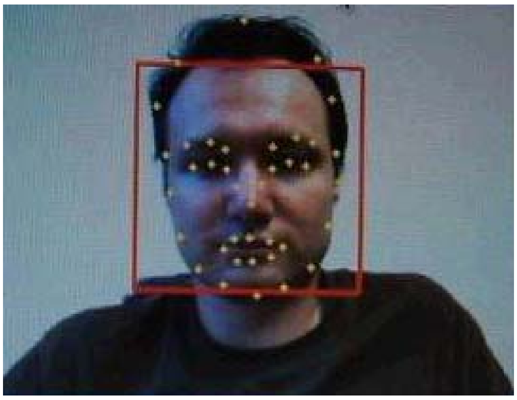


Figure 4.29: Fiducial points and face rectangle localized by Luxand FaceSDK

4.3.7 FaceART face recognition system

The FaceART is organized as a standard biometric system consisting of enrollment and authentication procedures. The first collects reference biometric data of users and the second compares query samples with user templates. The general workflow comprises four basic modules: acquisition, pre-processing, feature extraction and matching. The software is implemented in C/C++ under Linux making use of the OpenCV library [115].

Acquisition

The input data of the acquisition module can be provided in form of an image, a list of images, a video stream from a file or a camera video stream. The output data is a grayscale image of limited size. Large images are automatically reduced in size prior to pre-processing.

Pre-processing

Pre-processing includes three steps: face localization, light normalization and masking. Light normalization is performed by the best-fit plane (BFP) subtraction and subsequent histogram equalization. Masking hides background parts usually appearing in the corners of a face image. The processing chain is visualized in Figure 4.30. The output of the pre-processing module is a cropped face image of 50x50 pixels.

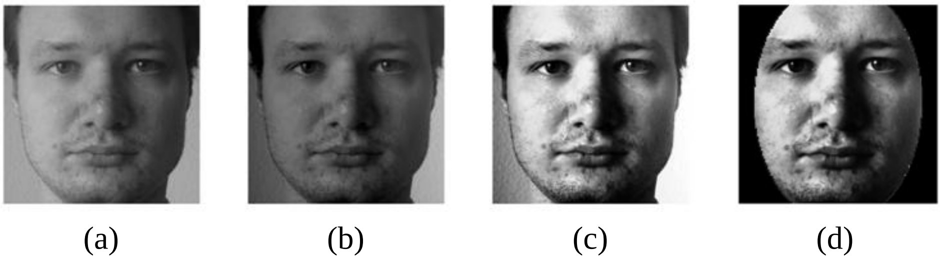


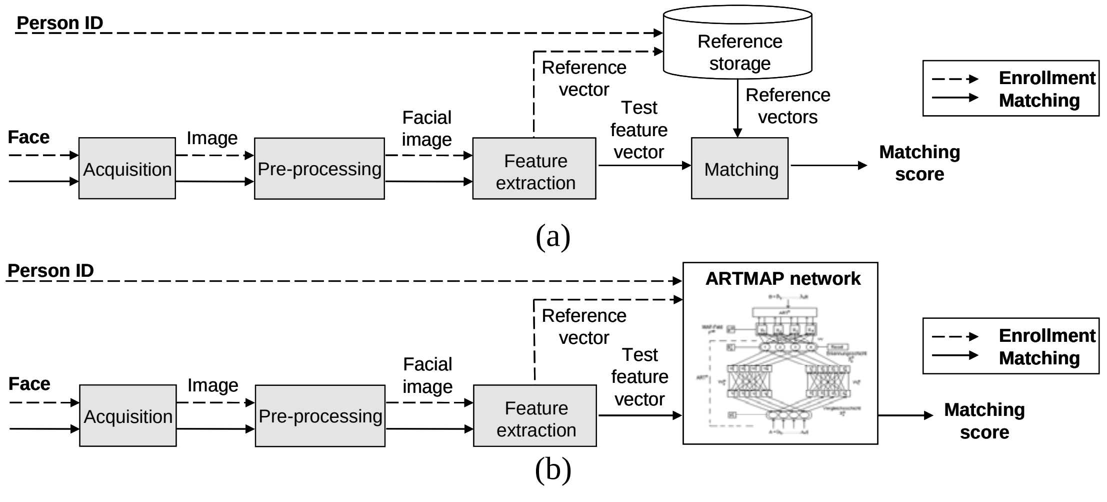
Figure 4.30: Face normalization: (a) original; (b) after BFP subtraction; (c) after histogram equalization; (d) after masking

Feature extraction

In the simplified version feature extraction is omitted and the normalized face image is used as a feature vector. The two-dimensional image is line-wise written to a one-dimensional array so that 50x50 pixel image becomes 2500-dimensional vector. For the reduction of dimensionality the PCA can be applied to transform face images to eigenfaces decomposition coefficients as described in Section 3.2.3. The insignificant dimensions can be reduced by determining the maximum dimensionality or by determining the amount of information that can be sacrificed. The output of the feature extraction module is a feature vector of predefined size.

Matching

The fourth and last module is matching. Here the test feature vector is classified using k-NN or Fuzzy ARTMAP. The matching can be provided either in verification or in identification mode. In case of verification the result is a matching score representing the dissimilarity between the test sample and the template. In case of identification the result is a rank list of users with corresponding matching scores. The k-NN applies majority voting rule as described in Subsection 3.3.2 so that the matching score is a percentage of neighbors in the particular class. This value rises in the interval [0,1] and can be considered as a membership likelihood. The ARTMAP network substitutes the reference storage and matching algorithm as it illustrated in Figure 4.31. The detailed description of the fuzzy ARTMAP is available in [33] and also briefly presented in Subsection 3.3.3. In FaceART the network is used as a black box with only one controlling parameter - vigilance $\rho \in [0, 1]$, which controls the degree of code compression. The lower value implies more compact class representation. In particular, ρ describes how well the model fits the training vectors. Low values lead to immoderate generalization and high values lead to overfitting. In the marginal case $\rho = 0$ each class is represented by the only one node and the network operates similarly to the Naive-Bayes classifier. In contrast, if $\rho = 1$ then each training vector obtains a node and the network degrades to the nearest neighbor search with a scalar product as the similarity measure. The network is trained using fast learning. The classification result of the ARTMAP is the set of likelihood values reflecting the membership in registered classes. A likelihood value is within the interval [0,1] and the sum of all likelihood values returned for a test sample is 1.



(a)

(b)

Figure 4.31: (a) Standard workflow of an automatic face recognition system; (b) The workflow modified by addition of ARTMAP network (reprinted from [161])

The FaceART can operate in one of four matching modes. In the first mode test face images are collected and directly matched with reference face images. The comparison is done by means of one of the three p-norms (p=1, 2 and ∞) or by cross-correlation between the clipped central part of the test face image and reference face image. The classification of test samples is done by means of the k-NN classifier. The second mode extends the first one by integrating the ARTMAP network for the classification of raw face images. In the third mode the raw face images are substituted by the coefficients of eigenfaces. The comparison of reference and test vectors is done in the same manner as in the first scenario except for the cross-correlation option. The comparison is extended by means of the Mahalanobis distance. The classification of test samples is performed with the help of the k-NN classifier as well. In the fourth mode, which is the most sophisticated one, the ARTMAP network is learned from the coefficients of eigenfaces. Figure 4.32 schematically illustrates all four matching modes. The dashed line in the diagram means that eigenfaces can be calculated based on the reference database (or provided by an external application that is not shown in the diagram). However, the eigenfaces used in this evaluation are derived from reference images.

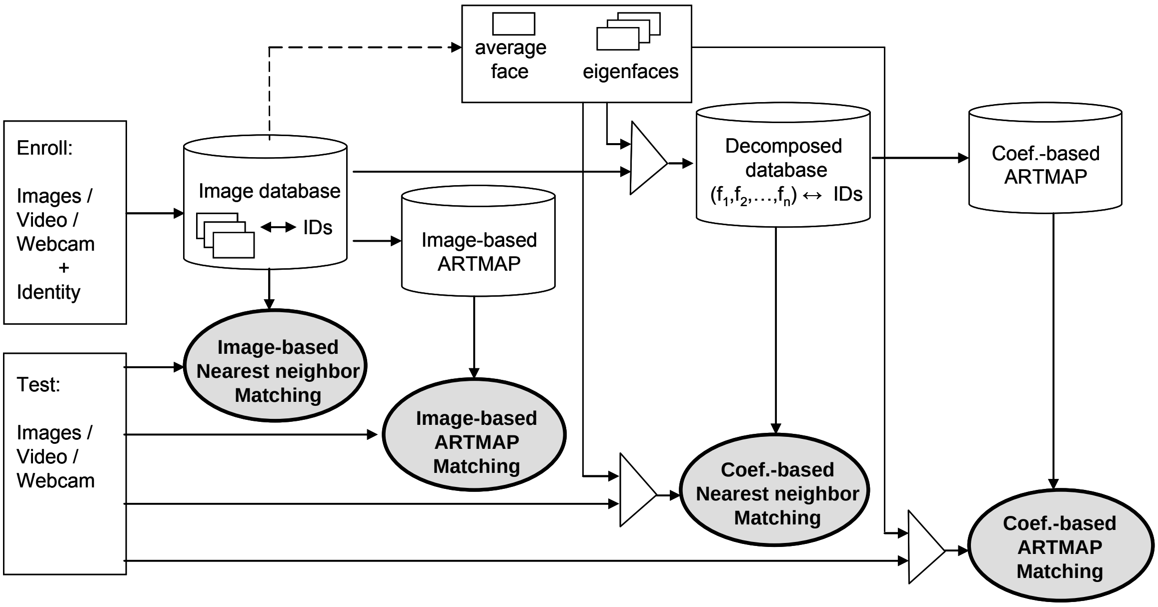

average face
eigenfaces
Enroll:
Images / Video / Webcam + Identity
Image database
IDs
Image-based ARTMAP
Decomposed database
(f1,f2,...,fn) ↔ IDs
Coef.-based ARTMAP
Test:
Images / Video / Webcam
Image-based Nearest neighbor Matching
Image-based ARTMAP Matching
Coef.-based Nearest neighbor Matching
Coef.-based ARTMAP Matching

Figure 4.32: General structure of the FaceART framework with four matching modes (reprinted from [161])

4.3.8 Evaluation objectives

The evaluation of the proposed face recognition algorithms (FaceART, Luxand FaceSDK) is performed in two stages. At first, the reference recognition performance is determined based on the XM2VTS database using the Lausanne protocol [154]. For the FaceART, the optimal matching mode along with the corresponding parameters is revealed. Secondly, the recognition performances of both systems are determined in the experimental vehicle with OpelDB database. Four operational scenarios are addressed: SDI-sec., SDI-conv., PDV-sec., PDV-conv. (see Section 4.3.3). The main evaluation objective is to reveal if high driver recognition accuracy can be achieved in any of the proposed scenarios. The secondary evaluation objective is to reveal which face recognition algorithm performs better in the simplified case (XM2VTS) with constant illumination and frontal faces, and in the close to real-life environment (OpelDB) with an uncontrolled illumination and non-guaranteed frontal faces.

An important issue of a realistic biometric application is the aging of the reference data. For this reason, the experimental data in OpelDB (reference and test samples) have been collected in five sessions with a time interval of at least one week between two subsequent sessions. Three test modes are addressed:

1. No-aging mode (non-realistic perfect situation): the reference and test data are taken from the same session.
2. Adaptive mode (simulation of the template update): the reference and test data are taken from two subsequent sessions, whereby the session with the reference data always precedes the session with the test data.
3. Aging mode (standard situation): the reference data from the first session is matched against the test data from the four remaining sessions.

During the evaluation the test modes are combined with the operational scenarios to discover how successful the considered algorithms are in perfect and real situations and how applicable the proposed operational scenarios are with regard to the current state of technology.

Chapter 5

5 Experiments

The experimental chapter is comprised of four sections. The first section addresses the experiments with the simplified version of seat occupancy detection system identifying whether the seat is occupied or not. The second section describes the experiments provided with the user discrimination system to determine who is currently interacting with the integrated center consol. The third and fourth sections are devoted to the facial driver identification. In the third section the introduced face recognition systems are tested with one common face database (XM2VTS) using the standard test protocol (Lausanne protocol) to obtain the general recognition performance and compare it against that of well-known academic face recognition systems. In the fourth section the experiments are carried out on the personally collected data in real-life environment (OpelDB) to explore the applicability and practicability of the proposed concepts.

5.1 Vision-based seat occupancy detection system

This section is comprised of the following subsections:

- Data collection
- Performance measures
- Evaluation of template matching applied to seat occupancy detection
- Evaluation of face detection applied to seat occupancy detection
- Feasibility of face detection for seat occupancy detection
- Conclusion

5.1.1 Data collection

As mentioned in section 4.1, the seat occupancy detection system has been developed within the framework of the VisMes project using the real car for experiments. The proper evaluation of an imaging system in a real-life environment is a very challenging task because of the variety of lighting conditions as well as the diversity of persons and objects in a passenger compartment. In fact, it is practically impossible to collect data representing all real-life situations. Nonetheless, by organizing the experiment, we seek to cover as many situations as possible. In doing so we try to:

- engage different persons and objects
- engage as many test persons as possible
- capture images under uniform and non-uniform illumination conditions
- strive for an equal number of men and women

- strive for an almost uniform age distribution

Test persons are asked to get into the car in an arbitrary manner, take a seat, fasten the seat belt, remain seating for at least three seconds and then change the seat. Every test person is expected to consecutively occupy all five seats. In a similar manner, objects and infant seats are placed to different seats for a time period of at least three seconds. The in-car camera is activated just before the first passenger has entered the car so that the first frame in each video can be used to cut templates of empty seats.

All collected video frames are manually annotated. The meta-information about the environment such as weather conditions, time of day, position of the sun regarding the car, car location etc. as well as the seat occupancy status for each seat is embedded directly into camera frames. The actual evaluation of the recognition performance is a frame-by-frame comparison of the algorithm's decision with meta-information and counting discrepancies.

Figure 5.1 shows several random cuts with different occupants under different illumination conditions at different locations.

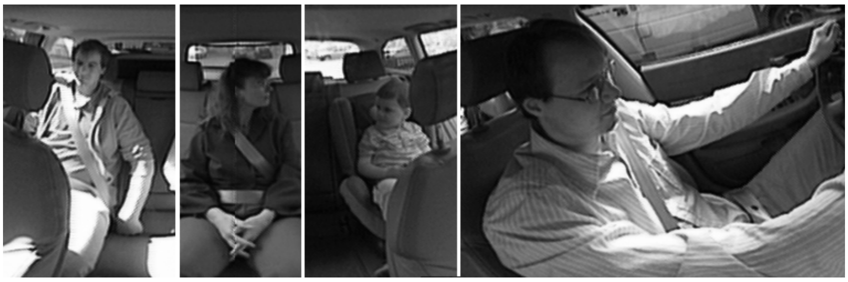


Figure 5.1: SOD, examples of occupied seats (from left to right): tall man, small woman, child and average-sized man (reprinted from [160])

The resulting experimental database includes 39 persons. Among them are males and females of different ages and body heights as well as children in child restraint seats. In addition, the database contains several objects such as bags, empty rear/forward-facing child restraint seats, and a beer box. The total number of video frames showing the complete car passenger compartment is 53928. The distribution of these frames regarding the type of occupancy is presented in Table 5.1 separately for uniform and non-uniform illumination conditions. Splitting into uniformly and non-uniformly illuminated scenes serves to study how robust the algorithms are when uncontrolled drastic changes in illumination arise. Due to the limited access to the experimental car, the database does not include night samples.

Table 5.1: SOD, distribution of frames in the experimental database (reprinted from [160])

Seat	Front left	Front right	Rear left	Rear center	Rear right
Uniform illumination conditions					
Children	0	2268	0	0	1648
Adults	12656	8822	9511	7092	8884
Objects	70	2974	2070	1246	3030
Empty seats	28195	26857	29340	32583	27359
Non-uniform illumination conditions					
Children	0	0	776	0	757
Adults	3061	4112	3020	1791	3933
Objects	0	0	77	0	331
Empty seats	9946	8895	9134	11216	7986

5.1.2 Performance measures

The evaluation of the seat occupancy detection system is provided in form of hypothesis testing as introduced in Subsection 2.3.2. The proposed simplified system distinguishes between two seat occupancy states: "the seat is occupied" and "the seat is empty". Hence, the null hypothesis for each frame states that the seat is occupied and the alternative hypothesis states that the seat is empty. The choice of the null hypothesis seems to be obvious here. The null hypothesis can be rejected only in case of very strong evidence against it and the false rejection of null hypothesis is reckoned as the critical failure. This so-called type I error can be interpreted as that the occupied seat has been recognized to be empty. As a consequence the airbag is suppressed and the occupant confronts the risk of serious injuries or even fatality. The opposite false decision, namely accepting the null hypothesis in case it is actually wrong, is considered here as non-critical. This so-called type II error can be interpreted as that the empty seat is recognized to be occupied. As a consequence the airbag is deployed on an empty seat leading only to unnecessary costs of replacing the airbag. However, for a complete occupant classification system this choice of the null hypothesis can be inappropriate, for details see Subsection 2.3.2.

In order to obtain intuitive performance indicators and to provide a direct link to the application, two errors are introduced: the false non-occupancy rate ($FNOR$) which is the relative frequency of frames where an occupied seat has been recognized as an empty seat, and the false occupancy rate (FOR) which is the relative frequency of frames there an empty seat has been recognized as an occupied seat. Both error rates are considered equally important here.

By applying template matching to seat occupancy detection, the cut of an empty seat (template) is searched for in the corresponding seat region. The decision to consider a seat to be empty or not is made based on the correlation coefficient (r) resulting from the matching, as defined in Subsection 4.1.7. A high value of the correlation coefficient ($|r| \rightarrow 1$) indicates an empty seat. By determining the threshold (τ) for the correlation coefficient, the decision boundary between classes of empty and occupied seats is specified. The numbers of false positive and negative decisions depend on the selection of the threshold. Hence $FNOR$ and FOR are reciprocal functions of the threshold. The formal notations of both error rates are given by equations 5.1 and 5.2.

$$FNOR(\tau) = Pr\left(r \geq \tau | occupied\right) \approx \frac{N_{oc}^{-}}{N_{oc}} \tag{5.1}$$

$$FOR(\tau) = Pr\left(r < \tau | non - occupied\right) \approx \frac{N_{non-oc}^{+}}{N_{non-oc}} \tag{5.2}$$

Values N_{oc} and N_{non-oc} denote numbers of frames with occupied and empty seats respectively just as values N_{oc}^{-} and N_{non-oc}^{+} denote numbers of falsely missed occupied seats and falsely detected empty seats.

In the experiments, the threshold is selected in a way that $FNOR$ and FOR become equal. The error rate at this point is referred to as equal error rate (EER). EER is used here as the performance indicator of the template matching approach. A lower EER implies better recognition performance.

By applying face detection to recognize whether a seat is occupied or not, it is more plausible to consider the opposite value of $FNOR$, namely $1 - FNOR$. In reference to face detection this value can be called detection rate (DR) and represents the relative frequency that a face has been found in the frame with an occupied seat. The second rate (FOR) remains unchanged but can be renamed to false detection rate (FDR) for better plausibility. This one represents the relative frequency that a face has been found in the frame with an empty seat. The formal notations of DR and FDR are given by equations 5.3.

$$DR \approx \frac{N_{oc}^{face}}{N_{oc}}, \quad FDR \approx \frac{N_{non-oc}^{face}}{N_{non-oc}} \tag{5.3}$$

The face detection algorithms utilized in the experiment have a fixed setup which means that the tolerance parameters for face sensitivity cannot be adjusted. Hence, the decision boundary between face and non-face classes is predefined and the values N_{oc}^{face} and N_{non-oc}^{face} are calculated only at one decision threshold. For a good detection algorithm, the DR should be as high as possible and the FDR as low as possible.

5.1.3 Evaluation of template matching applied to seat occupancy detection

The first technique introduced for the seat occupancy detection is template matching. Here the task is solved in an inverted way, namely the algorithm looks for an empty seat and not for an occupied seat. Since an empty seat is indicated by its part (template), the questions arising during the evaluation are:

- What is the best template selection strategy,
- How different pre-processing techniques and fusion approaches influence the performance of template matching.

There are three template selection strategies introduced in Subsection 4.1.7:

- A_1: Select a template almost as large as a seat (large template),
- A_2: Select a template significantly smaller than a seat (small template),
- A_3: Select several small templates to determine whether numerous small templates perform better than just one (two small templates are taken as an example).

The template selection strategies are combined with two pre-processing techniques: local normalization (B_1) and edge detection (B_2) to counter non-uniform illumination conditions and complemented by the case of no image enhancement (B_0).

In order to study whether pre-processing techniques can be reasonably combined and whether considering a sequence of frames instead of a single frame improves the recognition performance of template matching, three matching score fusion strategies are proposed:

- Multi-algorithmic matching-score fusion (B_{1+2}),
- Temporal matching-score fusion (B_1^T, B_2^T),
- Multi-algorithm temporal matching-score fusion (B_{1+2}^T).

Note that each seat is treated individually so that the template selection strategy, image enhancement technique and fusion method can vary from seat to seat.

The results of the experiments on template matching are assembled in tables 5.2 and 5.3 showing the EER values resulting from videos with uniform illumination and non-uniform illumination conditions respectively. The best EER values are highlighted for each of the five seats.

For uniformly illuminated frames (see Table 5.2), considering front seat and two sided rear seats the small template (A_2) significantly outperforms the large template (A_1). However, increasing the number of templates from one to two (A_3) only occasionally improves the results and the improvement is even insignificant. In contrast, the two small templates strategy applied to the rear center seat notably improves the recognition performance of the algorithm. Applying the large template, however, is reasonable only in case edge detection is performed.

Unfortunately, the experiment does not allow identifying any tendency regarding the utilization of image enhancement techniques. The number of cases when edge detection outperforms local

Table 5.2: SOD, template matching, EER values for uniformly illuminated frames (modified from [160])

Seat	Front left	Front right	Rear left	Rear center	Rear right
A_1: large template					
B_0: raw	13.33%	22.27%	17.78%	25.63%	25.89%
B_1: local normalization	5.33%	9.31%	14.65%	17.68%	14.80%
$B^T{}_1$	5.29%	9.41%	14.34%	17.56%	14.47%
B_2: edge detection	12.63%	5.02%	7.75%	8.56%	14.32%
$B^T{}_2$	11.97%	5.04%	7.82%	7.70%	13.37%
B_{1+2}	5.14%	8.90%	8.53%	13.53%	14.37%
$B^T{}_{1+2}$	5.19%	9.23%	8.52%	13.19%	14.06%
A_2: small template					
B_0: raw	5.59%	11.58%	9.12%	21.38%	13.70%
B_1: local normalization	4.08%	2.82%	7.32%	17.20%	12.44%
$B^T{}_1$	2.81%	2.95%	6.71%	16.95%	11.35%
B_2: edge detection	7.93%	2.81%	9.10%	18.86%	6.33%
$B^T{}_2$	7.42%	2.55%	8.11%	16.17%	5.43%
B_{1+2}	4.66%	2.60%	6.94%	20.16%	6.43%
$B^T{}_{1+2}$	3.45%	2.53%	6.78%	16.99%	5.60%
A_3: two small templates					
B_0: raw	2.60%	15.61%	8.70%	12.93%	10.56%
B_1: local normalization	4.86%	12.67%	6.23%	13.74%	7.88%
$B^T{}_1$	4.82%	12.71%	6.53%	13.15%	8.09%
B_2: edge detection	3.30%	17.38%	7.33%	13.17%	7.83%
$B^T{}_2$	3.68%	17.22%	7.69%	10.25%	7.67%
B_{1+2}	3.89%	12.06%	6.30%	11.30%	6.94%
$B^T{}_{1+2}$	3.63%	11.88%	6.60%	10.56%	6.70%

normalization is almost the same as the number of cases when the opposite is true. Even the matching-score fusion of both does not always lead to the improvement of template matching results.

Temporal fusion of matching scores leads to slight improvement of the results regardless of the image enhancement technique and the template selection strategy in most cases.

For the front left seat, the best result (EER = 2.60%) is achieved with two small templates and without any pre-processing. For the front right seat, the best performance (EER = 2.53%) is obtained with one small template and multi-algorithm temporal matching-score fusion. For the rear left seat, the two small templates strategy with local normalization leads to the best result (EER = 6.23%). For the rear right central seat, the large template strategy with edge detection and temporal fusion is the best one (EER = 7.70%). For the rear right seat, the best result (EER = 5.43%) is gained with one small template, edge detection and temporal fusion.

The asymmetry in results between the front left and the front right seats is caused by the presence of children in a child restraint system on the front right seat. The 2268 children frames constitute approximately 5.5% of the whole number of examined frames. A child restraint system always covers the part of the seat containing the template selected in the A_2 strategy. Since the child restraint system appears clearly different from the template, the recognition results for the front right seat are better. In contrast, when the large template or two small templates are searched for, the recognition results for the front right seat are worse compared to that for the front left seat. This happens because the child restraint system usually does not fully cover the seat back, but the seat back is the significant part of the large template, and one of two small templates, which has been cut from the seat back, remains uncovered.

There is the same reason for asymmetry when the rear right and rear left seats are compared. There are 1648 frames with a child restraint system on the rear right seat (approximately 4% of the total amount of examined frames) and no single frame with a child restraint system on the rear left seat. However, the results are more or less symmetrical because regardless of the template selection strategy the child restraint system almost fully covers the seat area.

The highest error rates are generated when template matching is applied to the rear center seat, which is especially apparent for the small template strategy (A_2). The rear center seat covers the smallest fraction of the camera frame of all seats, so that the templates selected for this seat are significantly smaller than those selected for the rear side seats. Moreover, the pattern of the templates is almost uniform exhibiting a low number of distinctive characteristics. In contrast, by selecting two small templates (A_3) and even the large template (A_1), the problem of template uniformity is overcome.

The situation for non-uniformly illuminated frames (see Table 5.3) is different. There is no reason to expect symmetric results for the front seats and rear side seats because the illumination is not controlled. Some frames include cast shadows, glares, and other detrimental illumination conditions which can significantly harm template matching results. These effects are mostly reflected in template matching results for the rear center seat with A_1 and A_2 template selection strategies. Here, the permanently overexposed head restraint is a part of the template. When no edge detection is applied, the EERs exceed 50% making the system worse than random guessing.

The most notable fact is that in all cases template matching greatly benefits from pre-processing. Considering the rear seats, edge detection performs significantly better than local normalization regardless of the template selection strategy. A combination of edge detection and local normalization (B_{1+2}) often cannot top the performance of single edge detection. For the front seats the superiority of one or another pre-processing approach cannot be derived from the results of the experiment. However, combining both pre-processing approaches (B_{1+2}) often leads to better EERs than when only one pre-processing approach is used.

Table 5.3: SOD, template matching, EER values for non-uniformly illuminated frames (modified from [160])

Seat	Front left	Front right	Rear left	Rear center	Rear right
A_1: large template					
B_0: raw	26.59%	13.42%	45.55%	67.11%	29.37%
B_1: local normalization	13.40%	0.91%	25.04%	54.19%	22.15%
B^T_1	13.53%	1.09%	24.31%	54.29%	22.60%
B_2: edge detection	10.91%	7.29%	9.41%	27.57%	18.15%
B^T_2	8.94%	6.30%	9.22%	27.09%	16.31%
B_{1+2}	12.80%	0.82%	9.20%	38.73%	20.22%
B^T_{1+2}	12.96%	1.14%	8.53%	38.41%	20.69%
A_2: small template					
B_0: raw	12.63%	27.36%	31.79%	51.52%	21.52%
B_1: local normalization	3.47%	2.73%	29.62%	50.90%	19.84%
B^T_1	2.01%	2.88%	30.94%	51.07%	17.56%
B_2: edge detection	5.34%	1.43%	14.06%	43.36%	13.86%
B^T_2	3.03%	1.13%	12.09%	39.16%	8.49%
B_{1+2}	2.83%	0.84%	18.93%	42.26%	11.02%
B^T_{1+2}	1.68%	0.97%	17.56%	42.62%	7.52%
A_3: two small templates					
B_0: raw	15.54%	1.68%	30.28%	33.14%	5.04%
B_1: local normalization	7.70%	0.69%	8.55%	22.37%	2.61%
B^T_1	5.49%	0.86%	9.26%	18.43%	3.84%
B_2: edge detection	7.54%	3.39%	4.45%	15.95%	3.20%
B^T_2	4.53%	2.71%	4.01%	11.07%	2.54%
B_{1+2}	4.54%	1.59%	6.00%	15.17%	2.38%
B^T_{1+2}	4.27%	1.55%	5.56%	11.48%	1.84%

For the front left seat, the best result ($EER = 1.68\%$) is gained with one small template and both pre-processing with temporal matching-score fusion. For the front right seat, the two small templates strategy with local normalization leads to the best result ($EER = 0.69\%$). For the rear left seat as well as for the rear center seat, the best results ($EER = 4.01\%$ and $EER = 11.07\%$ correspondingly) are achieved with two small templates and edge detection with temporal matching-score fusion. For the rear right seat, the best result ($EER = 1.84\%$) is gained with two small templates, multi-algorithmic fusion of pre-processing approaches and temporal fusion. So, except for the front left seat, the best results are achieved with two small templates.

It is important to note that template matching on the front seats and rear side seats under non-uniform illumination conditions performs better than on the same seats under uniform illumination conditions. These outcomes of the experiments can be explained by advantageous coincidences where the sided or frontal illumination through bright sunlight may actually increase the contrast of seat patterns and, therefore, simplify the search for the template.

For non-uniform as well as for uniform illumination conditions, temporal fusion of matching scores leads in most cases to a slight improvement of the results regardless of image enhancement and template selection approaches. The multi-algorithmic matching score fusion improves the recognition performance only in case both local normalization and edge detection perform equally well and the corresponding EERs are fairly low.

When small templates (A_2/A_3) are utilized, template matching seems to be more reliable in comparison to the case when the large template (A_1) is utilized. However, the general suggestion about the superiority of one or another template selection strategy cannot be made. Similarly, one pre-processing approach cannot be favored over another.

5.1.4 Evaluation of face detection applied to seat occupancy detection

The detection and localization of occupants' faces is probably the most plausible way to determine whether a seat is occupied by a person. However, the research question examined here is whether the face detection can be exclusively used for seat occupancy detection. In other words, a detected face reveals an occupied seat and the absence of a face reveals an empty seat.

Three exemplary face detection algorithms are taken into consideration. These are face detectors of Viola-Jones [242], Kienzle et al. [127] and Nilsson et al. [180]. The algorithms are compared based on the database described in Subsection 5.1.1. In order to avoid ambiguity in the test outcomes, the video frames with objects are considered to be non-occupied. The performance measures are described in Subsection 5.1.2.

The results of the experiments are illustrated in tables 5.4, 5.5 and 5.6. Table 5.4 demonstrates the results of face detection applied to the complete database. Tables 5.5 and 5.6 address experiments with uniform and non-uniform illumination conditions separately. Since DR and FDR are considered equally important the half-total error rate $HTER = 0.5((1 - DR) + FDR)$ can be taken as a scalar indicator of the recognition performance as described in Subsection 2.3.2. The best pairs of DR and FDR are highlighted for each of the five seats.

For the front seats, the complete breakdown of all three face detectors can be stated. The best detection rate (13.56%) is achieved with Nilsson's algorithm, but the corresponding false detection rate (4.58%) is unacceptably high. The algorithms of Viola-Jones and Kienzle yield far better false detection rates not exceeding 0.5%, but the corresponding detection rates are less than 5%, which is too low to consider the algorithms useful. The lowest detection rates by far are gained with the Viola-Jones algorithm. However, the algorithm yields near-perfect false detection rates.

For the rear side seats, the utilization of face detection can be considered reasonable. Here, the algorithms of Kienzle and Nilsson yield comparable detection rates from 9.66% to 14.43%

Table 5.4: SOD, face detection results for the complete test including uniformly and non-uniformly illuminated frames (modified from [158])

Seat	Front left	Front right	Rear left	Rear center	Rear right
Viola-Jones					
DR	0.69%	0.80%	15.79%	5.85%	20.13%
FDR	0.06%	0.07%	0.15%	0.04%	1.58%
Kinzle et al.					
DR	1.93%	4.20%	9.66%	4.07%	14.01%
FDR	0.36%	0.50%	0.40%	1.01%	1.07%
Nilsson et al.					
DR	12.49%	13.56%	11.62%	12.52%	14.43%
FDR	18.78%	4.58%	1.48%	1.35%	2.33%

at sufficiently low false detection rates which do not exceed 2.33%. The Viola-Jones algorithm clearly outperforms both other algorithms with detection rates of 15.79% and 20.13% and the corresponding false detection rates of 0.15% and 1.58% for the left and right seats respectively.

For the rear center seat, Nilsson's algorithm yields the best results with the detection rate of 12.52% and the corresponding false detection rate of 1.35%. Although the false detection rate is slightly higher than that of both competitors, the detection rate is by far the best of all three face detectors.

The separate evaluation of uniformly and non-uniformly illuminated frames has shown the same trend as observed in the evaluation of the complete database. The Viola-Jones face detector proves to be the most suitable one for the rear side seats. Nilsson's algorithm performs best at the rear center seat. All three algorithms perform very poorly at the front seats. Here, Nilsson's algorithm shows the best detection rates, but yields unacceptably high false detection rates. The false detection rates of the Viola-Jones and Kienzle algorithms are low but the corresponding detection rates are unacceptable.

Table 5.5: SOD, face detection results for uniformly illuminated frames (modified from [158])

Seat	Front left	Front right	Rear left	Rear center	Rear right
Viola-Jones					
DR	0.57%	0.82%	18.38%	4.29%	18.84%
FDR	0.04%	0.06%	0.18%	0.03%	1.99%
Kinzle et al.					
DR	2.01%	4.82%	11.89%	4.16%	16.33%
FDR	0.38%	0.36%	0.27%	0.68%	0.78%
Nilsson et al.					
DR	12.14%	13.73%	12.57%	11.79%	14.41%
FDR	21.55%	5.06%	1.31%	1.14%	2.30%

By comparing the results of the experiments with uniformly and non-uniformly illuminated frames, one notable observation has been made. Kienzle's algorithm performs much better with uniformly illuminated frames, which means that the algorithm is highly sensitive to illumination variations. Since the plain gray-scale values of images are utilized as features for classification, the algorithm obviously suffers of a lack of pre-processing. Two other algorithms do not show any significant differences in recognition rates between uniformly and non-uniformly illuminated frames and, therefore, seem to be quite insensitive to varying illumination conditions.

Table 5.6: SOD, face detection results for non-uniformly illuminated frames (modified from [158])

Seat	Front left	Front right	Rear left	Rear center	Rear right
Viola-Jones					
DR	1.21%	0.73%	9.30%	12.05%	23.01%
FDR	0.14%	0.13%	0.05%	0.06%	0.08%
Kinzle et al.					
DR	1.63%	2.50%	4.08%	3.74%	8.81%
FDR	0.28%	0.97%	0.85%	1.99%	2.13%
Nilsson et al.					
DR	13.96%	13.11%	9.24%	15.40%	14.46%
FDR	10.91%	2.97%	2.08%	2.00%	2.43%

5.1.5 Feasibility of face detection for seat occupancy detection

The current safety standard regulating the deployment/suppression of smart airbags is solely based on weight measuring. In fact, it is unclear whether an airbag should deploy when the seat is occupied by a heavy object. The standard weight measuring system forces an airbag to deploy on seats occupied by heavy objects or by empty child restraint systems. In contrast, a face detection system will find no faces at these seats and, therefore, can prevent airbag deployment. Hence, face detection can be considered a reasonable supplement to standard weight measuring devices.

However, the complete substitution of commonly used seat occupancy detection systems by face detection is not feasible. In fact, the practical implementation of face detection in a car is impeded by several inherent limitations. The most notable limitations are:

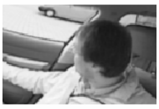
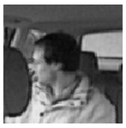
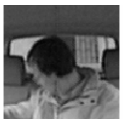
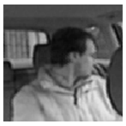
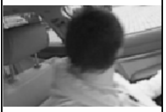
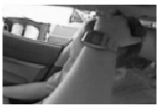
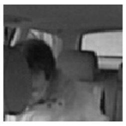
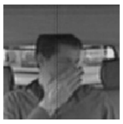
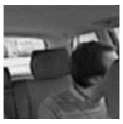
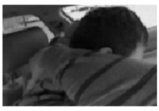
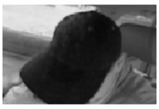
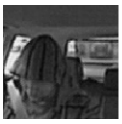
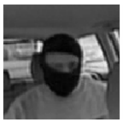
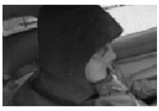
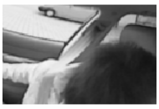
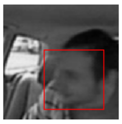
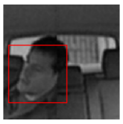
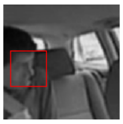
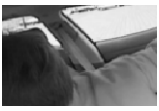
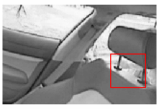
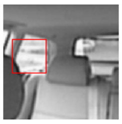
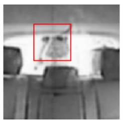
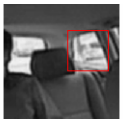
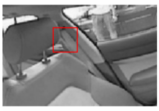
- Incapability of optical sensing systems to observe all regions of a passenger compartment in the same manner. Despite the 360-degree camera applied in the experimental car, there are many situations in which occupants' faces are not visible to the camera,
- Very high variability of face appearance caused by rotations and permanently changing illumination conditions.

Another reason for the poor face detection performance is the fact that the tested face detectors are trained based on images of frontal or almost frontal faces. Therefore, they perform well only for near-frontal faces which can be observed by comparing the results for front and rear seats. While at rear seats, the camera usually captures frontal faces, faces of driver and front-seat passengers are viewed by the camera as profiles. Table 5.7 demonstrates five classes of situations which can be considered major reasons for missing faces and false detections. These situations are arranged regarding their probability of occurrence in descending order and summarized as follows:

1. An extremely turned occupant's head. Considering the vehicle ingress and egress scenario, it is very likely to observe turned heads where the occupants' faces are invisible to the camera. Such frames make up the majority of test frames. Bearing in mind, face detection is performed for each video frame, the low face detection rates become obvious.
2. Occupants' faces are covered by parts of interior or by occupants' gestures. Even though this situation is not as common as the former one, coverings occur periodically.
3. An occupant wears excessive or obsolete clothing. A face that is partially hidden by a cap, hat or hood causes serious problems for face detectors. The frames presented in Table 5.7 are rather artificial, but their quantity shows that such cases must be considered.
4. An extremely tilted occupant's body causing two errors simultaneously: face detection at an adjacent empty seat and missing a face at an addressed occupied seat. Due to an extremely tilted body, a head can be located outside of the camera's view. These situations are quite seldom and occur in case an occupant changes a seat or looks back from a front seat.

5. Occurrence of face-like objects in windows or in the interior. The likelihood of this situation depends on a face detector. Nilsson's algorithm, for instance, provides unacceptably many false detections especially for the front seats. In contrast, the Viola-Jones and Kienzle's algorithms demonstrate relatively low false detection rates.

Table 5.7: SOD, major reasons for missing faces and false detection, situations 1 to 4 lead to missing faces at occupied seats; situations 4 and 5 lead to false detections at empty seats (modified from [158])

	Situation	Problem	Examples				
			Front right	Rear right	Rear center	Rear left	Front left
1.	Extremely turned head	Occupant's face is not visible to the camera					
2.	Interior details or gestures hide an occupant's face	Occupant's face is only partially visible to the camera					
3.	Excessive / obsolete clothing	Occupant's face is only partially visible to the camera			No sample in the database		
4.	Tilted body	Occupant's face is in the region of a wrong seat or out-of-region					
5.	Face-like objects in windows or in the interior	Face-like images lead to detection of absent faces					

Situations 1 to 4 constitute the main reasons for very low detection rates. Situations 4 and 5 are the main reasons for false detections. Regardless of the quality of the face detection algorithm, the majority of frames presented in Table 5.7 cannot be correctly classified.

5.1.6 Conclusion

The experiments with template matching have demonstrated that empty front seats can be reliably detected with the EERs ranging from 0.69% to 11.07% when the optimal pre-processing and fusion techniques are utilized. Unfortunately, the experiments have not revealed the superiority of one of the addressed pre-processing techniques. Similarly, a clear conclusion about the superiority of any template selection strategy cannot be made. However, it can be definitely stated that in case of non-uniform illumination conditions, the application of pre-processing is essential and that the temporal matching-score fusion almost always slightly improves the recognition performance. The recognition rates of the proposed optical approach to seat occupancy detection cannot compete with that of the established non-optical approaches (e.g. weight sensing) where missing of persons never occurs.

The experiments with face detectors have shown that face detection cannot replace conventional seat occupancy detection approaches because of its inherent limitations, but can support them by distinguishing between persons and objects. Moreover, face detection may form a basis for the subsequent occupant classification e.g. face-based gender and age recognition. The low detection rates resulting from the experiments are caused by factors which are independent from a face detection algorithm because faces on most of frames are invisible to the camera.

All three face detectors addressed in the experiments are incapable of reliable detection of face profiles at front seats because they are primarily tailored to detect nearly frontal faces. For the rear seats, the detection rates are significantly better, but seldom higher than 20%. The Viola-Jones algorithm performs best at the rear side seats and Nilsson's algorithm at the rear center seat.

The technical specifications of the camera, namely low resolution (720x576 pixels) and low frame rate (7.5 FPS), also pose some limitations to image processing algorithms. Faces of occupants at rear seats are seldom larger than 30x30 pixels. Rapidly moving faces cannot be tracked. So, several high-speed high-resolution cameras individually used for different seats could significantly improve face detection performance.

5.2 Visual distinguishing between driver and front-seat passenger hands

This section is comprised of the following subsections:

1. User discrimination in a real car (UDiS1)
2. User discrimination in the AMSL car simulator: first experiment (UDiS2)
3. User discrimination in the AMSL car simulator: second experiment (UDiS3)
4. Conclusion

5.2.1 User discrimination in a real car (UDiS1)

Visual user discrimination when driver and front-seat passenger interact with center console components is the next application developed within the framework of the VisMes project as a meaningful addition to seat occupancy detection. The same omni-directional camera is utilized for observing driver and front-seat passenger arms (see Section 4.2.4) in the same experimental car. However, the images produced by the camera are different: the bottom part of the image is a zoomed out version of that used for occupancy detection showing a 360-degree view of the complete passenger compartment just as the top part is an enlarged section of the center console region (see Figure 4.18). The bottom part is irrelevant for user discrimination.

In order to properly evaluate an imaging system in a real-life environment, a high amount of experimental data is required due to the diversity of illumination conditions and permanently changing appearances of hands. Moreover, the way passengers interact with center console components depends on how experienced they are. Younger people are usually more familiar with new technologies and, therefore, more confident when providing actions. In brief, by collecting the experimental data we try to cover as many situations as possible. In particular we try to:

- engage as many test persons as possible,
- capture images under uniform and non-uniform illumination conditions,
- capture images where persons wear different clothes and gloves,
- strive for an equal number of men and women,

- strive for an almost uniform age distribution.

Data collection

Persons are asked to take the driver or front passenger seat and to touch center console components in a certain order. Then the person changes seats and repeats the procedure. The sequences of components are different for driver and front-seat passenger. The in-car camera starts recording just before the first interaction happens so that the first frame in each video shows a motionless person. The main drawback of the data gathering scenario is that the opposite seat to the active one is only occasionally occupied, drastically simplifying the identification of an active person. In order to avoid ambiguity the simultaneous interaction with center console components by driver and passenger is precluded.

All collected video frames have been manually annotated. The meta-information about the environment such as weather conditions, time of day, position of the sun regarding the car, car location etc. as well as the currently touched component is embedded directly into video frames. The actual evaluation of the recognition performance is a frame-by-frame comparison of the algorithm's decision with meta-information and counting discrepancies.

The experimental database contains 16 persons. Among them are males and females of different ages and body heights. Persons wear different coats, jackets, shirts and gloves, sometimes in the color of the vehicle interior. Totally, 23825 video frames have been collected including 10054 frames with uniform and 13771 frames with non-uniform illumination conditions. Some samples from the database are presented in Figure 5.2.

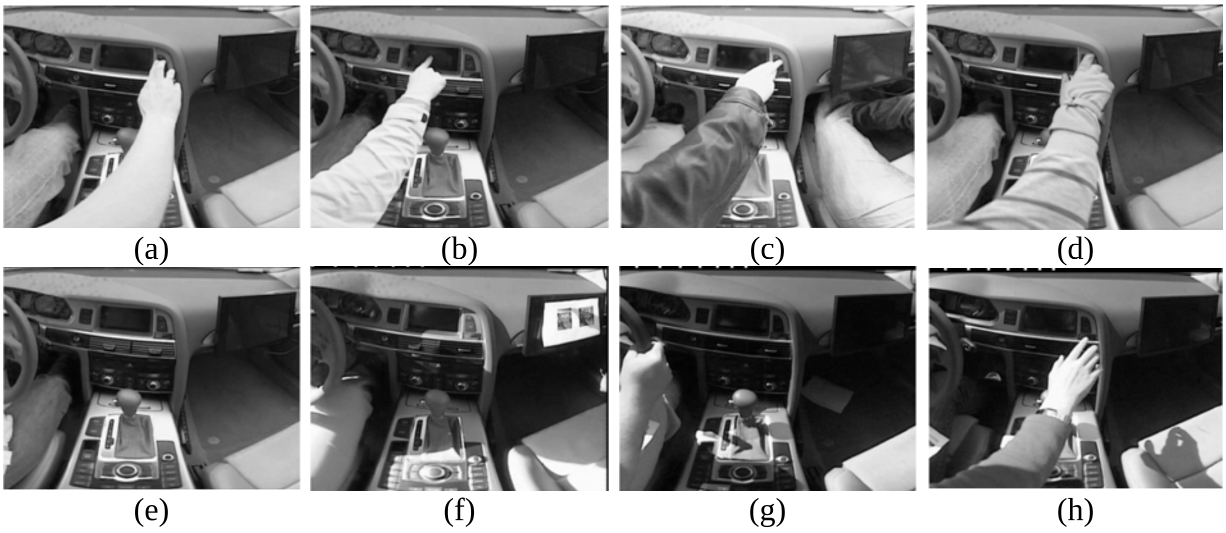


Figure 5.2: UDiS1, top row – different clothing: (a) t-shirt, (b) white textile jacket, (c) black leather jacket, (d) glove and jacket in interior color; bottom row – illumination variations: (e) uniform illumination/diffuse lighting; (f) intensive sunlight from the rear left; (g) intensive sunlight from the left; (h) shadow on passenger seat (reprinted from [108])

Since the real car from the VisMes project is utilized outdoors where the illumination conditions are uncontrolled, splitting into uniformly and non-uniformly illuminated scenes is important to study the reactions of algorithms to drastic changes of environmental lighting. Night samples are not included in the experimental database.

Performance measures

In this very first experiment the frame-based evaluation model is utilized. The system simultaneously decides for each frame firstly whether a driver touches any of the components and secondly whether a passenger does. Hypothesis testing is adopted to evaluate the outcomes of the motion-based hand segmentation proposed in Section 4.2.8. Missing the active hand is reckoned to be critical error because the vehicle is not aware of the originator of the action which may lead to an incorrect reaction of the system. In contrast, falsely recognized actions are not critical because if no components are touched, no reaction is required and the falsely recognized actions can be ignored. Hence, the null hypothesis for a driver is posed as: “the driver touches one of the center console components” and the alternative hypothesis is that he/she does not. Similarly, the null hypothesis for a passenger is posed as: “the passenger touches one of the center console components” and the alternative hypothesis is that he/she does not.

False positive rate (FPR), which is the rejection of the null hypothesis when it is true, determines how often the system misses actions. False negative rate (FNR), which is the acceptance of the null hypothesis when it is false, determines how often the system detects absent actions. The formal notations of both error rates are given by equations 5.4 and 5.5.

$$FPR(\tau) = Pr(m < \tau | action) \approx \frac{N^{-}_{action}}{N_{action}} \tag{5.4}$$

$$FNR(\tau) = Pr(m \geq \tau | no - action) \approx \frac{N^{+}_{no-action}}{N_{no-action}} \tag{5.5}$$

Values N_{action} and $N_{no-action}$ denote numbers of frames with and without actions respectively just as values N^{-}_{action} and $N^{+}_{no-action}$ denote numbers of frames with falsely missed actions and falsely detected absent actions respectively. The decisions are made based on the amount of motion in a special region denoted by m. The motion threshold τ is derived from the training data and hard-coded in the algorithm. Hence, both error rates can be calculated only at this threshold. Despite the reasoning about the importance of one or another error rate, both FPR and FNR are considered in the following to be equally important, so that the half-total error rate ($HTER$), which is an average value of FPR and FNR, is used as the scalar measure of the recognition performance.

Experiments

Hypothesis testing for the calculation of error rate is applied twice, individually for driver and front-seat passenger. The results are presented in tables 5.8, 5.9 and 5.10. Table 5.8 contains the error rates resulting from the experiments with the complete data set. Here, the FPR for the driver is 3.20% lower than that for the passenger just as the FNR for the driver is 2.05% higher than that for the passenger. The $HTER$ values yield 15.62% and 16.19% for the driver and passenger respectively. The table additionally shows the numbers of experimental frames providing an idea about the confidence of the resulting error rates.

Table 5.8: UDiS1, results of evaluation of complete data set

	FNR	N_{no_action}	*FPR*	N_{action}	*HTER*
Driver	14.41%	20175	16.82%	3650	15.62%
Passenger	17.61%	22369	14.77%	1456	16.19%

Table 5.9 contains the results of the experiments with uniform and non-uniform illumination conditions separately. Splitting uniformly and non-uniformly illuminated scenes is important to reveal the eventual robustness of the algorithms against drastic lighting variations. The $HTER$ values resulting from the tests with uniformly illuminated scenes approach 11.31% and 11.43% for driver and passenger respectively. For non-uniformly illuminated scenes the corresponding $HTER$ values approach 18.77% and 19.68%. Based on these values it can be asserted that the motion-based hand segmentation is not robust enough, and the recognition performance significantly suffers from non-uniform illumination conditions.

Table 5.9: UDiS1, results of the separate evaluation of uniformly and non-uniformly illuminated scenes

Illumination		*FNR*	N_{no_action}	*FPR*	N_{action}	*HTER*
Uniform	Driver	7.35%	8671	15.26%	1383	11.31%
	Passenger	10.70%	9437	12.16%	617	11.43%
Non-uniform	Driver	19.75%	11504	17.78%	2267	18.77%
	Passenger	22.66%	12932	16.69%	839	19.68%

Summary of the experiments on UDiS1

The results of the experiments are not encouraging. The reason for this is the simplicity of the decision making algorithm considering only the motion in small regions assigned to driver and passenger. These regions are shown in Figure 5.3 in red. The regions of the center console components are marked with green frames. It can be clearly seen in Figure 5.3f that the hand in the MMI region does not trigger any motion in driver and passenger regions. The proposed algorithm is not able to make a correct decision for camera frames showing the interaction with the MMI knob.

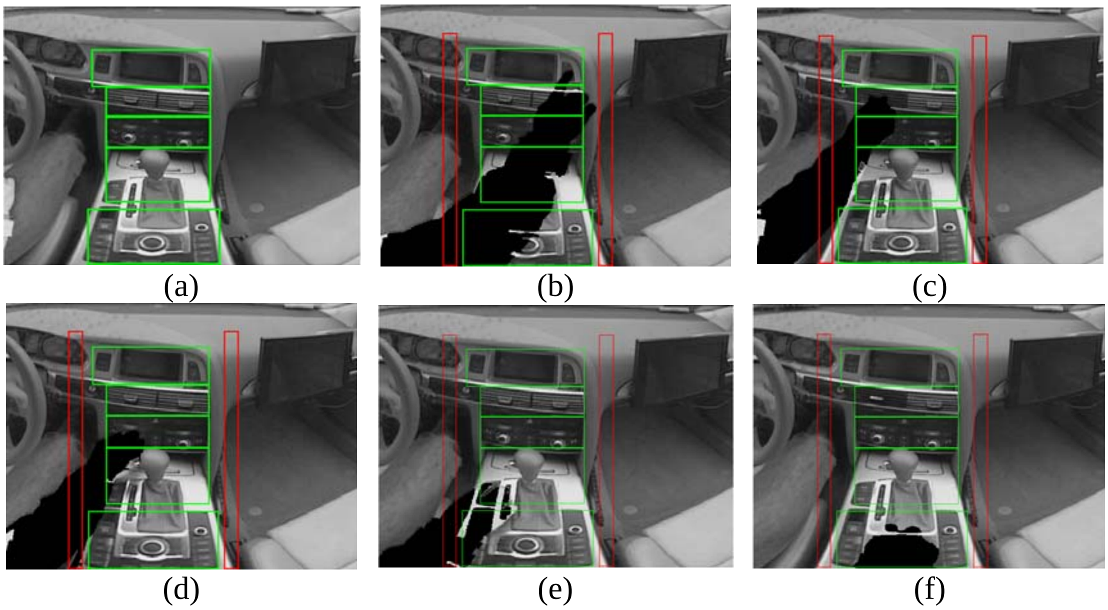

(a) (b) (c) (d) (e) (f)

Figure 5.3: UDiS1, (a) Regions of center console elements; Segmented hand and forearm shape (b) in navigation region, (c) in fan region, (d) in air-conditioning region, (e) in transmission region and (f) in MMI region (reprinted from [108])

Another source of errors is intensive sunlight from the side causing dynamic shadows of the moving hands in driver and passenger regions (see Figures 5.2f and 5.2h). These shadows are often misinterpreted as interactions of driver/passenger with center console components. Unfortunately, shadows can neither be filtered by the motion detection engine nor by background subtraction.

The major reason for observing high error rates is, however, the frame-based decision making which causes ambiguous interpretation of the term “interaction” bearing in mind interaction between hand and center console components. Only video frames showing a person touching a console component have been annotated as an interaction. All other frames have been annotated as no interaction occurs. In fact, the hand moving towards the center console is a part of an interaction. Video frames showing these moving hands are correctly recognized by an algorithm as an interaction. However, these actually correct decisions mismatch the ground truth and are, therefore, judged as false decisions resulting in the high FNR. Consequently, the frame-based evaluation is recognized to be improper. This is why next versions of the user discrimination systems are designed with a focus on the action-based decision making and proper unambiguous evaluation.

5.2.2 User discrimination in the AMSL car simulator: first experiment (UDiS2)

Aiming at overcoming the drawbacks of the first experimental setup, by further developing the user discrimination system the focus is on three aspects:

- switch from uncontrolled illumination conditions to the controlled simulation of intensive frontal and side sunlight as well as uniform environmental lighting,
- improve the imaging algorithm by fusing motion and texture analysis,
- replace frame-based decision making with action-based decision making to avoid any ambiguities by evaluating the recognition performance.

In order to satisfy the first aspect, the experiments are carried out in the laboratory using the AMSL car simulator (for details see Section 4.2). The daylight in the laboratory is simulated with the help of two ceiling lamps and the sunlight coming from the window. For the twilight scenario the ceiling lamps are turned off and the sunblinds are partially closed, so that only faint sunlight passes into the room. The night scenario implicates the complete absence of lighting outside the car simulator. The strong side illumination is simulated by the 100W light bulb. The lamp is placed at the same level with the car roof and carried out at a distance of approximately one meter from the car simulator facing the passenger compartment.

The second aspect is gained by complementing the motion-based decision making algorithm with determining the principal axis of the bounding box around the active arm. The principal axis points out to the active person.

For the implementation of the third aspect, the semi-automatic annotation of camera frames by highly reflective tokens is carried out. By recording evaluation videos, the tokens are uncovered when a person provides an action and covered otherwise. The interaction with center console components is replaced by the interaction with a radio navigation system which is used to simulate the dual-view touch screen. One important limitation is assumed, namely that driver and passenger do not interact with the radio navigation system simultaneously. This assumption discards the need for two parallel decisions for driver and passenger and allows for each action to decide whether the driver or passenger interacts. The points at time when an action starts and ends as well as the action originator are known thanks to highly reflective tokens. I realize that parallel interaction with the dual-view touch screen is a realistic scenario, but the analysis of overlapping hands is beyond the scope of this thesis.

Data collection

One experiment implies the participation of three persons, two test persons in driver and passenger seats and a supervisor. Test persons are instructed to provide the following actions: first, the driver touches four different corner buttons of the radio navigation system and the passenger remains motionless, then the passenger touches these four buttons and the driver remains motionless. The same procedure is repeated one more time. The supervisor covers and uncovers the highly reflective tokens in corners of the camera frame to indicate an active person. Afterwards the test persons change seats and provide the same actions. Hence, an evaluation video contains 16 actions from each person, 8 in a driver and 8 in a passenger seat. There are 10 test persons engaged. Thus, the final database consists of 10 videos and 160 (80 driver and 80 passenger) actions for each evaluation scenario. In practice, some errors during recording have occurred, so that the actual number of assessable actions is slightly less. Apart from the evaluation videos, 7 training videos with 112 actions have been recorded. These are used to define decision threshold τ introduced in Section 4.2.9. With the threshold set up to 0, all actions in training videos have been recognized without error.

Performance measures

Since the user discrimination system reacts to the action, the null hypothesis is posed as "the driver is the originator of the action" and the alternative hypothesis as "the passenger is the originator of the action". Theoretically, the hypotheses can be swapped because I-type and II-type errors are equally important. Since the terms false positive rate and false negative rate are confusing here, the errors rates are denoted as driver error rate (ER_d) and passenger error rate (ER_p) implying that a passenger interaction is recognized given that a driver is actually interacting, and a driver interaction is recognized given that a passenger is interacting, respectively. Formal notations of error rates are given by equations 5.6.

$$ER_d(\tau) \approx \frac{E_d(\tau)}{N_d}, \quad ER_p(\tau) \approx \frac{E_p(\tau)}{N_p} \tag{5.6}$$

The value E_d denotes the number of driver actions interpreted as passenger actions and the value E_p denotes the number of passenger actions interpreted as driver actions. Values N_d and N_p denote the total numbers of driver and passenger actions respectively. Decision threshold τ is a hard-coded parameter derived from the experiments with training samples. For the more convenient estimation of the recognition performance, two aforementioned error rates are combined in average error rate (ER) given by Equation 5.7.

$$ER(\tau) \approx \frac{E_d(\tau) + E_p(\tau)}{N_d + N_p} \tag{5.7}$$

Experiments

The recognition performance of an imaging system in a car passenger compartment can be affected by numerous environmental conditions. It could be, for instance, color and material of the car interior, passenger clothing, obstructive movements in the background and personal habits of car occupants to interact with the console. However, the most important factor is the uncontrolled illumination which is the inherent problem of imaging systems influencing the algorithm's ability to successfully link the action to a driver or front-seat passenger. For this reason and because of the impossibility to cover all factors, the evaluation focuses solely on different illumination conditions.

The experiments are organized in a way to address three research questions:

1. Do we need infrared illumination during daytime? The test reveals the difference in the recognition performances with and without active near-infrared illumination sources in the interior.
2. Does the algorithm demonstrate similar recognition performances at different day times (day, twilight and night)? The test reveals how the error rates change when twilight lighting and the night are simulated.
3. Does the intensive illumination from the side diminish the algorithm's recognition performance? Here, the results of the tests with diffuse illumination conditions are compared with those of tests with strong directional illumination. In particular, the reaction of the algorithm to dynamic shadows is investigated.

Test 1: Infrared illumination during daytime The first scenario evaluates the difference of recognition performance during daytime with and without active near-infrared (NIR) illumination. It is assumed, that the skin reflectance brought about by NIR light makes a hand brighter and, therefore, more distinctive in comparison to the background. Consequently, the image processing algorithm is assumed to segment a hand more easily. However, the turned on NIR LED lamp consume additional power. Thus, the employment of the lamp has to provide notable improvement of the recognition rates. Figure 5.4 demonstrates an effect of NIR illumination during daytime.

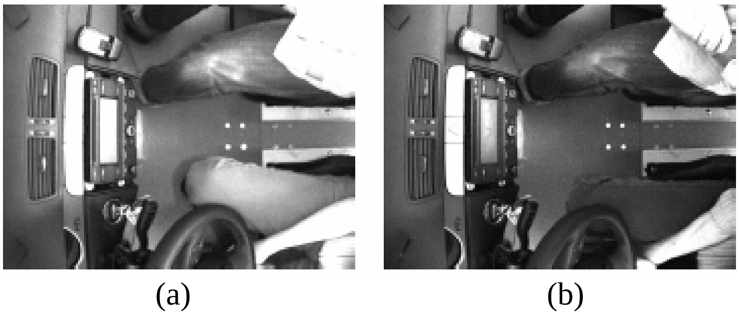


Figure 5.4: UDiS2, effect of near-infrared illumination: (a) lamp turned on, (b) lamp turned off (reprinted from [159])

The results of the first test are presented in Table 5.10. The error rates for a driver are significantly higher than those for a passenger regardless of the status of the NIR lamp. This difference is the result of different behaviors of driver and passenger. While the driver instinctively rotates the steering wheel, the passenger stays motionless most of the time. Additional movements give rise to unconventional shadows and motions in unexpected regions. The average error rates (as defined in Subsection 5.2.2) yield 10.42% and 18.47% with and without NIR illumination respectively. Due to this drastic reduction of the recognition performance to approximately 75%, NIR illumination proves to be necessary around the clock. Based on these results, the NIR lamp is permanently turned on in both succeeding tests.

Test 2: Daytime Since the user discrimination system is designed to operate around the clock, the equally high recognition performance is expected even if sunlight is limited or absent. Figure 5.5 shows samples of day, twilight and night camera views. In general, these situations differ by the amount of outside light penetrating into the car passenger compartment.

Table 5.10: UDiS2, results of the first test: "Infrared illumination during daytime"

		Tests	Correct	Fault		*ER*
NIR lamp on	Driver	73	62	11	ER_d = 15.07%	10.42%
	Passenger	71	67	4	ER_p = 5.63%	
NIR lamp off	Driver	80	58	22	ER_d = 27.50%	18.47%
	Passenger	77	70	7	ER_p = 9.09%	

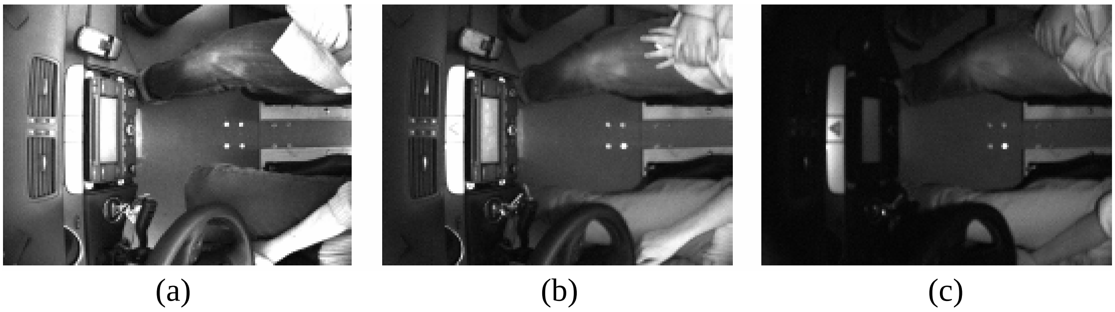


Figure 5.5: UDiS2, camera views at different daytimes: (a) day, (b) twilight and (c) night (reprinted from [159])

The results of the second test are presented in Table 5.11. Similarly to the first test, the asymmetry in driver and passenger error rates can be stated. The average error rate grows from 10.42% in the day to 15.38% in the twilights, which corresponds to the increase of approximately 50%. The night scenario brings the recognition algorithm to its limits. The edges of forearm and hand cannot be detected because of the lack of illumination. In contrast, the strongly reflecting materials are easily distinguishable. A motion-based decision algorithm detects only noise, which is mostly presented in the passenger area. Due to these facts the recognition algorithm constantly assigns all actions to a passenger. Hence, the error rate for the passenger's actions is always 0%, while the error rate for the driver's actions is always 100%. The main conclusion from this test is that the

Table 5.11: UDiS2, results of the second test: "Daytime"

		Tests	Correct	Fault		*ER*
Day	Driver	73	62	11	ER_d = 15.07%	10.42%
	Passenger	71	67	4	ER_p = 5.63%	
Twilight	Driver	66	51	15	ER_d = 22.73%	15.38%
	Passenger	64	59	5	ER_p = 7.81%	
Night	Driver	1	0	1	ER_d = 100.00%	50.00%
	Passenger	1	1	0	ER_p = 0.00%	

Test 3: Intensive side lighting Straight sunlight, when it intensively comes from one direction or while driving along a tree-lined avenue, provokes dynamic shadows and irregular illumination in a car passenger compartment. Due to the overexposure of large areas, the edges of objects cannot be correctly detected and the silhouettes of moving objects are often distorted. This test addresses three cases: front, left and right side lighting (see Figure 5.6 for examples). Rear side lighting is not considered because the sunlight usually does not reach the dashboard area.

Table 5.12 presents the results of the third test. Bizarrely, the error rates resulting from the test with the left side lighting are significantly better than those with front and right side lighting.

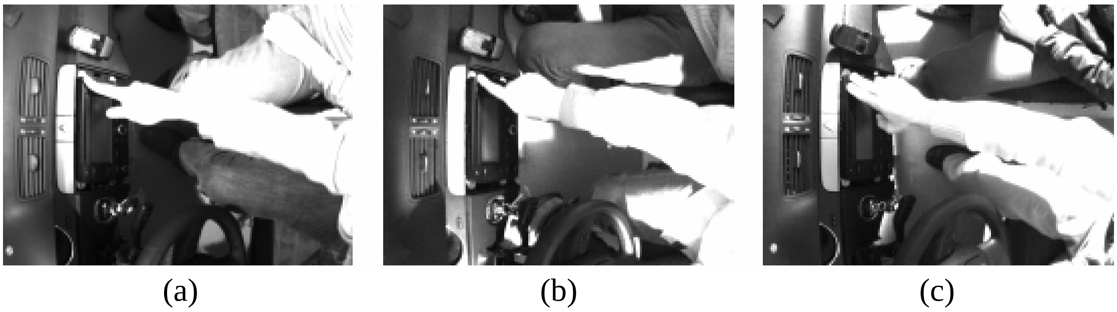


Figure 5.6: UDiS2, intensive lighting from: (a) front side, (b) left side, (c) right side (reprinted from [159])

Unexpectedly, they are even better than in the second test with the twilight scenario and also better than in the first test with the NIR lamp switched off. In contrast, the error rates resulting from the tests with front and right side lighting are similarly high and significantly worse than those in the twilight or day scenario. The asymmetrical error rates for driver and passenger can be stated for the front and left side lighting scenarios. Here, the error rates for a passenger are significantly lower. This asymmetry is caused by the extra shadows from the moving right hand of a driver, which is normally placed on the steering wheel.

Table 5.12: UDiS2, results of the third test: “Intensive side lighting”

		Tests	Correct	Fault		*ER*
Frontal lighting	Driver	69	50	19	ER_d = 27.54%	19.59%
	Passenger	79	69	10	ER_p = 12.66%	
Left side lighting	Driver	80	67	13	ER_d = 16.25%	11.69%
	Passenger	74	69	5	ER_p = 6.76%	
Right side lighting	Driver	86	69	17	ER_d = 19.77%	20.00%
	Passenger	84	67	17	ER_p = 20.24%	

Summary of the experiments on UDiS2

The second version of the used discrimination system has two major drawbacks. The first one is its inability to properly operate at night, requiring an adjustment of active illumination sources inside of the car simulator. The second drawback is the semi-automatic annotation of camera streams by the manual covering and uncovering of highly reflective tokens. In fact, the tokens are not as reflective as assumed, so that they cannot be found in some frames diminishing the number of assessable actions in the corresponding scenarios.

The imaging algorithm has to be improved as well. The discrete matching scores of the implemented system yield values -1, -0.5, 0, 0.5 and 1. The value 0 means the inability of the system to choose between driver and passenger. Values ±0.5 designate the breakdown of one of two algorithms and values ±1 correspond to a confident decision of the system. In fact, the fusion of both motion and texture analysis algorithms is provided on the decision level. Fusion of real-valued matching scores, derived from the amount of motion in a special region and from the angle of the principal axis of an arm, can lead to more flexible decision making because the uncertainty of one algorithm can be better compensated by the confident decision of another algorithm.

5.2.3 User discrimination in the AMSL car simulator: second experiment (UDiS3)

The third version of the user discrimination system is developed to overcome the drawbacks of the second version by addressing the following aspects:

- solve the night image acquisition problem by improving active in-car illumination,
- improve the acquisition setup avoiding any ambiguities when evaluating recognition performance,
- improve the imaging algorithm,
- extend the experimental database by engaging more persons.

The first aspect is realized by installing the 880 nm lamp that is a valuable addition to the 940 nm lamp because the imaging sensor has a low sensitivity in the high range of near-infrared diapason (see Figure 4.15), so that the light produced by 940 nm LEDs is imperceptible for the most part.

The second aspect is satisfied by replacing the radio navigation system with a touch screen. The touch screen is synchronized with the acquisition system. Buttons and slide bars are consecutively shown on the touch screen in accordance with predefined protocol. The sequences regarding how items appear correspond to real interactions. Driver and passenger align their seat positions, adjust the air conditioning and seat heaters, assign a route in the navigation system, select a song in the infotainment system, and adjust the volume. The system is aware who is currently interacting with the touch screen because drivers have been instructed to touch only green items and passengers only red items. All touch screen interactions are stored in the events log that is synchronous with the video stream.

The third aspect is gained by adding the hand detection functionality based on the Viola-Jones object detector [242]. The imaging system detects driver or passenger hands while moving towards the touch screen. In the period between a control component appearing on the touch screen and the touch screen being contacted by the user, the recognition algorithm registers the position of the first and the last hand motion frames as well as the amount of movement in driver and passenger regions. This information allows identifying which user is currently active.

Performance measures are identical to those addressed for UDiS2. For details see Subsection 5.2.2.

Data collection

The number of persons taking part in the experiments is increased to 20. Each person acts once as a driver and once as a passenger. Since both front seats are occupied, each evaluation scenario includes 20 videos. Driver and passenger interchangeably interact with the touch screen maximally approaching real-life interactions defined in the new acquisition protocol (see Subchapter 4.2.4). Every video contains 45 actions, 28 from a driver and 17 from a passenger. The videos are divided into training and test subsets. The training set includes the videos with 6 persons and the test set the videos with the remaining 14 persons. The training videos are used for cropping hand patterns to train the hand classification cascade as well as for tuning motion thresholds. The test videos form the basis for the estimation of the recognition performance. The total number of actions in test videos per scenario is 630 including 392 driver actions and 238 passenger actions.

Experiments

The research questions do not deviate from those posed in Subsection 5.2.2 for the second version of the user discrimination system (UDiS2).

Test 1: Infrared illumination during daytime Table 5.13 presents the results of the first test. In both scenarios with and without active NIR illumination the error rates are at the same level. So, the NIR lamps do not have any tangible influence on the recognition performance of the hand detection algorithm and can be theoretically switched off. Notably, the algorithm tends to vote rather for a driver providing significantly more misclassifications for a passenger. The hand movement from the steering wheel to the touch screen is easily detectable. In contrast, the left hand of a passenger is often out of camera view (see Figures 5.7 and 5.9). When a fast movement toward a touch screen arises, the first frame with the detected hand can be located very closely to the touch screen revealing no direction of movement. If the picture to touch is located in the left half of the touch screen, the action will be probably misclassified to belong to a driver. Figure 5.7 shows the samples of camera frames with and without active NIR illumination.

Table 5.13: UDiS3, results of the first test: "Infrared illumination during daytime"

	Driver as passenger (ER_d)	Passenger as driver (ER_p)	Average error rate (ER)
NIR lamps turned on	16/392 (4.08%)	36/238 (15.13%)	52/630 (8.25%)
NIR lamps turned off	15/392 (3.83%)	41/238 (17.23%)	56/630 (8.89%)

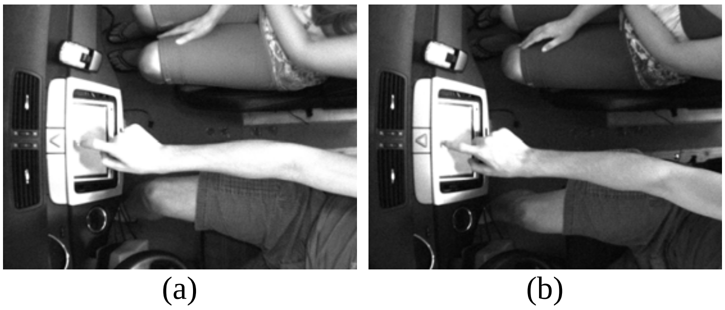


Figure 5.7: UDiS3, Effect of active near-infrared illumination: (a) lamps turned on; (b) lamps turned off (reprinted from [107])

Test 2: Daytime The results of the second test are presented in Table 5.14. The 880 nm lamp, installed as an addition to the 940 nm lamp, has a great effect on the twilight and night images as shown in Figure 5.8. Thanks to NIR illumination the error rates in the twilight scenario ($ER =$ 10.79%) and in the night scenario ($ER =$ 10.00%) are only slightly higher than those in the day scenario ($ER =$ 8.25%) which proves the utilization of NIR lamps to be a suitable solution for the case of deficient sunlight. Unexpectedly, the recognition performance at night is slightly better compared to that in the twilight. However, the difference of 0.79% is insignificant. The tendency of the algorithm to recognize passenger actions as driver actions rather than otherwise remains unchanged.

Test 3: Intensive side lighting Table 5.15 presents the results of the third test. Unexpectedly, the error rates resulting from scenarios with left and right side lighting are significantly different. While with the right side lighting, the tendency to misclassify passenger actions to belong to a driver is preserved, with the left side lighting both driver and passenger error rates are similarly high, exceeding 13%. This phenomenon is reflected in asymmetrical average error rates which yield ER of 14.60% with left side and ER of 8.41% with right side lighting. It can be explained only by

Table 5.14: UDiS3, results of the second test: “Daytime”

	Driver as passenger (ER_d)	Passenger as driver (ER_p)	Average error rate (ER)
Day	16/392 (4.08%)	36/238 (15.13%)	52/630 (8.25%)
Twilight	20/392 (5.10%)	48/238 (20.17%)	68/630 (10.79%)
Night	26/392 (6.63%)	37/238 (15.55%)	63/630 (10.00%)

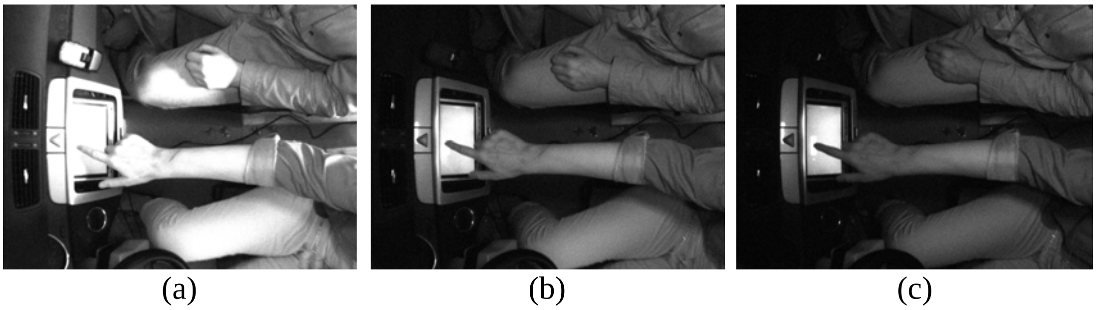


Figure 5.8: UDiS3, Different times of the day: (a) day; (b) twilight; (c) night (reprinted from [107])

some unusual driver activities in one or several test videos. Notably, the recognition performance with the strong left side lighting is similar to that with the uniform lighting. This fact indicates the user discrimination algorithm based on hand detection as being fairly robust against intensive lateral lighting. Figure 5.9 shows sample frames with diffuse, left side and right side lighting.

Table 5.15: UDiS3, results of the third test: “Intensive side lighting”

	Driver as passenger (ER_d)	Passenger as driver (ER_p)	Average error rate (ER)
Diffuse	16/392 (4.08%)	36/238 (15.13%)	52/630 (8.25%)
Left side	53/392 (13.52%)	39/238 (16.39%)	92/630 (14.60%)
Right side	10/392 (2.55%)	43/238 (18.07%)	53/630 (8.41%)

Summary of the experiments on UDiS3

The experiments have shown that distinguishing between driver and passenger interactions based solely on hand detection is not reliable. Combining hand detection with the amount of motion in driver and passenger regions improves recognition performance, but error rates are still too high. The classifier trained for hand detection is by far not optimal. Misdetections of hands as well as missing hands occur very often requiring additional filtering of the detected regions. Hand tracking may significantly contribute to the reliable determination of the trajectory of a moving hand.

Conclusion

Three prototypes of an optical recognition system for distinguishing between driver and front-seat passenger while interacting with the center console have been evaluated.

The first version of the system (UDiS1) is implemented in a real car. The experiments are carried out outdoors under non-controlled uniform and non-uniform illumination conditions. The recognition performance is evaluated separately for driver and passenger by counting frames with missed and falsely detected interactions. The decision if an interaction occurs is made based on the

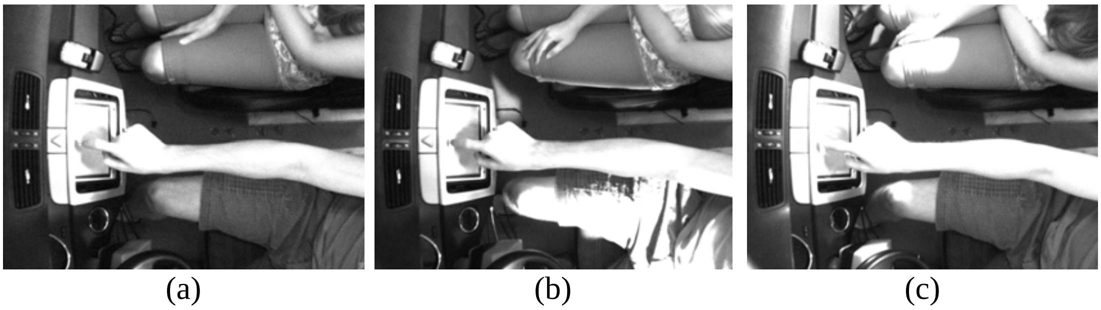


Figure 5.9: UDiS3, Intensive side lighting: (a) diffuse lighting as a reference, (b) left side lighting, (c) right side lighting (reprinted from [107])

amount of motion in specific driver and passenger regions. Recognition performance is measured in terms of $HTER$ yielding 11.31% and 11.43% for driver and passenger respectively.

In contrast to the first version, the second version of the system (UDiS2) is designed to enable the control over illumination conditions. The acquisition system is implemented in a car simulator in the laboratory and the different lighting conditions are simulated by laboratory lamps and sunlight coming from windows. In tests, the day, twilight and night lighting conditions are reproduced as well as intensive side lighting from left and right. The imaging algorithm is based on the fusion of the motion analysis with edge detection and the subsequent silhouette analysis. Here, the system decides whether driver or passenger provides an action excluding simultaneous actions of both. By evaluating recognition performance the frame-based approach is replaced by the action-based approach. The average error rate gained with uniform lighting yields 10.42%.

The third version of the system (UDiS3) is an enhancement of the second one by providing the credible evaluation setup. This includes the automatic annotation of camera frames and implementing realistic scenarios for the interaction between occupants and the touch screen. The imaging algorithm is based on hand detection supported by measuring the amount of motion in driver and passenger regions. The average error rate of 8.25% resulting from the test with diffuse lighting is clearly too far away from the required standards to admit the integration of the proposed system in an actual automobile, but the achieved recognition performance with simple imaging algorithms can be considered a very promising result for an optical system.

Future work should be devoted to fusing image processing algorithms, migrating the prototypic user discrimination system from the car simulator to a real car, and carrying out experiments with the moving car.

5.3 General face recognition performance on XM2VTS

This section is comprised of the following subsections:

1. XM2VTS database and Lausanne protocol
2. Performance measures
3. Experiments with FaceART
4. Experiments with Luxand FaceSDK
5. Performance of alternative face recognition systems
6. Conclusion

5.3.1 XM2VTS database and Lausanne protocol

The XM2VTS database is a widely accepted standard within academia for the evaluation of face recognition systems. This database has been created at the University of Surrey using a high-quality camcorder and digital video cassette recorder. The data is stored in color videos with a resolution of 720x576 pixels (for more details check [170]). The database includes 295 persons whose faces have been captured with uniform illumination conditions and a monotone constant background. The persons have visited the recording studio four times at intervals of approximately one month. On each visit (session) two recordings (shots) have been made. The first recording is a talking head shot and the second is a rotating head shot. The database is intended for multi-modal user authentication with three modalities: face, speech and lip movement. At the moment, XM2VTS comprises one of the highest numbers of users among public face databases.

It is important to note that the acquisition conditions set up by gathering the data in XM2VTS obviate the need for pre-processing such as face localization, light normalization, masking of the background or even face scale adjustment. In fact, there are some variations in pose and location, but the vast majority of images contain frontal faces of the same size at the same location. These facts let researchers focus on feature extraction and matching steps. This is why XM2VTS has been chosen to evaluate the general recognition performance of two face recognition systems (FaceART, Luxand FaceSDK) and compare it to the recognition performances of some established academic face recognition systems.

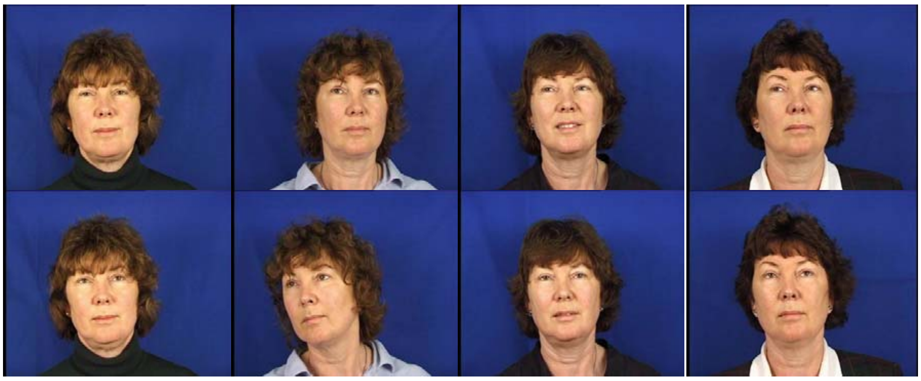


Figure 5.10: Samples from the Lausanne protocol: top row - talking head shot, bottom row - rotating head shot

The Lausanne protocol has been developed by IDIAP Research Institute in 1998 to provide a standard for evaluation of face recognition algorithms based on common performance assessment methodology. The protocol addresses frontal faces with blank expressions. Two images are cut from XV2VTS recordings, the first one at the beginning of the talking head shot and the second one from the middle of the head rotation shot. Hence, the protocol operates with 2360 color images that include faces of approximately 200x300 pixels (see Figure 5.10 for sample images). The users are partitioned randomly into one of three groups: 200 clients, 25 impostors for evaluation and 70 impostors for testing. One person cannot belong to two or more groups. The listing of clients and impostors can be found in [154].

There are two configurations within the Lausanne protocol combining client images in training and evaluation subsets each in a different manner. However, the subsets of impostor images for evaluation and testing are the same in both configurations. Configuration 1 addresses “good expert training”. Here, the first shots from the first three sessions are used for training, the second shots from the first three sessions for evaluation, and both shots from the fourth session for testing. So, the training set contains 600 images. Configuration 2 addresses “inferior expert training”. Here, both shots from the first two sessions are used for training, both shots from the third session for evaluation, and both shots from the fourth session for testing. So, the training set contains 800 images.

Table 5.16: XM2VTS, Lausanne protocol, partitioning of client and impostor samples

Configuration 1				
Session	Shot	200 Clients	25 Impostors	70 Impostors
1	1	Training	Eval.	Test
	2	Eval.		
2	1	Training		
	2	Eval.		
3	1	Training		
	2	Eval.		
4	1	Test		
	2			

Configuration 2				
Session	Shot	200 Clients	25 Impostors	70 Impostors
1	1	Training	Eval.	Test
	2			
2	1			
	2			
3	1	Eval.		
	2			
4	1	Test		
	2			

Table 5.16 visualizes the partitioning of client and impostor samples regarding protocol configurations. Training samples form the basis for training user models stored in the reference database. Evaluation samples are utilized for computing decision thresholds and estimating optimistic recognition performance. Test samples are applied for computing effective recognition performance. Regarding the verification scenario proposed by the Lausanne protocol, performance evaluation is a cross-matching of reference user models and evaluation/test samples. The numbers of matching scores resulting in each configuration can be calculated as shown in Table 5.17.

Table 5.17: XM2VTS, Lausanne protocol, numbers of matching scores

Configuration 1	Configuration 2
Evaluation:	
Client scores: 200 clients models * 3 evaluation shots = **600** Impostor scores: 200 clients models * 25 impostors * 8 impostor shots = **40000**	Client scores: 200 clients models * 2 evaluation shots = **400** Impostor scores: 200 clients models * 25 impostors * 8 impostor shots = **40000**
Test:	
Client scores: 200 clients models * 2 evaluation shots = **400** Impostor scores: 200 clients models * 70 impostors * 8 impostor shots = **112000**	

5.3.2 Performance measures

Statistical hypothesis testing is a perfect evaluation model for a biometric system which operates in the verification mode. The null hypothesis is posed as “the person is an impostor” and the alternative hypothesis as: “the person is a client”. The critical error (reject null hypothesis when it is true) is called false accept rate (FAR) implying the probability to accept an impostor. Less critical error (accept null hypothesis when it is false) is called false reject rate (FRR) implying the probability to reject a client. The thorough description of these error rates is given in Subsections 2.3.2 and 2.3.2.

As mentioned in Subsection 2.3.2, when talking about error rates, we actually mean false matching of impostors and false non-matching of clients rather than false acceptance or false rejection. Notice, the matching is done by the matching algorithm just as the decision to accept/reject a user is made by the application. There are applications e.g. border crossing control where the match in the ban list is equal to the rejection of a user. However, usually a biometric system is handled as an access control system where the matched users have to be accepted by the system and the non-matched users rejected. So, in our case the false match and false non-match can be called false accept and false reject respectively without loss of generality.

A matching score is considered the test statistic and the decision to accept/reject is made based on the comparison of the matching score with the decision threshold τ. Hence, FAR and FRR are functions of the threshold τ. Equations 5.8 are used to estimate these functions. Values $EI(\tau)$ and I denote the number of accepted impostor trials and the total number of impostor trials respectively. Values $EC(\tau)$ and C denote the number of rejected client trials and the total number of client trials respectively.

$$FAR(\tau) \approx \frac{EI(\tau)}{I}, \quad FRR(\tau) \approx \frac{EC(\tau)}{C} \tag{5.8}$$

The Lausanne protocol prescribes that the evaluation datasets are utilized for determining three thresholds and the corresponding evaluation error rates. With FAE to denote the false accept rate in evaluation and FRE to denote the false reject rate in evaluation, the formal notations for the thresholds are given as follows:

$$\begin{aligned}
&\tau_{FAE=0} = argmin_\tau(FRE(\tau)|FAE(\tau) = 0)\\
&\tau_{FRE=0} = argmin_\tau(FAE(\tau)|FRE(\tau) = 0)\\
&\tau_{FAE=FRE} = (\tau|FAE(\tau) = FRE(\tau))
\end{aligned}$$

In our experiments only the third threshold is addressed. In reference to Subsection 2.3.2, this threshold is denoted by τ_{eer} and the corresponding error rate is referred to as the equal error rate (EER).

During testing, the τ_{eer} is fixed and the FAR and FRR are calculated only at this threshold. The sum of $FAR(\tau_{eer})$ and $FRR(\tau_{eer})$ is used as the scalar indicator of the recognition performance and referred to as the total error rate (TER). The TER resulting from the evaluation is equal to the double EER.

5.3.3 Experiments with FaceART

By carrying out the experiments with the FaceART face recognition system with XM2VTS, three research objectives are in focus:

- separately evaluate matching modes to discover the best internal parameters within each mode,
- compare matching modes to each other to determine the best FaceART configuration in reference to the Lausanne protocol,
- dispose FaceART into the state of the art of face recognition systems by comparing the recognition performance of FaceART (with the best configuration) to that of its competitors.

As introduced in Subsection 4.3.6, FaceART can operate in one of four matching modes (see Figure 4.32). In the first mode, test face images are directly matched with reference face images. The second mode extends the first one by integrating the ARTMAP network for the classification of raw face images. In the third mode, the raw face images are substituted by the coefficients of eigenfaces. In the fourth mode, the coefficients of eigenfaces are learned by the ARTMAP network.

It is important to notice that pre-processing is reduced to face detection which is implemented on the basis of the Viola-Jones algorithm [242]. Best-fit-plane subtractions and histogram equalization are omitted because of the perfect acquisition setup of XM2VTS.

First matching mode: raw images classified by k-NN

By evaluating FaceART in the first matching mode, two parameters are addressed that seem to be essential here: distance metric and the size of facial images. In fact, the matching algorithm (k-NN) has two parameters: the number of neighbors (k) and the distance metric. Due to the low number of reference images (three in the 1st configuration and four in the 2nd configuration) and the absence of outliers in the training set, there is no need for k being higher than 1. In contrast, the distance metric is worth evaluation. For this test the size of facial images is set to 50x50 pixels. Four metrics implemented within FaceART (for details see Subsection 4.3.7) are compared: city-block distance, Euclidean distance, max-distance and cross-correlation. According to the results reported in Table 5.18, the city-block distance is the best metric followed by the Euclidean distance delivering comparable results. The error rates with cross-correlation are significantly higher than those with city-block and Euclidean distance. The max-distance is inappropriate for this matching mode which is proven by the resulting TER values over 69%.

Table 5.18: FaceART, first matching mode, evaluation of distance metrics (reprinted from [161])

Metric	Configuration 1						Configuration 2					
	Evaluation			Test			Evaluation			Test		
	FAR	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*
city-block	13.88%	13.88%	27.76%	11.72%	12.50%	**24.22%**	13.50%	13.50%	**27.00%**	11.32%	12.75%	**24.07%**
Euclidean	13.71%	13.71%	**27.42%**	11.27%	15.75%	27.02%	15.06%	15.06%	30.13%	12.43%	15.25%	27.68%
max	37.81%	37.81%	75.62%	35.13%	34.00%	69.13%	37.24%	37.24%	74.49%	36.50%	34.50%	71.00%
x-corr.	16.03%	16.03%	32.07%	12.33%	21.25%	33.58%	18.00%	18.00%	36.00%	13.88%	19.00%	32.88%

Since pre-processing and feature extraction are omitted in the first matching mode and the feature vector is represented by a gray-scale facial image of particular size, the size of facial images is the only parameter which can be derived from all steps preceding the matching step. The optimal size of facial images is discovered only with the city-block distance which is considered the best metric. The results of the test are presented in Table 5.19. Among the addressed sizes from 20x20 pixels to 100x100 pixels, the error rates differ very slightly. These minor variations can be associated with rounding errors during the computation of matching scores. To conclude, the size of facial images seems to be of no matter in the first matching mode.

Table 5.19: FaceART, first matching mode, evaluation of the size of facial images (reprinted from [161])

Scale	Configuration 1						Configuration 2					
	Evaluation			Test			Evaluation			Test		
	FAR	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*
20x20	13.38%	13.38%	**26.76%**	11.38%	13.25%	24.63%	13.58%	13.58%	27.15%	11.47%	12.50%	**23.97%**
30x30	13.71%	13.71%	27.42%	11.59%	12.75%	24.34%	13.15%	13.15%	**26.29%**	11.13%	13.25%	24.38%
50x50	13.88%	13.88%	27.76%	11.72%	12.50%	**24.22%**	13.50%	13.50%	27.00%	11.32%	12.75%	24.07%
75x75	13.88%	13.88%	27.76%	11.75%	12.50%	24.25%	13.50%	13.50%	27.00%	11.34%	13.00%	24.34%
100x100	13.83%	13.83%	27.65%	11.67%	12.75%	24.42%	13.50%	13.50%	27.00%	11.33%	12.75%	24.08%

Second matching mode: raw images classified by ARTMAP

In the second matching mode, there are two significant parameters: vigilance of the ARTMAP network and the size of facial images. As introduced in Subsections 3.3.3 and 4.3.7, vigilance is the basic parameter forming the internal structure of the ARTMAP network during training. However, the vigilance parameter has not been evaluated within this matching mode, but is intuitively set to 0.85. The reason is that the training of the ARTMAP network based on images is a computationally intensive task. Testing of several values of the vigilance parameter would take thousands of hours. Due to the compliment coding and quadratic growth of the feature vector size relating to the linear size of facial images, the feature vector space becomes very high-dimensional. For instance, having very small facial images of 20x20 pixels, the dimension of a feature vector becomes 20x20x2=800. Since the first configuration of the Lausanne protocol includes 600 and the second 800 training patterns, with the vigilance of 0.85 the ARTMAP network expands with a vast amount of internal nodes, which leads to extremely slow training. Hence, for facial images larger than 50x50 pixels, the second matching mode becomes impractical. Table 5.20 presents the results of experiments with three sizes of facial images: 20x20, 30x30 and 50x50. Despite the insignificant difference in error rates, it can be stated that larger facial images lead to better verification performance. In comparison to the first matching mode, the error rates become slightly lower (see Table 5.19).

Table 5.20: FaceART, second matching mode, evaluation of the size of facial images (reprinted from [161])

Scale	Configuration 1						Configuration 2					
	Evaluation			Test			Evaluation			Test		
	FAR	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*
20x20	11.25%	11.25%	**22.51%**	10.19%	13.00%	23.19%	11.48%	11.48%	22.96%	11.09%	11.75%	22.84%
30x30	11.59%	11.59%	23.18%	10.33%	13.75%	24.08%	11.37%	11.37%	22.75%	10.26%	12.50%	22.76%
50x50	11.49%	11.49%	22.99%	10.49%	12.50%	**22.99%**	11.21%	11.21%	**22.41%**	10.20%	12.25%	**22.45%**

Third matching mode: eigenfaces coefficients classified by k-NN

As introduced in Subsection 3.2.3, principal component analysis (PCA) is an approach for feature reduction. In appearance-based face recognition, features are single pixels in the image. PCA transforms pixels to the coefficient of the basis transformation. The basis vectors are eigenvectors of the covariance matrix created from some set of training images. In face recognition, these basis vectors are referred to as eigenfaces. Referring to the information criterion, the importance of each eigenface in the new representation of data is proportional to its eigenvalue. Eigenfaces with high eigenvalues determine the general face appearance while eigenfaces with low eigenvalues determine fine details. So, eigenfaces are sorted by importance and only the most informative ones are utilized for data transformation while the remaining ones are ignored. In such a manner, the loss of information is controlled when the reduction of the data dimensionality is required. Nevertheless, regarding a single feature, the amount of information and the discrimination power are not the equal terms. Hence, the most informative eigenfaces do not necessarily possess the best discrimination power.

There are three parameters to be determined in the third matching mode: size of facial images, number of eigenfaces and distance metric for k-NN. Similarly to the first matching mode k is set to 1.

The experiments have shown that the size of facial images is of minor importance when PCA is utilized. This fact is indirectly proven by calculating the amount of information preserved in the

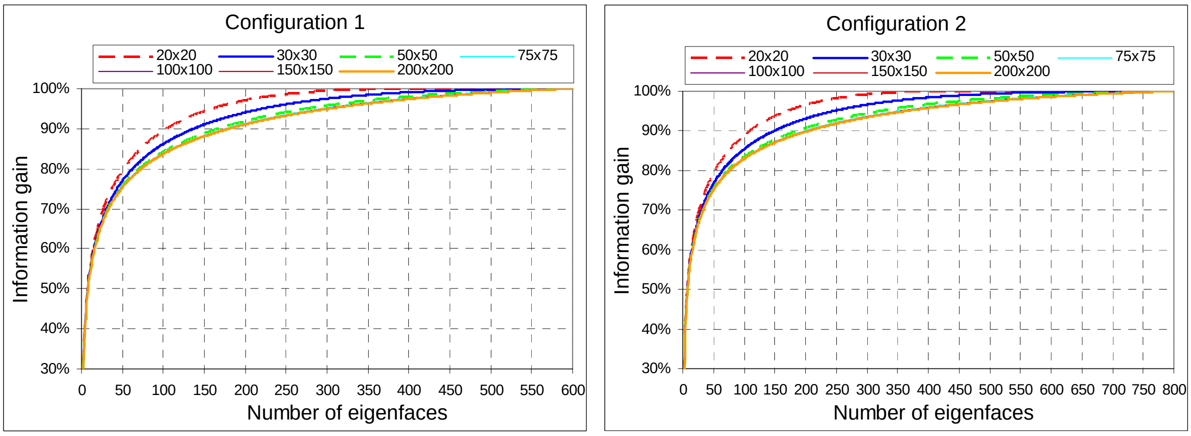


Figure 5.11: FaceART, third matching mode, relationship between the number of eigenfaces and the cumulative information gain brought by eigenfaces with regard to different sizes of facial images (modified from [161])

first eigenfaces. Figure 5.11 illustrates the information gain as a function of the number of first eigenfaces for seven different sizes of facial images (20x20, 30x30, 50x50, 75x75, 100x100, 150x150 and 200x200 pixels). In both configurations of the Lausanne protocol, in case facial images are larger than 50x50 pixels, the curves are almost equal. In fact, the curves corresponding to image sizes of 75x75, 100x100 and 150x150 pixels are covered by the 200x200 curve. For small images the curves are slightly different. From the information theoretical point of view this means that the size of facial images has no significant influence on the PCA transformation coefficients. Due to practical reasons, namely the high operating speed and reasonable face scale, the size of facial images is fixed to 50x50 pixels.

The next question is how many first eigenfaces should be utilized. Considering the example of 20x20 pixel images, 25 eigenfaces preserve approx. 70% of information, 50 eigenfaces preserve approx. 80% of information and 100 eigenfaces preserve approx. 90% of information. For larger facial images the percentage of the information preserved in 25, 50 and 100 eigenfaces is slightly lower. Intuitively, 25 eigenfaces are chosen as a good trade-off between the feature space dimensionality and information reduction. According to Figure 5.11 the amount of information preserved after decomposing 50x50 pixel images by means of 25 first eigenfaces is approximately 66% in both configurations.

The best distance metric is determined by carrying out experiments. Table 5.21 presents the results of the experiments with four metrics: city block distance, Euclidean distance, max-distance and Mahalanobis distance. All metrics except for max-distance generate comparable error rates with almost insignificant differences. In any case, the Mahalanobis distance seems to generate the best results.

By comparing first, second and third matching modes (always considering the best parameterization) it can be stated that eigenfaces coefficients outperform raw gray-scale images independently of whether 1-NN or ARTMAP is applied for classification.

Fourth matching mode: eigenfaces coefficients classified by ARTMAP

The fourth matching mode involves the PCA transformation for feature reduction and the ARTMAP network for classification. Due to the relatively low dimensionality of feature vectors, the ARTMAP network becomes more practical here. Three parameters are investigated: size of facial images,

Table 5.21: FaceART, third matching mode, evaluation of distance metrics (reprinted from [161])

Metric	Configuration 1						Configuration 2					
	Evaluation			Test			Evaluation			Test		
	FAR	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*
city-block	10.04%	10.04%	**20.07%**	9.23%	12.50%	21.73%	11.50%	11.50%	23.00%	10.54%	11.50%	22.04%
Euclidean	12.04%	12.04%	24.08%	11.35%	11.75%	23.10%	11.72%	11.72%	23.44%	10.98%	12.00%	22.98%
max	18.18%	18.18%	36.36%	16.70%	16.00%	32.70%	16.50%	16.50%	33.00%	15.52%	14.75%	30.27%
Mahalano.	10.66%	10.66%	21.33%	9.99%	11.50%	**21.49%**	10.28%	10.28%	**20.57%**	9.78%	12.00%	**21.78%**

number of eigenfaces, and vigilance of the ARTMAP network.

The statement from the previous subsection that the size of facial images is not significant, when eigenfaces decomposition is provided, is confirmed in the first experiment. In accordance with Figure 5.12, if facial images are larger than 50x50 pixels, the TER values remain almost constant or even slightly grow with the increasing image size. Hence, the size of 50x50 pixels is asserted to be optimal and fixed in all further experiments.

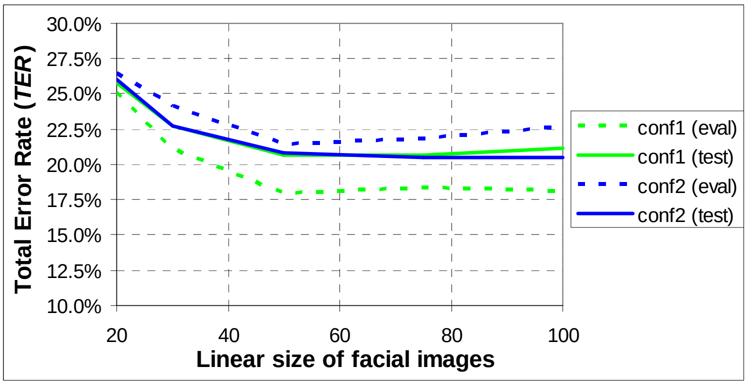


Figure 5.12: FaceART, fourth matching mode, relationship between verification performance (TER) and the size of facial images. The number of eigenfaces is 10 and the vigilance of the ARTMAP network is 0.85 (modified from [161])

The results of the next experiment are presented in Figure 5.13. The diagrams demonstrate that the optimal number of eigenfaces is between 10 and 15. Choosing fewer eigenfaces leads to the rapid reduction of recognition performance caused by the probable loss of user discriminative information preserved by discarded eigenfaces. Similarly, choosing more eigenfaces leads to a gradual increase of error rates due to involving superfluous non-discriminative information preserved by less informative eigenfaces.

The last experiment inside of the fourth matching mode addresses the vigilance of the ARTMAP network. The results of the experiment are shown in Figure 5.14. The curves representing TER as a function of the vigilance parameter possess two meaningful local minima. The first one is resided between 0.75 and 0.85 just as the second one approaches the right margin and can be formally designated as 0.99. As introduced in Subsection 3.3.3 and visualized in Figure 3.20, the vigilance of 1.0 implies the similar partitioning of the decision space as it happens with 1-NN. In this case each training sample becomes one node in the network making the utilization of ARTMAP impractical. Due to the marginal difference in TER values when the vigilance parameter is set to 0.75 or 0.99, the vigilance of 0.75 is asserted to be the best choice leading to the more compact network. The TER values resulting from the test with the vigilance of 0.75 are 16.01%/19.31% and

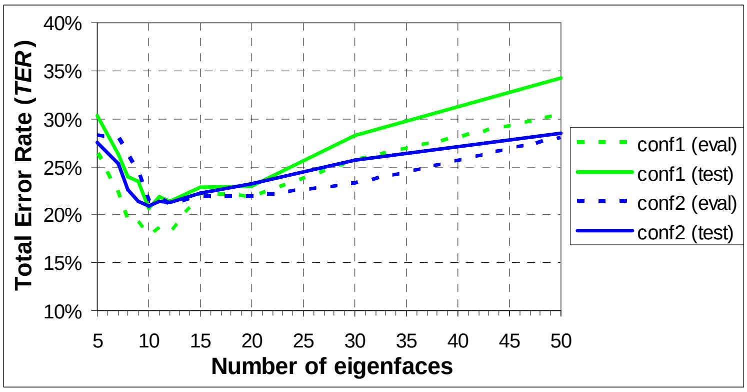


Figure 5.13: FaceART, fourth matching mode, relationship between the verification performance (TER) and the number of eigenfaces. The size of facial images is 50x50 pixels and the vigilance of the ARTMAP network is 0.85 (modified from [161])

21.05%/20.68% for evaluation/testing in the 1st and 2nd configurations of the Lausanne protocol respectively. Note, choosing a high vigilance leads to the risk of overfitting. In contrast, lower vigilance implies better generalization.

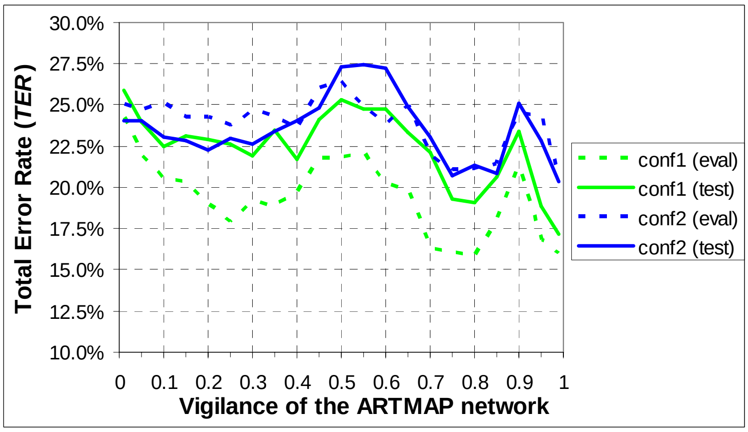


Figure 5.14: FaceART, fourth matching mode, relationship between the verification performance (TER) and vigilance parameter. The size of facial images is 50x50 pixels and the number of eigenfaces is 10 (modified from [161])

Another important point is the compression factor reached by forming categories within the network. In case the 1-NN classifier is applied, all training patterns must be preserved posing the maximum memory requirement. By choosing vigilance lower than 1.0, several training patterns can be described by one node. Figure 5.15 shows the relationship between the vigilance and the average number of category nodes (C-Nodes) in the ARTMAP network. With the vigilance of 0.75 the compression ratio in comparison to the 1-NN classifier is 5.115 (116 vs. 600 references) in the 1st configuration and 5.891 (135 vs. 798 references) in the 2nd configuration of the Lausanne protocol. The number of reference patterns in the 2nd configuration deviates from 800 because the face localization algorithm has a breakdown for two training images.

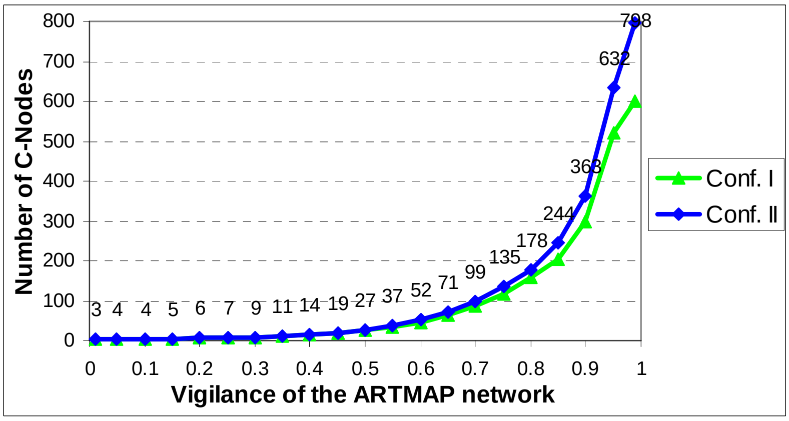


Figure 5.15: FaceART, fourth matching mode, relationship between the average number of C-Nodes and the vigilance parameter (modified from [161])

The best configuration of FaceART with XM2VTS

After comparing the FaceART matching modes with its best parameterizations, it can be stated that verification performance gradually improves from the first to the fourth mode. The corresponding error rates are listed in Table 5.22. The abbreviation “Mode X” stands for the following:

- Mode 1 represents gray-scale facial images of the size of 50x50 pixels directly matched using the city-block distance.
- Mode 2 represents gray-scale facial images of the size of 50x50 pixels matched using the ARTMAP network with the vigilance of 0.85.
- Mode 3 represents gray-scale facial images of the size of 50x50 transformed to eigenfaces coefficients based on 10 most informative eigenfaces and matched using the Mahalanobis distance.
- Mode 4 represents gray-scale facial images of the size of 50x50 transformed to eigenfaces coefficients based on 10 most informative eigenfaces and matched using the ARTMAP network with the vigilance of 0.75.

Table 5.22: Comparison of matching modes within FaceART (modified from [161])

Metric	Configuration 1						Configuration 2					
	Evaluation			Test			Evaluation			Test		
	FAR	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*
Mode 1	13.88%	13.88%	27.76%	11.72%	12.50%	24.22%	13.50%	13.50%	27.00%	11.32%	12.75%	24.07%
Mode 2	11.49%	11.49%	22.99%	10.49%	12.50%	22.99%	11.21%	11.21%	22.41%	10.20%	12.25%	22.45%
Mode 3	10.66%	10.66%	21.33%	9.99%	11.50%	21.49%	10.28%	10.28%	**20.57%**	9.78%	12.00%	21.78%
Mode 4	8.00%	8.00%	**16.01%**	7.81%	11.50%	**19.31%**	10.52%	10.52%	21.05%	9.93%	10.75%	**20.68%**

Generally, it might be concluded that eigenfaces coefficients outperform raw images and the ARTMAP network outperforms the nearest neighbor classifier regardless of the selected distance metric. Figure 5.16 shows ROC curves for Mode 4. ROC curves for other modes as well as FAR/FRR diagrams can be found in Appendix A.1.

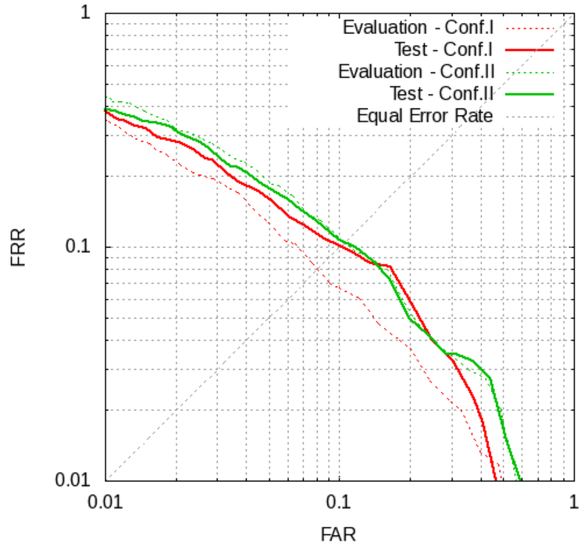


Figure 5.16: FaceART with XM2VTS, ROC curves for Mode4 with logarithmic scale

5.3.4 Experiments with Luxand FaceSDK

Luxand FaceSDK is an alternative piece of face recognition software addressed in this thesis (for more details see Subsection 4.3.5). This is a continually developing commercial product that implements the principles of feature-based face recognition. The current version of the software development kit (SDK) is 5.0. However, the experiments with XM2VTS were carried out in late 2009 with the 2.0 version of SDK.

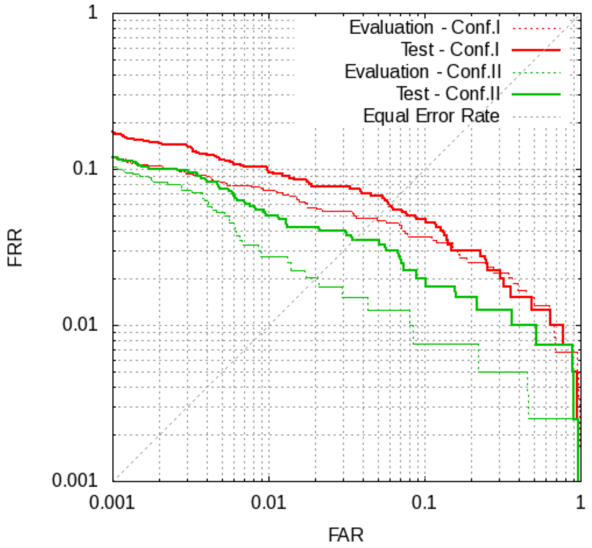


Figure 5.17: Luxand FaceSDK with XM2VTS, ROC curves with logarithmic scale

The library includes all relevant functions for fully automatic face localization and matching of faces. During the evaluation all functions are executed with the standard parameters intentionally avoiding any adjustment to the images from XM2VTS. Figure 5.17 shows the ROC curves resulting from the experiments with 1st and 2nd configurations of the Lausanne protocol. By analyzing

ROC curves one can observe that the FAR approaches 0 when FRR resides between 10% and 18% meaning that Luxand FaceSDK is well suited for security-oriented applications. The distributions of matching scores for genuine users and impostors as well as FAR/FRR diagrams can be found in Appendix A.2. The TER values for evaluation/testing in the 1st and 2nd configurations are 9.67%/12.55% and 3.88%/6.02% respectively.

5.3.5 Comparison to alternative face recognition systems

Considering the quantity of publications, the XM2VTS with Lausanne protocol seems to be the mostly established combination for evaluating verification performance of face recognition systems in an academic domain.

Three public face verification contests were held based on the Lausanne protocol. The first one was a part of the ICPR 2000 conference [165] with four participants:

- Aristotle University of Thessaloniki (AUT),
- Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP),
- University of Sydney (Sydney),
- University of Surrey (UniS).

The second verification contest took place in conjunction with the AVBPA 2003 conference [171] with participants from seven institutions:

- Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP),
- Universidad Politecnica de Valencia (UPV),
- Tübitak Bylten (TB),
- Université Catholique de Louvain (UCL),
- Unknown commercial system (Commercial),
- University of Surrey (UniS),
- Mahanakorn University of Technology (MUT).

However, only IDIAP, UPV, Commercial and UniS provided face verification systems enabled for automatic registration. The last face verification competition was organized within the ICBA 2006 conference [169]. This contest attracted one additional participant:

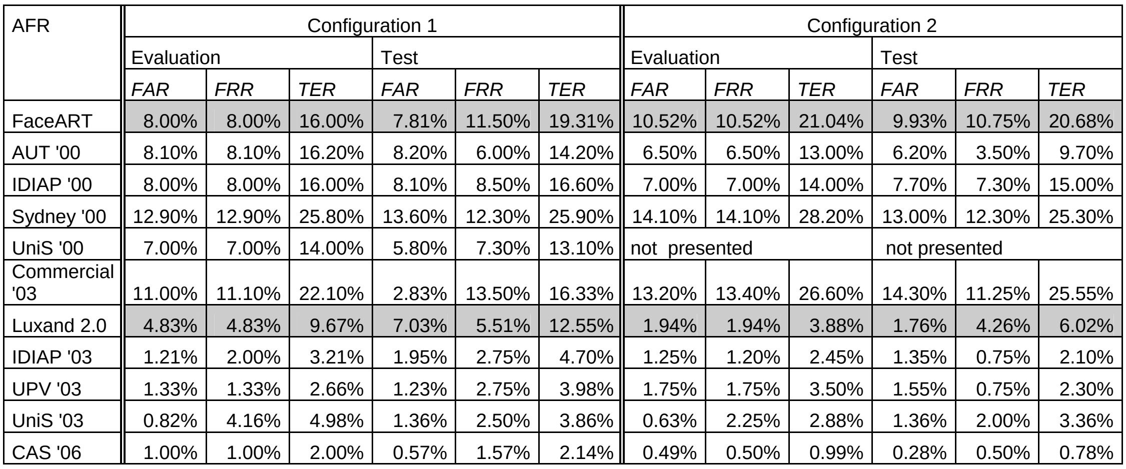
- Chinese Academy of Science (CAS)

The verification performances of algorithms gradually improve from contest to contest. The rivals seem to adjust the algorithms to the data to achieve the almost perfect results that can be seen in Table 5.23. Be advised that only the results with automatic registration are mentioned. Apart from the standard competitors, the table presents the error rates of FaceART and Luxand FaceSDK. FaceART has obviously deficient verification performance in comparison to the best systems e.g. CAS, IDIAP and UniS. The error rates are at the same level with the systems that participated in ICPR 2000. In contrast, Luxand FaceSDK has significantly lower error rates. The system is at the same level with algorithms from AVBPA 2003. It is important to note that neither FaceART nor Luxand FaceSDK is in any sense adjusted to the data in XM2VTS. In fact, the slight improvement of face localization subsystems of both systems could significantly reduce error rates.

5.3.6 Conclusion

FaceART is by far not an optimal piece of face recognition software. It has two principal drawbacks. The first one is inaccurate face localization provided by the external library. In fact, the accurate

Table 5.23: FaceART and Luxand FaceSDK 2.0 versus academic face recognition systems participated in the face verification contests ICPR 2000, AVBPA 2003 and ICB 2006 (modified from [161])



AFR	Configuration 1						Configuration 2					
	Evaluation			Test			Evaluation			Test		
	FAR	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*	*FAR*	*FRR*	*TER*
FaceART	8.00%	8.00%	16.00%	7.81%	11.50%	19.31%	10.52%	10.52%	21.04%	9.93%	10.75%	20.68%
AUT '00	8.10%	8.10%	16.20%	8.20%	6.00%	14.20%	6.50%	6.50%	13.00%	6.20%	3.50%	9.70%
IDIAP '00	8.00%	8.00%	16.00%	8.10%	8.50%	16.60%	7.00%	7.00%	14.00%	7.70%	7.30%	15.00%
Sydney '00	12.90%	12.90%	25.80%	13.60%	12.30%	25.90%	14.10%	14.10%	28.20%	13.00%	12.30%	25.30%
UniS '00	7.00%	7.00%	14.00%	5.80%	7.30%	13.10%	not presented			not presented		
Commercial '03	11.00%	11.10%	22.10%	2.83%	13.50%	16.33%	13.20%	13.40%	26.60%	14.30%	11.25%	25.55%
Luxand 2.0	4.83%	4.83%	9.67%	7.03%	5.51%	12.55%	1.94%	1.94%	3.88%	1.76%	4.26%	6.02%
IDIAP '03	1.21%	2.00%	3.21%	1.95%	2.75%	4.70%	1.25%	1.20%	2.45%	1.35%	0.75%	2.10%
UPV '03	1.33%	1.33%	2.66%	1.23%	2.75%	3.98%	1.75%	1.75%	3.50%	1.55%	0.75%	2.30%
UniS '03	0.82%	4.16%	4.98%	1.36%	2.50%	3.86%	0.63%	2.25%	2.88%	1.36%	2.00%	3.36%
CAS '06	1.00%	1.00%	2.00%	0.57%	1.57%	2.14%	0.49%	0.50%	0.99%	0.28%	0.50%	0.78%

localization is an essential issue for appearance-based methods since the images are matched literally pixel-to-pixel. FaceART includes no mechanism to validate the localized face region as well as to examine the accuracy of the localization results. Hence, some localized areas do not include a face or include it only partially. The occurrence of such localization errors among training images leads to the drastic degradation of recognition performance. The second drawback is a lack of analyzing discrimination power of single features. The PCA transformation does not guarantee that the resulting features possess high discrimination power. Thus, employing coefficients of first eigenfaces seems to provide very limited improvement. The recognition performance can be eventually gained by applying linear discriminant analysis (LDA) on raw images or even on eigenfaces [12]. The experiments have shown that the ARTMAP network applied instead of the matching algorithm and reference storage outperforms the nearest neighbor classifier and leads to more compact preservation of biometric templates. However, the best results are achieved with the marginal value of the vigilance parameter. With such parameterization the ARTMAP network operates similarly to the nearest neighbor classifier with an alternative similarity metric (which is the slightly modified scalar product). As a matter of fact, for appearance-based approaches to operate well, many reference images for each user are required. These images have to cover all possible face appearances. With the Lausanne protocol providing 3 and 4 images for training per user in the 1st and 2nd configuration respectively, the high recognition performance cannot be expected with FaceART.

Luxand FaceSDK has significantly better recognition performance with the Lausanne protocol compared to that of FaceART. Nonetheless, the feature extraction algorithm is often incapable of precisely localizing fiducial points on faces rotated too much in the second head shot. By requiring that the localization of fiducial points takes place exclusively for reliably detected faces and ignoring images with the excessively rotated heads, recognition performance can be drastically improved, reducing error rates to those shown by competing algorithms provided by other academic institutions.

5.4 Face recognition performance in a car

This section is comprised of the following subsections:

1. Opel databases (OpelDB2, OpelDB3, OpelDB2+3)
2. Performance measures
3. Experiments with FaceART
4. Experiments with Luxand FaceSDK
5. Conclusion

5.4.1 Opel databases

A recognition performance similar to that resulting from the experiments with XM2VTS cannot be expected in a real car. In fact, XM2VTS provides images collected in a photo studio with uniform illumination conditions and a monotone constant background. In contrast, the illumination conditions in a car can drastically vary from completely dark to very bright and the intensive side lighting can provide light artifacts on faces. The background is constant except for the window parts, but the light penetrating into the car can drastically change the appearance of the interior. Hence, there is a need for experimental data collected in automotive environment to reveal a realistic recognition performance of the addressed face recognition systems (FaceART, Luxand FaceSDK).

By creating the experimental database, the following aspects are in focus:

- reproduce realistic environmental conditions,
- address the proposed operational scenarios (single driver identification, permanent driver authentication),
- address aging of the reference data,
- carry out not only closed-set but also open-set identification tests.

In order to satisfy the first aspect, an experimental system is implemented in a conventional vehicle Opel Vectra B (for details see Subsections 4.3.4 and 4.3.5. Data are captured at different times of day, at different locations, with different illumination conditions which sometimes lead to overexposed parts of images. Thanks to the camera set up and scrupulously adjusted active illumination sources, frontal faces can be easily detected even at night and under intensive side lighting. Figure 5.18 shows some examples with difficult illumination conditions.

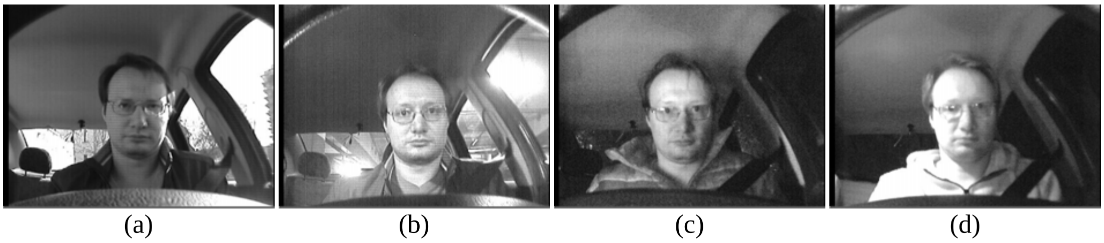

(a) (b) (c) (d)

Figure 5.18: OpelDB3 samples: (a) intensive side lighting leading to a light artifact on a face, (b) underground parking with deficient illumination and outdoor lamps facing the camera, (c) at night without outdoor lighting and (d) at night with conventional interior lamp turned on

Two databases OpelDB2, OpelDB3 are created to address the operational scenarios proposed in Subsections 4.3.3 and 4.3.3. OpelDB2 is constructed to focus on short-time aging of the reference data not necessarily engaging many users. By constructing OpelDB3, the focus is on testing discriminative performance of face recognition systems which demands a high number of users. All users in these databases are considered as clients so that the evaluation is provided first in a closed-set identification mode and second in verification mode. Combining both databases in OpelDB2+3 enables carrying out an open-set identification test. Here, five common users are considered as clients and the remaining users as impostors.

5.4.2 OpelDB2

The OpelDB2 database was created from July to September 2010. By gathering data for the database, twelve persons are engaged, but only six complete all acquisition sessions. The goal is to simulate the simultaneous car usage by family members so that every registered driver accesses the car several times, always introducing a new face appearance. There is no true support from active NIR illumination. The lamp is not properly adjusted so that face shots become overexposed at night. Hence, all recordings are collected during daytime with sufficient amount of outdoor lighting and an NIR lamp switched off.

The recordings are made in five sessions at the interval of approximately one week between two subsequent sessions. Every session contains three videos:

- Reference video (R) with head rotations. The recording begins with a person looking at the camera. Then he/she lifts the head and looks at the street, then rotates the head left and right, then up and down, the tilts the head left and right, and finally once more looks at the street and at the camera (see Figure 5.19 for samples).
- First test video (T_1) of the car entrance with fastening seat belt and short look to the camera (see Figure 5.20 for the final video frames).
- Second test video (T_2) with the unconstrained car entrance and an approximately two minute drive. The drive includes the following maneuvers: leave a parking space and immediately turn right, follow the street for 50 meters and turn right, follow the street for 150 meters and make a broken U-turn using the by-street on the left, drive 150 meters, take a left turn, drive another 50 meters and back into a parking space which is on the left side of the car. Some example frames can be found in Figure 4.27.

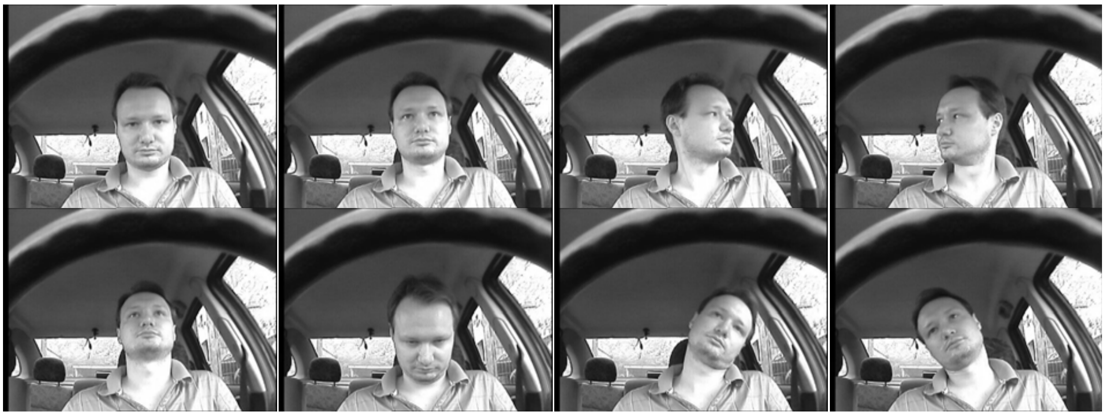


Figure 5.19: OpelDB2 head rotation samples from a reference video: look at the camera, at the street, left, right, up, down, left tilt, right tilt

The reference video is used for enrollment. The driver registration algorithm can learn different face appearances and create a statistical model. The first and second test videos are created to address two operating scenarios: single driver identification (SDI) and permanent driver verification (PDV) respectively. As a reminder, SDI is required to activate the memory function (comfort) and individual driver assistance systems (safety), and PDV is required to assure that the car is steered by a registered driver (anti-theft protection).

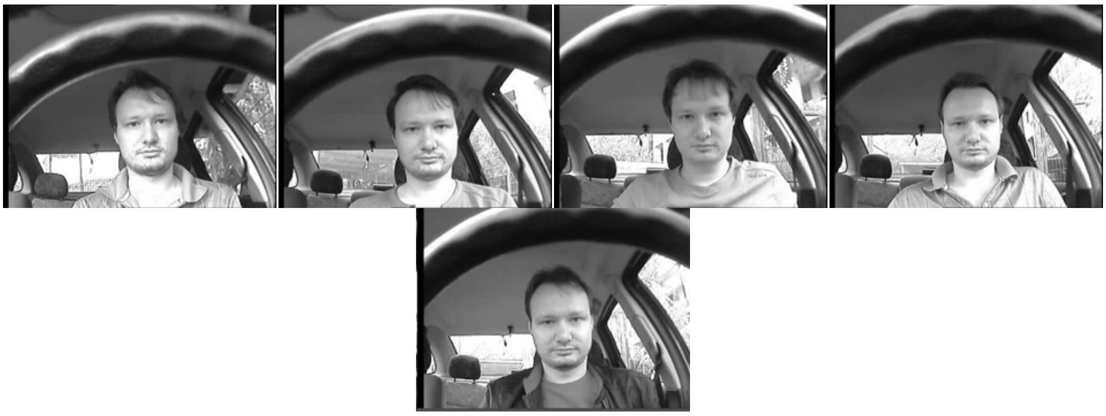


Figure 5.20: OpelDB2 samples of frontal faces from five sessions

Collecting data at intervals longer than one week helps to address short-time aging of the reference samples. It is claimed that the longer the time interval between the point in time when the reference data is gathered and the point in time when the authentication takes place, the worse the recognition performance will be. Three evaluation modes are proposed to investigate this assumption:

- Standard mode simulating the typical situation where reference data undergo aging. Enrollment is based on the reference data from the first session and authentication is based on the test data from the remaining four sessions. When matched with each next session, the reference data becomes older.
- Adaptive mode simulating the dynamic template preservation. The reference data is permanently updated based on authentication results. Enrollment is based on the reference data from the sessions which forego the session of the test data. By operating in the adaptive mode, the system obviates the re-enrollment because the reference and test samples become a shorter time-gap.
- No-aging mode simulating the non-realistic perfect situation. Reference and test data are taken from the same session. Evaluating a system in this mode helps to reveal the upper bounds of recognition performance.

Table 5.24 illustrates the distribution of reference and test videos in accordance with the evaluation modes and operating scenarios.

5.4.3 OpelDB3

The OpelDB3 database was created from April to May 2013 to extend the number of drivers. There are 51 persons engaged whereby five of them have participated in creating OpelDB2. The data is gathered during only one session and includes a reference and a test video. Reference videos are similar to those recorded for OpelDB2 and test videos are similar to the first test videos from OpelDB2. Hence, OpelDB3 supports only the closed-set SDI test.

The camera viewing angle is slightly adjusted to avoid situations in which parts of a face are covered by a steering wheel. Active NIR lighting has been adjusted. The LED ring around the

Table 5.24: OpelDB2, evaluation modes, R_n - reference video from the session n, T_{1-n} and T_{2-n} - first and second test videos from the session n respectively

	SDI test	PDV test
Standard mode	R_1 vs. T_{1_2} R_1 vs. T_{1_3} R_1 vs. T_{1_4} R_1 vs. T_{1_5}	R_1 vs. T_{2_2} R_1 vs. T_{2_3} R_1 vs. T_{2_4} R_1 vs. T_{2_5}
Adaptive mode	R_1 vs. T_{1_2} R_1+ R_2 vs. T_{1_3} R_1 + R_2 + R_3 vs. T_{1_4} R_1 + R_2 + R_3 + R_4 vs. T_{1_5}	R_1 vs. T_{2_2} R_1 + R_2 vs. T_{2_3} R_1 + R_2 + R_3 vs. T_{2_4} R_1 + R_2 + R_3 + R_4 vs. T_{2_5}
No-aging mode	R_1 vs. T_{1_1} R_2 vs. T_{1_2} R_3 vs. T_{1_3} R_4 vs. T_{1_4} R_5 vs. T_{1_5}	R_1 vs. T_{2_1} R_2 vs. T_{2_2} R_3 vs. T_{2_3} R_4 vs. T_{2_4} R_5 vs. T_{2_5}

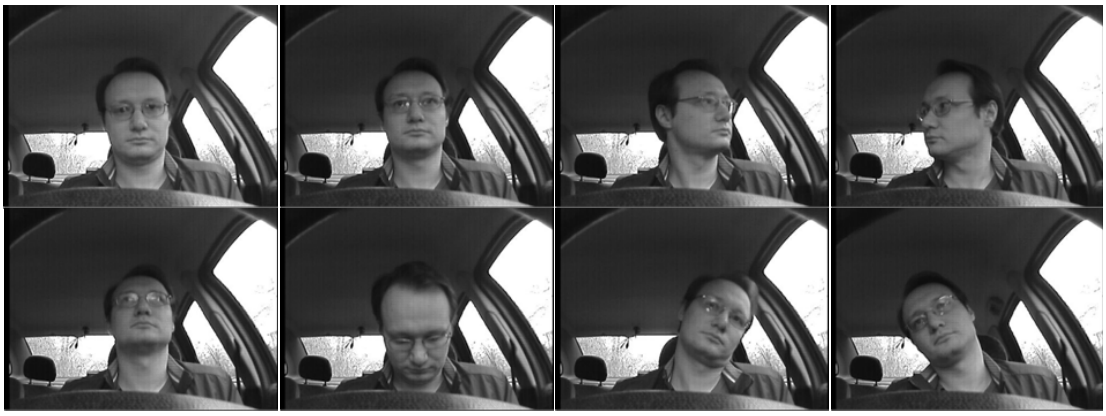


Figure 5.21: OpelDB3 head rotation samples from a reference video: look at the camera, at the street, left, right, up, down, left tilt, right tilt

camera lens is permanently switched on, but covered by a diffuser. On the one hand, this enables night acquisition without overexposing faces. On the other hand, its lighting compensates intensive sunlight from the rear. Another adjustment regarding the interior illumination is the 875 nm NIR lamp in the ceiling next to the windscreen at the place of the conventional interior lamp. This lamp helps to compensate sunlight from the driver window making faces more homogeneous. Figure 5.21 shows head rotation samples from one reference video in OpelDB3.

5.4.4 OpelDB2+3

In OpelDB2+3 two aforementioned driver databases are combined to carry out the open-set SDI test. Five common users from OpelDB2 and OpelDB3 are considered as clients. Each of them was acquired in six sessions (5 sessions from OpelDB2 + 1 session from OpelDB3). The remaining 46 users are considered to be impostors. While collecting data for the OpelDB2 several persons were acquired in one session. Three persons, which were acquired in 2010 but included neither in OpelDB2 nor in OpelDB3, are added to OpelDB2+3 as impostors. In the case of impostors, only the first test video is relevant because they are prevented from driving a car. Impostor videos are further referred to as $T_{1\_impostor}$. Table 5.25 introduces how client and impostor matching trials are distributed in accordance with the evaluation modes.

Table 5.25: OpelDB2+3, evaluation modes, R_n - reference video from the session n, T_{1-n} - genuine test videos from the session n, $T_{1-impostor}$ - impostor test videos

	Genuine trials	Impostor trials
Standard mode	R_1 vs. T_{1_2} R_1 vs. T_{1_3} R_1 vs. T_{1_4} R_1 vs. T_{1_5} R_1 vs. T_{1_6}	R_1 vs. $T_{1\_impostor}$
Adaptive mode	R_1 vs. T_{1_2} $R_1 + R_2$ vs. T_{1_3} $R_1 + R_2 + R_3$ vs. T_{1_4} $R_1 + R_2 + R_3 + R_4$ vs. T_{1_5} $R_1 + R_2 + R_3 + R_4 + R_5$ vs. T_{1_6}	R_1 vs. $T_{1\_impostor}$ $R_1 + R_2$ vs. $T_{1\_impostor}$ $R_1 + R_2 + R_3$ vs. $T_{1\_impostor}$ $R_1 + R_2 + R_3 + R_4$ vs. $T_{1\_impostor}$ $R_1 + R_2 + R_3 + R_4 + R_5$ vs. $T_{1\_impostor}$
No-aging mode	R_1 vs. T_{1_1} R_2 vs. T_{1_2} R_3 vs. T_{1_3} R_4 vs. T_{1_4} R_5 vs. T_{1_5} R_6 vs. T_{1_6}	R_1 vs. $T_{1\_impostor}$ R_2 vs. $T_{1\_impostor}$ R_3 vs. $T_{1\_impostor}$ R_4 vs. $T_{1\_impostor}$ R_5 vs. $T_{1\_impostor}$ R_6 vs. $T_{1\_impostor}$

5.4.5 Performance metrics

With closed-set identification addressed by OpelDB2 and OpelDB3, all test persons are registered. A result of each identification trial is a rank list of clients and the true identity is guaranteed to be in the list. The most common metric of recognition performance in the identification mode is identification accuracy expressing the probability to meet the correct driver at the first position in the rank list without considering matching scores (see Subsection 2.3.4). However, if the numbers of test samples for different classes are significantly different or the recognition performance is not symmetrical regarding the classes, identification accuracy can be strongly biased (for explanation see Subsection 2.3.4). The Kappa statistic calculated based on the confusion matrix helps to overcome the problem of biased estimation (for details see Subsection 2.3.4). Another typical indicator of the recognition performance is the cumulative matching curve (CMC) which represents the probabilities to meet the true identity among N first entities of the rank list (for details see Subsection 2.3.4). The CMC is a very useful performance indicator for human-aided identification systems. For automatic systems, however, only the first position in the rank list is important. In experiments, CMC curves are created to compare the recognition performances of FaceART and Luxand FaceSDK in different evaluation modes.

In contrast to identification, a result of each verification trial is a matching score. This is why, in the permanent driver verification scenario, the client scores and impostor scores are the issues determining the recognition performance. Since closed-set identification does not imply real impostors, the client matching scores are intra-class scores resulting from matching reference and test samples of the same user, and impostor matching scores are inter-class scores resulting from matching of reference samples of one user with test samples of another user. The distributions of client and impostor matching scores are used to create FAR/FRR diagrams as well as to determine the threshold where EER is reached. FAR and FRR are estimated so as to reside within confidence intervals at the confidence level of 95% ($\alpha = 0.05$). The maximum confidence interval of both is adopted for EER (for details see Subsection 2.3.3).

In open-set identification, which is an issue of OpelDB2+3, only some persons are registered, but the vast majority is not. The unregistered persons can be considered true impostors. In impostor trials, the rank list cannot include true identity. This problem can be overcome by introducing one additional class “non-clients” or by involving matching scores in the rank list. The first solution requires reference data for the “non-clients” class. These reference data, however, can never be representative because the set of possible impostors is almost unlimited. By considering matching scores, a person is identified as a client at the first position in the rank list only if the corresponding matching score is less than the threshold (score is a distance metric), or higher than the threshold (score is a similarity metric). Consequently, there are two types of false identifications: the correct identity is not on the first position in the rank list or the correct identity is on the first position but the matching score is too high (for distance) or too low (for similarity). So, the experiments with OpelDB2+3 include both, calculation of EER with the corresponding threshold τ_{eer} and further identification with regard to the matching score. Additionally, the Kappa statistic is calculated to address the eventual bias of the identification accuracy.

5.4.6 Experiments with FaceART

Before starting to describe the results of the experiments, let us come back to Subsection 4.3.3 and take another look at abbreviations. The SDI is the operational scenario with semi-automatic authentication and the PDV is the operational scenario with automatic authentication. The “-sec.” or “-conv.” addition to the name of the operational scenario refers to manual or semi-automatic enrollment respectively. In manual mode, a human supervisor scrupulously selects frames for enrollment. It can be one frame (1frame) or several frames (Nframes). In semi-automatic mode, a computer extracts all frames containing faces (multi-view) and a human supervisor quickly removes falsely detected non-faces. Automatic enrollment is not addressed in the experiments.

FaceART, as a representative of appearance-based face recognition approaches, requires collecting different face appearances during enrollment to achieve high recognition performance. Hence, enrollment from one frame does not fit the concept of FaceART. High identification rates cannot be expected with one reference frame. In contrast, learning FaceART from several frames with almost frontal faces can lead to adequate identification accuracy. Furthermore, FaceART is parameterized to operate in the third mode (modes are described in Subsection 4.3.7, providing eigenfaces decomposition and utilizing the 1-NN classifier. Eigenfaces are created on the fly from reference samples. This is why the number of eigenfaces depends on the number of reference samples and varies from one evaluation mode to another. Based on the results with XM2VTS, the city-block distance is claimed to be the optimal metric and chosen for all FaceART experiments.

In order to avoid identification errors caused by inaccurate face localization, faces found outside the special region are ignored. Figure 5.22 shows these special regions for both databases: OpelDB2 and OpelDB3. At the same time, the lower and upper margins for the face size are defined so as to ignore too small and too big face candidates. The size and coordinates of potential face regions are derived from analyzing distributions of these values in the set of correctly localized faces.

FaceART with OpelDB2

Table 5.26 presents the results for closed-set SDI with OpelDB2. The identification results are evaluated in a frame-based manner. With the enrollment from one frontal face, the identification accuracies yield 42.25%, 59.46% and 84.86% in standard, adaptive and no-aging evaluation modes respectively. The significant difference between identification accuracy and a kappa statistic indicates the asymmetrical recognition performance regarding different users. With the enrollment from several almost frontal faces, the identification accuracy exceeds 90% yielding 90.04%, 94.88% and 99.27% in standard, adaptive and no-aging modes respectively. With the enrollment from

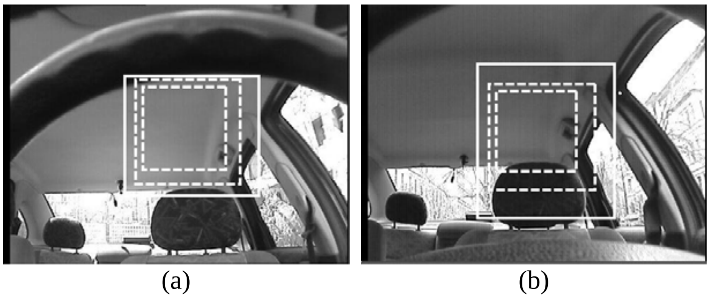


Figure 5.22: FaceART, white solid rectangles mark valid regions for a face in: (a) OpelDB2, (b) OpelDB3; dashed squares visualize the minimal and maximal face size

"face multi-views", the accuracy is higher than that with one frame but lower than that with several frames. The effect of aging reference data becomes apparent by comparing evaluation modes. In standard mode the recognition rates are significantly lower that those in adaptive mode and recognition rates in adaptive mode fall far behind those in no-aging mode. An accuracy of approximately 95% in adaptive mode proves that even a trivial face recognition system such FaceART can be successfully applied for SDI when the enrollment and identification are controlled, and the reference data become permanently updated.

Table 5.26: FaceART, closed-set single driver identification with OpelDB2 (6 clients)

Operational scenario	Param.	Evaluation mode	Identification accuracy	Kappa statistic
SDI-sec., 1frame	1NN, 5 eig	standard	297/703 (42.25%)	32.10%
		adaptive	418/703 (59.46%)	49.49%
		no-aging	695/819 (84.86%)	80.42%
SDI-sec., Nframes	1NN, 50 eig	standard	633/703 (90.04%)	86.92%
		adaptive	667/703 (94.88%)	93.30%
		no-aging	**813/819 (99.27%)**	**99.02%**
SDI-conv., multi-view	1NN, 50 eig	standard	424/703 (60.31%)	50.21%
		adaptive	not tested	not tested
		no-aging	783/819 (95.60%)	94.15%

Table 5.27 presents the results for PDV with OpelDB2. Apart from EER, the table contains the threshold (τ_{eer}) where EER is reached and the confidence interval of EER at the confidence level of 95%. The EER values over 20% are too high to consider FaceART applicable for PDV. These error rates imply that at least every fifth verification trial is erroneous. The driver will be warned that he/she cannot be verified which can be very annoying. Unexpectedly, with the enrollment from one frame, the EER value in no-aging mode is lower than that with the enrollment from several frames. The reason for this is probably the number of eigenfaces which means that 50 eigenfaces are superfluous. This is also the reason for different thresholds with 1frame and with Nframes/multi-view.

FaceART with OpelDB3

Table 5.28 presents the results for SDI with OpelDB3. Here, the difference between identification accuracy and a kappa statistic is very low meaning that recognition performance is almost symmetrical for all users.

The first issue investigated in the experiments is a proper number of eigenfaces. The table shows

Table 5.27: FaceART, permanent driver verification with OpelDB2 (6 clients)

Operational scenario	Param.	Evaluation mode	*EER*	τ_{eer}	Confidence interval
PDV-sec., 1frame	1NN, 5 eig	standard	31.41%	13.87551	±0.97%
		adaptive	28.64%	11.13637	±0.94%
		no-aging	**20.38%**	**12.079025**	±0.81%
PDV-sec., Nframes	1NN, 50 eig	standard	27.54%	37.065817	±0.93%
		adaptive	24.68%	36.184981	±0.90%
		no-aging	24.74%	35.838516	±0.86%
PDV-conv., multi-view	1NN, 50 eig	standard	30.07%	40.312078	±0.96%
		adaptive	not tested	not tested	not tested
		no-aging	27.57%	36.175244	±0.89%

that 50 eigenfaces outperform 30 eigenfaces and the latter outperform 15 eigenfaces in all cases. The application of eigenfaces for multi-view enrollment encounters the memory expansion problem. Due to the very high number of reference samples (> 25000), the database decomposition function breaks down. Hence, multi-view enrollment is tested with FaceART operating in the first mode (see Subsection 4.3.7) in which raw gray-scale values are classified by k-NN.

The second issue is which classifier is superior. With the enrollment from one frame, only the 1-NN classifier is feasible. With the enrollment from several frames, the performance indices indicate that 1-NN outperforms 5-NN, the latter outperforms 9-NN and 9-NN outperforms ARTMAP, independently from the number of eigenfeces. The best accuracy (92.11%) is achieved with multi-view enrollment and 1-NN applied for classification of raw gray-scale values.

Table 5.28: FaceART, closed-set single driver identification with OpelDB3 (51 clients)

Operational scenario	Parameters	Identification accuracy	Kappa statistic
SDI-sec., 1frame	1NN, 5 eig	2605/5502 (47.35%)	46.15%
	1NN, 15 eig	3727/5502 (67.74%)	66.95%
	1NN, 30 eig	4027/5502 (73.19%)	72.54%
	1NN, 50 eig	4165/5502 (75.70%)	75.12%
SDI-sec., 10frames	1NN, 5 eig	2711/5502 (49.27%)	48.05%
	1NN, 15 eig	3872/5502 (70.37%)	69.64%
	1NN, 30 eig	4336/5502 (78.81%)	78.26%
	1NN, 50 eig	**4522/5502 (82.19%)**	**81.74%**
	5NN, 5 eig	2709/5502 (49.24%)	48.02%
	5NN, 15 eig	3781/5502 (68.72%)	67.94%
	5NN, 30 eig	4169/5502 (75.77%)	75.15%
	5NN, 50 eig	4425/5502 (80.43%)	79.93%
	9NN, 5 eig	2600/5502 (47.26%)	45.99%
	9NN, 15 eig	3705/5502 (67.34%)	66.53%
	9NN, 30 eig	4060/5502 (73.79%)	73.12%
	9NN, 50 eig	4342/5502 (78.92%)	78.38%
	ARTMAP, 5 eig	2556/5502 (46.46%)	45.18%
	ARTMAP, 15 eig	3816/5502 (69.36%)	68.56%
	ARTMAP, 30 eig	4051/5502 (73.63%)	72.93%
	ARTMAP, 50 eig	4269/5502 (77.59%)	77.02%
SDI-conv., multi-view	**1NN, raw**	**5068/5502 (92.11%)**	**91.90%**
	5NN, raw	5026/5502 (91.35%)	91.12%
	9NN, raw	4988/5502 (90.66%)	90.41%

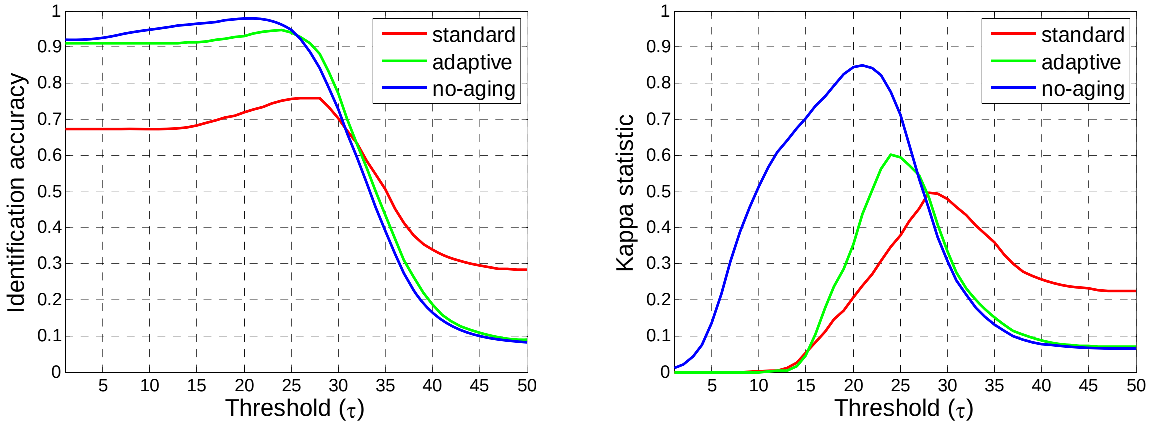
FaceART with OpelDB2+3

In the experiments with OpelDB2+3, the open-set identification with 5 clients and 49 impostors is addressed. The 1-NN classifier is chosen for tests because it outperforms 5-NN, 9-NN and ARTMAP classifiers in previous experiments and also provides a matching score expressed by the city-block distance between a test sample and a reference sample. As mentioned in Subsection 5.4.2, involving matching scores is essential for open-set identification. Aiming at analyzing distributions of intra-class, inter-class and impostor matching scores, tests are first carried out in verification mode. An EER value indicates how successful the identification could be. Table 5.29 presents the results of the verification test. The EER values of 5% and 10.8% in no-aging and adaptive modes respectively indicate potentially high identification accuracy. In standard mode, EER of 17.69% indicates rather limited recognition performance.

Table 5.29: FaceART, verification test with OpelDB2+3 (5 clients and 49 impostors)

Operational scenario	Param.	Evaluation mode	*EER*	τ_{eer}	Confidence interval
SDI-sec. Nframes	1NN, eig50	standard	17.69%	33.22284	±1.55%
		adaptive	10.80%	31.56233	±1.26%
		no-aging	5.00%	28.30156	±0.85%

The τ_{eer} is not an optimal threshold for identification. Reducing the threshold can help to reject a high number of impostor trials and only a low number of clients trials and, therefore, to get to a more secure system. However, too low thresholds make the system useless rejecting all users just as with too high thresholds all impostor trials lead to successful identification. Figure 5.23 illustrates the relationship between identification accuracy/kappa statistic and the threshold, revealing the values 28, 24 and 21 as optimal thresholds for standard, adaptive and no-aging evaluation modes respectively. The accuracy values at marginal thresholds actually represent the balance between the number of client and impostor trials.



(a) (b)

Figure 5.23: FaceART, relationship between the threshold for the matching score and (a) identification accuracy and (b) kappa statistic

Table 5.30 presents the results of open-set SDI with OpelDB2+3. Here, the values of kappa statistic strongly deviate from accuracy values due to the fact that the number of impostor trials is approximately 10 times higher than the number of client trials. Consequently, the ability to correctly reject impostors has higher influence on identification accuracy. Nonetheless, the values of kappa statistic here represent the recognition performance in a more appropriate way. A very good agreement with kappa of approximately 85% can be stated in no-aging mode and a good agreement with kappa of approximately 60% in adaptive mode. In standard mode, the kappa of approximately 50% indicates a moderate agreement.

Table 5.30: FaceART, open-set single driver identification with OpelDB2+3 (5 clients and 49 impostors)

Operational scenario	Parameters	Evaluation mode	Identification accuracy	Kappa statistic	Optimal threshold
SDI-sec., Nframes	1NN, 50 eig	standard	5419/7143 (75.86%)	49.53%	28
		adaptive	24935/26347 (94.64%)	60.10%	24
		no-aging	30674/31341 (97.87%)	84.91%	21

5.4.7 Experiments with Luxand FaceSDK

Luxand FaceSDK can be considered a representative of feature-based face recognition approaches whereby the individual face model is constructed based on fiducial points localized on a face (for details see Subsection 4.3.6). One sharp frame with a frontal face is enough for creating a robust face model. Since manual extraction of a key frame from the reference video (1frame) works well, semi-automatic enrollment (Nframes) becomes superfluous. In a similar way to which it is done for FaceART, faces found outside the special region are ignored to reduce errors caused by inaccurate face localization. Figure 5.24 shows these regions along with squares representing lower and upper margins for the face size in both databases: OpelDB2 and OpelDB3.

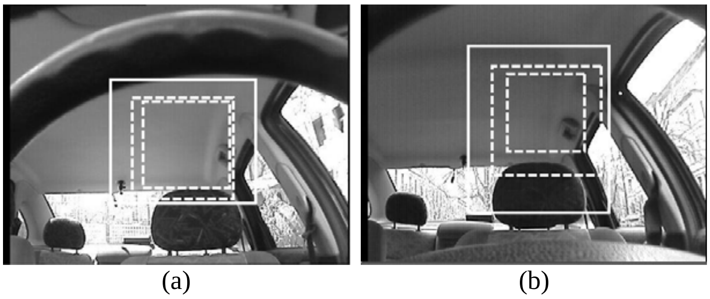


Figure 5.24: Luxand FaceSDK, white solid rectangles mark valid regions for a face in: (a) OpelDB2, (b) OpelDB3; dashed rectangles visualize the minimal and maximal face size

Luxand FaceSDK with OpelDB2

The experiments with OpelDB2 were carried out using version 4.0 of Luxand FaceSDK. Table 5.31 present the result for SDI. In all three evaluation modes, the identification accuracy exceeds 99%. The values of kappa statistic do not deviate much from accuracy values indicating very high level of agreement.

In the PDV test (see Table 5.32), the EER values are similarly high as they are with FaceART varying between 26.16% in the optimal case and 27.85% in the worst case. Unexpectedly, updating reference data does not improve verification performance. In contrast to FaceART, the matching score in FaceSDK is a degree of similarity. In client trials, the scores are expected to approach 1. Scores lower than 0.9 indicate significant differences in faces. The τ_{eer} value of 0.6 or even 0.7

Table 5.31: Luxand FaceSDK, closed-set single driver identification with OpelDB2 (6 clients)

Operational scenario	Operating mode	Identification accuracy	Kappa statistic
SDI-sec., 1frame	standard	701/707 (99.15%)	98.93%
	adaptive	706/707 (99.86%)	99.82%
	no-aging	775/778 (99.61%)	99.51%

reveals that many client images contain rotated faces. Hence, the PDV scenario requires additional image processing for estimating the quality of localized faces and filtering of non-frontal ones.

Table 5.32: Luxand FaceSDK, permanent driver verification with OpelDB2 (6 clients)

Operational scenario	Operating mode	EER	τ_{eer}	Confidence interval
PDV-sec. 1frame	standard	26.16%	0.593221	±1.21%
	adaptive	26.92%	0.702811	±1.22%
	no-aging	27.85%	0.607654	±1.09%

Luxand FaceSDK with OpelDB3

In the experiments with OpelDB3 and OpelDB2+3, the version 5.0 of Luxand FaceSDK is utilized. The closed-set SDI test includes 51 clients. The enrollment is provided from one key frame with a frontal face. The resulting identification accuracy yields 91.05% (5664 successes in 6221 trials) with a kappa statistic of 90.83%.

Luxand FaceSDK with OpelDB2+3

Since the open-set identification supported by OpelDB2+3 with 5 clients and 49 impostors requires considering matching scores, the first test is provided in verification mode aiming at analyzing intra-class, inter-class and impostor distributions. The EER values together with thresholds and confidence intervals are presented in Table 5.33. ERRs of 7.64%, 6.46% and 4.93% in standard, adaptive and no-aging modes respectively indicate very high potential for identification.

Table 5.33: Luxand FaceSDK, verification test with OpelDB2+3 (5 clients and 49 impostors)

Operational scenario	Operating mode	EER	τ_{eer}	Confidence interval
SDI-sec., 1frame	standard	7.64%	0.216301	±1.01%
	adaptive	6.46%	0.3946	±0.94%
	no-aging	4.93%	0.418833	±0.78%

The optimal threshold for identification can be derived from diagrams in Figure 5.25 introducing the relationship between identification accuracy/kappa statistic and the threshold. In contrast to FaceART, choosing too low thresholds lets the system accept all trials and choosing too high thresholds lets the system reject all trials. Hence, increasing a threshold makes a system more secure by rejecting significantly more impostor trials and only a low number of client trials. The thresholds of 0.46, 0.75 and 0.91 are optimal for standard, adaptive and no-aging evaluation modes respectively.

The results of open-set SDI with OpelDB2+3 are presented in Table 5.34. Identification accuracies are very high yielding 92.04%, 97.53% and 98.32% in standard, adaptive and no-aging modes respectively. However, due to a significant difference between kappa statistic and accuracy, kappa

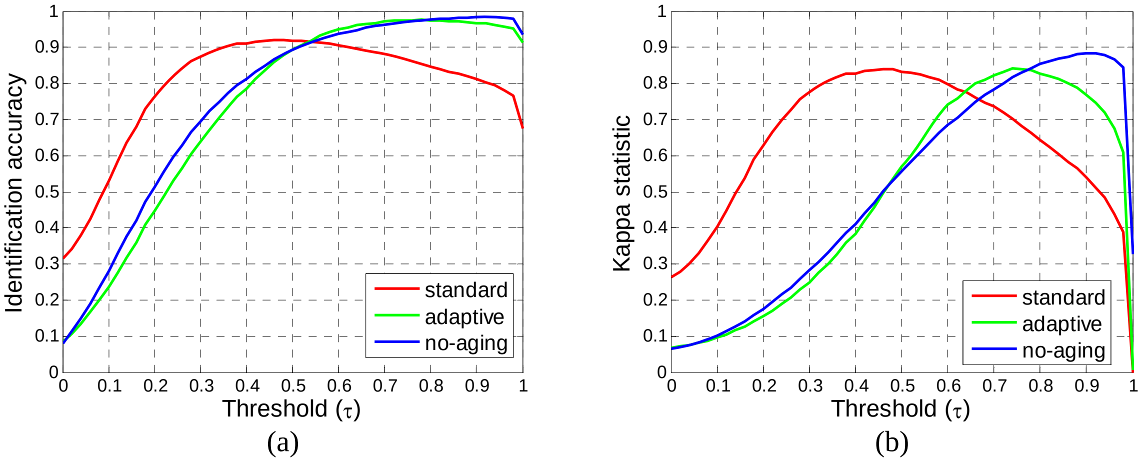


Figure 5.25: Luxand FaceSDK, relationship between the threshold for the matching score and (a) identification accuracy and (b) kappa statistic

values are preferable for designating recognition performance. In all three evaluation modes a very good agreement can be stated with kappa values of 83.94%, 84.02% and 88.45%.

Table 5.34: Luxand FaceSDK, open-set single driver identification with OpelDB2+3 (5 clients and 49 impostors)

Operational scenario	Operating mode	Identification accuracy	Kappa statistic	Optimal threshold
SDI-sec., 1frame	standard	7471/8117 (92.04%)	83.94%	0.46
	adaptive	29289/30032 (97.53%)	84.02%	0.75
	no-aging	35199/35799 (98.32%)	88.45%	0.91

5.4.8 Conclusion

Generally, the recognition performances of both algorithms confirm my theoretical considerations about the importance of the control over enrollment and authentication scenarios in biometric systems. The high degree of control (e.g. guarantee frontal faces in reference and test data) leads to relatively low error rates. The recognition performance of Luxand FaceSDK is significantly higher compared to that of FaceART even considering the fact that the former is learned from only one frame. In contrast, the presence of several slightly different frontal face images is essential for FaceART to demonstrate high identification accuracy. Distributions of matching scores, FAR/FRR diagrams, ROC and CMC curves for both systems FaceART and Luxand FaceSDK can be found in Appendix B, whereby for FaceART, only the best parameterization is considered.

In the SDI test, the results with OpelDB2 can be considered optimistic. Identification accuracy over 99% in all evaluation modes with Luxand FaceSDK is a perfect result which unfortunately cannot be expected in real life. The results of experiments with OpelDB2+3 can be considered more realistic. In case FaceART is applied, the accuracy/kappa is significantly better with up-to-date reference data (adaptive/no-aging versus standard mode). Hence, for successful operating, FaceART requires a mechanism for a permanent updating of biometric reference data. In contrast, Luxand FaceSDK only slightly benefits from updating reference data, but the accuracy/kappa is very convincing even in standard evaluation mode. Generally, the experiments show that a practical integration of face recognition into a car is feasible for at least SDI with the enrollment requiring human-aided quality control of the reference data and the identification requiring a driver takes a glimpse at the camera.

The PDV test was carried out only with OpelDB2 and even in this fairly optimistic test, the results are disappointing. For automatic unconstrained verification, the error rates drastically increase with both: FaceART and Luxand FaceSDK. There are two factors leading to high error rates. First, the moving car gives rise to permanently changing lighting and dynamic shadows and, second, the rotated heads are often inaccurately localized. The improvement of recognition performance can be achieved by controlling the image quality or, to be more precise, by filtering high-contrast and well-illuminated frontal faces. However, the required image processing approaches are not committed to a particular face recognition system.

The experiments do not systematically address the influence of illumination conditions to the recognition performance. Although the data in OpelDB3 was collected at different locations with different illumination conditions, the one acquisition session does not allow systematical study of changing face appearances in various automotive environments. For future experiments, the data of a limited set of users has to be collected in as many automotive environments as possible involving many acquisition sessions.

Chapter 6

6 Discussion of experimental results

This chapter is devoted to the discourse and generalization of the results of the experiments introduced in the previous chapter. Firstly, the individual recognition performance of the proposed automotive applications (SOD, UDiS and FDR) is discussed, namely which hardware and software solutions are felicitous and which are not. Some further algorithmic solutions are proposed which can theoretically help to improve recognition performance and to overcome complications occurring with currently used algorithms. However, the implementation and evaluation of these solutions are beyond the scope of this thesis. Secondly, it is discussed which algorithms or technical findings from one application can be successfully utilized in others and how all three applications can be implemented in one car, which components can be common and which not.

6.1 Achievements and drawbacks

For designing an automotive application, a concept consisting of five elements is proposed. Firstly, a new application is discovered together with an application scenario. Latter answers the question, what additional value is created by implementing the application in terms of improving safety, security or comfort. For each application scenario, the operational scenarios are specified reflecting how the system interprets the video stream and which reactions are provided. Based on operational scenarios, the requirements regarding hardware components are derived. After the camera setup is chosen, the active illumination sources are located and adjusted. Finally, based on predefined operational scenarios and hardware setup, the image processing and pattern recognition algorithms are specified. Here, each preceding step generates constraints for the succeeding one.

For three automotive applications that have been recognized to be worth implementation (seat occupancy detection (SOD), distinguishing of driver and front-seat passenger hands during the interaction with the center console (UDiS) and facial driver recognition (FDR)), the lists of components can be found in Subsection 3.1 (Table 3.1). The proposed software solutions are listed in Table 3.2, described in detail in Section 3.2 and combined to recognition systems in Chapter 4. Chapter 5 introduces experiments with the recognition systems revealing recognition performances.

6.1.1 Seat occupancy detection

Two basically different approaches are proposed for visual SOD: template matching and face detection. For template matching, the investigation goals are to check how the recognition performance changes by switching from uniform to non-uniform illumination and to find the best parameterization regarding pre- and post-processing. The experiments are organized to determine the proper template selection strategy, to choose the best one from local normalization and edge detection, and to check whether multi-algorithmic and temporal matching score fusion help to reduce error rates. Unfortunately, the experiments do not reveal the clear superiority of any template selection strategy. Two small templates generally perform best. Nonetheless, the way of choosing templates and their quantity remains an open question. Currently, template selection is based on intuition. A methodological approach for choosing template candidates is required. Similarly, a clear conclusion about the superiority of one of the addressed pre-processing techniques cannot be made. However, it can be stated that in case of non-uniform illumination conditions, the utilization of pre-processing is essential and that the temporal matching-score fusion almost always slightly improves recognition performance. Generally, the results show that template matching proves to be a suitable approach. With well selected pre- and post-processing, empty front seats can be recognized with EERs not exceeding 2.60%. For side rear seats, EERs are slightly higher not exceeding 6.23%. The results with non-uniform illumination are in some cases even better than with uniform illumination.

Template matching is based on cross-correlation which is a greedy search for a piece of an empty seat in a seat region. The critical point is that the search for a template is superfluous because after stabilization the location of the template in the seat region is known. Another critical point is that the template is represented by a single image (even if this image is permanently updated). When the lighting quickly changes, the probability of missing the template is high. In order to ensure that a template can be reliably found even if illumination conditions dramatically differ from the reference conditions, the high number of template samples has to be gathered from empty vehicle frames with various illumination conditions. This generalization of the template representation leads to a classification task. The collected template samples represent the set of positive examples. The set of negative examples can be formed from the samples of the same region cut from occupied seats. A classifier is learned from these samples to decide whether the image cut from a certain location belongs to the positive set (empty seat) or to the negative set (occupied seat). Any well-proven classification method e.g. SVM would fit.

Another improvement can be found in long-term temporal fusion or rather switching from considering single frames to considering a seat state over a long period of time. A significant event would be a changing of the state. So long as the state is classified correctly, the falsely classified marginal frames do not destroy recognition performance. The latency of several frames when the state is changed is not critical. Regarding the fact that the state changes only in three cases, a person gets in, gets out, or changes seats, a plausibility model can be created.

Aiming at replacing established non-optical devices such as weight sensors, the error rates of the optical system must approach zero. The EER of even 2% is too high to assert that the system has a potential to replace weight sensors soon. From this point of view, an optical system has to bring an additional value to be installed into a car. Face detection, for instance, can be utilized for recognizing types of occupants.

Three face detection algorithms were tested in an automotive environment. The detection rates are very low when considering frame-based decision making. All face detectors have a total break-down while detecting face profiles at front seats. The detection rates for rear seats are significantly higher, but even they seldom exceed 20% for the best algorithm, the Viola-Jones face detector. The reason for low detection rates is that faces on many frames are invisible to the camera. This can be considered an inherent limitation and does not depend on a particular face detection algorithm. Hence, face detection used alone cannot replace conventional seat occupancy detectors. However, face detection is the most plausible way to maintain conventional systems by providing occupant classification e.g. face-based gender and age recognition as well as localizing head position for out-of-position detection and body size estimation.

Future work should be devoted to designing the combined occupant classification system with errorless recognition of the seat state and the consequent determination of the occupation type (big/medium/small adult or object) and the detection of out-of-position situations.

6.1.2 Distinguishing between driver and front-seat passenger hands during interaction with the center console

Three prototypes of a visual system for distinguishing between driver and front-seat passenger hands while interacting with the center console were designed, implemented and evaluated. The algorithms are modified from version to version, but most important is that each subsequent version has an objective to eliminate shortcomings in the evaluation of the previous one so that resulting recognition rates become plausible and trustworthy.

The first version of the system is implemented in a real car and the experiments are carried out outdoors. With the purpose of covering as many situations as possible, it is endeavored to engage as many test persons as possible, capture images where persons wear different clothes and gloves and all this under non-controlled uniform and non-uniform illumination conditions. Unfortunately, the collected data do not allow a methodological study of how a particular situation influences recognition performance. For instance, we cannot say if wearing gloves increases or decreases error rates. The first step towards methodological evaluation was splitting the data according to uniform and non-uniform illumination conditions and the separate estimation of recognition performances. However, the frame-based evaluation in the first version has proven to be inappropriate. This is why in the second version of action-based evaluation has been applied, counting not frames but missed and falsely detected interactions. Moreover, the second version is designed to enable the control of illumination conditions. A car simulator in the laboratory is more appropriate here. Lighting conditions are simulated by laboratory lamps and sunlight coming from windows. Day, twilight and night illumination conditions as well as intensive side lighting from left and right are reproduced. Another difference is that the system decides whether a driver or passenger provides an action excluding simultaneous actions of both, just as before two simultaneous decisions first for a driver and second for a passenger were made. One problem of the first and second versions is not a quite realistic interaction with the center console and the manual subjective annotation of video frames. In contrast, the third version includes a touch screen and provides a credible and realistic acquisition setup. It includes the simulation of realistic interactions with e.g. radio or navigation and the automatic annotation of video frames.

The imaging algorithms are also developed from the first to the second and to the third versions of UDiS. In the first version, the interactions are indicated solely by the amount of motion in specific driver and passenger regions. In the second version, the motion-based approach is fused with edge detection and followed by silhouette analysis. In the third version, the motion-based approach is supported by hand detection and trajectory analysis. The error rates decrease from the first to the third version, but those improvements cannot be considered significant. Generally, the average EER values are around 10% which is clearly too far away from the required standards to admit the technology mature for integration in a real automobile.

In future works, the latter acquisition system implemented in the car simulator can be used to collect more data addressing not only illumination but also variations in clothing in a controlled methodological way. The image processing algorithms need to be improved or at least better fused within an intelligent decision making system. After the error rates of the new combined UDiS approach zero, the discrimination system should be migrated from the car simulator to a real car followed by carrying out experiments in a moving car.

6.1.3 Facial driver recognition

Two basic objectives are pursued in the experiments on FDR: to generally compare appearance-based and feature-based approaches, and to show that recognition performance rather depends on the operational scenario than on the face recognition algorithm. The two face recognition algorithms involved in the experiments are by far not the best ones, but they are fair representatives of their families: FaceART of appearance-based and Luxand FaceSDK of feature-based approaches. Their positions among state-of-the-art approaches are revealed in the tests with the XM2VTS database using the Lausanne protocol. High recognition performance is not expected with FaceART because appearance-based approaches generally require many reference images for each user covering all possible face appearances. The Lausanne protocol prescribes three training images per user in the 1st and four in the 2nd configuration respectively. Hence, FaceART yields 8% EER in the best case. Luxand FaceSDK has significantly lower error rates yielding 1.94% EER in the best case.

With face databases collected in a real car, two operational scenarios are tested: single driver identification (SDI) and permanent driver verification (PDV). Two datasets are created: OpelDB2 and OpelDB3. They are also combined to OpelDB2+3. While OpelDB2 mainly addresses short-term aging by engaging few users and collecting data over time in five sessions, data in OpelDB3 is collected in one session aiming at engaging as many users as possible. Hence, OpelDB2+3 comprises data from six acquisition sessions with five genuine drivers and 49 impostors enabling an open-set identification test.

In an SDI test with OpelDB2, both face recognition systems demonstrate high identification accuracy. It exceeds 99% with Luxand FaceSDK and 90% with FaceART in all evaluation modes. However, these results are optimistic because all recordings are made during daytime without extreme illumination conditions. With SDI and OpelDB3, the identification accuracy of Luxand FaceSDK approaches 91% and of FaceART even 92% for the best parameterization. These results can be considered more realistic because the adjusted interior illumination allows for being independent from outdoor lighting. The experiments with OpelDB2+3 simulate real-life car usage with few registered drivers and a large set of impostors. The recognition rates in different evaluation modes demonstrate that FaceART requires a mechanism for a permanent update of reference data to operate well. In contrast, Luxand FaceSDK only slightly benefits from updating reference data. Since the number of impostor and genuine trials are very different, the kappa statistic is a more appropriate measure of success here. In the no-aging evaluation mode, kappa approaches 88% with Luxand FaceSDK and 85% with FaceART.

The PDV test is carried out only with OpelDB2. The results are dramatically worse compared to those of the SDI test. For both FaceART and Luxand FaceSDK, the EER values are higher than 20%. Such high errors indicate the inapplicability of face recognition. In fact, a moving car implies varying lighting and dynamic shadows. Moreover, a driver permanently rotates his head for left and right turns, for a broken U-turn and to back into a parking space. As a result, a face is often inaccurately localized. For appearance-based methods, the accurate localization is essential since the images are matched literally pixel-to-pixel. FaceART makes use of an external face detector and includes neither a mechanism to validate the localized face region nor to examine the accuracy of the localization results. Similarly, the feature extraction function of Luxand FaceSDK is incapable of precisely localizing fiducial points on excessively rotated faces.

Generally, the results of experiments with both XM2VTS and Opel databases confirm my theoretical considerations about how important it is to control the quality of enrollment and authentication samples in biometric systems. Simply by guaranteeing frontal faces in reference and test data, recognition performance can be very high, which is proven in the SDI experiment with OpelDB2. The lack of quality control leads to dramatically high error rates, which is proven in the PDV experiments with OpelDB2 and XM2VTS. In fact, the relatively high TER values with Luxand FaceSDK are caused by excessively rotated heads in the second head shot and inability to precisely localize fiducial points. Recognition performance would be drastically improved, reducing error rates to those shown by the best competitors simply by ignoring excessively rotated faces.

The experiments show that a practical integration of face recognition into a car is feasible for at least SDI with enrollment requiring human-aided quality control of the reference data and identification requiring a driver to take a glimpse at the camera. Note, the requirement for high-contrast and well-illuminated frontal faces can be more easily realized by technical solutions (e.g. additional illumination sources or several cameras [157]) than by image processing algorithms. And even these image processing algorithms are not part of a face recognition system. Hence, even non-optimal face recognition software such as FaceART can be successfully utilized in a car.

Future work needs to be devoted to methodological study of how extreme illumination conditions influence recognition accuracy of SDI. The experimental dataset needs to be extended through collecting data in as many automotive environments as possible. The number of drivers can be kept low, but the same data need to be collected in many acquisition sessions.

6.2 Links between hardware and software solutions in different applications

This subsection addresses the common points of the proposed automotive applications. Here, it is discussed which algorithmic or technical findings in one application suit another application. Requirements for integration of all three applications into one car are also discussed. Not only software solutions but also hardware components and eventual joint operational scenarios are addressed.

6.2.1 Image processing and pattern recognition algorithms

As mentioned in Section 6.1, the template matching in SOD can be extended to a classification task making use of any conventional classifier. The first candidate is the SVM approach that is applied in Kienzle's face detection algorithm [127]. The face detection applied for SOD can greatly benefit from several cameras individually used for each seat to observe frontal faces. As shown in FDR, if a camera frontally observes a passenger, the risk to miss a face is significantly lower than if an omni-directional camera observes face profiles. The methodological testing of illumination conditions is another important issue that needs to be addressed in SOD similarly as was done for UDiS. The current error rates do not seem quite trustworthy and certainly include some randomness, which is reflected in EER values with non-uniform illumination which are often lower than those with uniform illumination.

UDiS can benefit from hand tracking. For this purpose template matching from SOD can be utilized. After a hand is localized, a template can be cut and searched for in the subsequent frames, whereby the search region is limited by reasonable surroundings of the initial hand region.

The key issue of FDR is precise face localization. As is the case in UDiS, smart background subtraction can be provided to simplify face detection. The binary silhouette of a foreground can be used as a validation feature for localized faces revealing the exact face contours. FDR fairly well addresses intra-person variations by gathering face images of same persons in several sessions, but still lacks the systematics in testing regarding illumination conditions, which is the main issue of the experiments with UDiS. In UDiS and SOD, however, the data from same persons need to be collected several times to address intra-person variations.

6.2.2 Hardware components and operational scenarios

The omni-directional camera, utilized in a real-car for the prototypic implementation and evaluation of SOD and UDiS, poses serious limitations on image processing algorithms due to the relatively low resolution of 720x576 pixels and the low frame rate of 7.5 FPS. Regarding face detection in SOD, faces of occupants in rear seat regions are seldom larger than 30x30 pixels. The tracking approaches in SOD (a passenger changes seats) and in UDiS (hand moving towards center console) have failed because rapidly moving objects cannot be tracked at the frame rate of 7.5 FPS. Another complication arising with the omni-directional camera is that the data cannot be captured simultaneously in two different modes required for SOD and UDiS respectively (see Subsections 4.1.4 and 4.2.4). While in UDiS the camera view angle fits the application well, in SOD the utilization of several cameras individually used for different seats is preferable. The omni-directional camera cannot even be utilized for FDR because of the impelling need for capturing frontal faces. The camera position in FDR is undisputed, which means if all three applications are implemented in one automobile, at least two cameras are required: an omni-directional camera for SOD and UDiS and a dashboard camera for FDR. The omni-directional camera needs either to possess very high resolution to simultaneously provide sufficiently large images of a center console and passenger seats or to be capable of quickly switching between acquisition modes. Another essential characteristic is a high frame rate enabling tracking of hands while moving towards center console and passengers while changing seats, getting in and getting out.

With regards to artificial illumination in a car cabin, the utilized NIR lamps usually emit light at a wavelength between 850 and 880 nm. Longer wavelengths are weakly perceptible for imaging sensors and shorter wavelengths can be perceptible to a human eye. SOD requires uniform illumination in the whole car passenger compartment. This can be realized by several lamps in A-, B- and C-pillars or by a wide angle lamp in the center of the ceiling. At the moment when the experiments were carried out, the NIR lamps were not installed in the experimental car. The car simulator with the UDiS comprises two wide angle NIR LED lamps (880 nm and 940 nm) in the ceiling next to the windscreen. The 940 nm lamp has proven to be incapable of sufficiently illuminating the cabin in complete darkness. To be more precise, the utilized industrial CCD camera requires more intensive illumination than that provided by the lamp. The 880 nm lamp has proven to be a suitable light source. However, both lamps are switched on to better exploit the effect of skin reflectance under NIR illumination. A hand becomes better visible not only at night but also during daytime. FDR makes use of two NIR lamps. One is used to compensate intensive sunlight from the driver window and located in the ceiling at the position of the conventional cabin lamp. This LED lamp emits light waves with wavelengths of 875 nm which is very similar to one of the lamps in UDiS. The second light source for FDR is the NIR LED ring around the camera lens. Due to the specific skin reflection, the faces become well distinguishable at night and more homogeneous during daytime. If all three applications are integrated in the same car, the lamp used for FDR has to be preserved and the pillar lamps need to be installed to ensure uniform illumination of passenger seats.

The operational scenarios are basically independent from each other except for the fact that SOD, when applied for a front passenger seat, can be used to activate the UDiS when a passenger gets into a car and to deactivate it when a passenger gets out.

Chapter 7

7 Summary, conclusions, ongoing and future work

This chapter concludes the thesis with a brief review of research challenges and research objectives. Here is summed up the extent to which the proposed concept of designing automotive applications as well as the proposed applications themselves cover research gaps and what research questions remain open. The scientific contribution of this work is critically discussed. The ongoing work in the domain of visual monitoring of a car cabin and car surroundings is introduced. Current trends are listed together with examples of ready-for-market systems and technologies. Finally, the perspectives of retrofitting a car with driver assistance and comfort systems that utilize looking-in-car cameras are discussed. Sensor technologies are mentioned that have recently become available on the market and can be superior to conventional cameras in tracking passenger movements and reconstructing the geometry of passenger's body or face.

7.1 Motivation, research challenges and objectives

While the automotive camera systems for monitoring car surroundings have recently become a widely established standard, an investigation of camera systems for monitoring a car passenger compartment is still in its early stages. In fact, front view cameras are currently used for lane departure warning (LDW), traffic sign recognition (TSR), automotive night vision (NV) and pedestrian recognition (PR). Rear view cameras support reverse parking. Side cameras eliminate blind spots. A looking-in-car camera is solely utilized for the Driver Attention Monitor that was introduced by Toyota in 2006 for Lexus and is currently used in their luxury models.

7.1.1 Contributions

This work represents applied research aiming at improvement of safety, security and comfort of an automobile by means of visual monitoring of a car passenger compartment. The contribution to science is in the proposed general methodology for designing automotive applications as a chain of five components: application scenarios, operational scenarios, acquisition devices, artificial illumination sources and computer vision algorithms. This chain can be used to describe and formalize any visual monitoring system. Another contribution is the proposed applications designed and practically implemented from scratch in accordance with the aforementioned methodology including concrete solutions for all five components.

7.1.2 Proposed automotive applications

The general research challenge addressed in this thesis is to indicate existing automotive applications that can benefit from an in-car camera, but currently make use of alternative sensors. A further challenge is to discover completely new applications, which can be realized exclusively using visual sensory devices.

One such application is driver drowsiness detection. Different factors can indicate human drowsiness. These are for example: an increase of electroencephalogram alpha waves, excessive yawning, head nodding, slow eyelid closures or increased number and duration of eye blinks. Hence, drowsiness can be detected not only directly by analyzing deviations in typical driver behavior [113] or by monitoring his eyes [99], but also indirectly by making use of a front view camera that registers lane departure [134]. The monitoring system from Toyota comprises a CCD camera on the steering column and NIR LEDs spread left and right from the camera lens. This system indicates driver drowsiness by analyzing frequency and duration of eye closure events. Driver drowsiness detection or more generally driver distraction detection is undisputedly the most significant safety application because it is the major factor of severe crashes caused by a human factor [130]. However, this application is not designed, implemented and evaluated in this thesis because it has already been well-addressed in other investigations (e.g. [16, 99]) and market solutions exist [99, 137].

Apart from driver drowsiness detection, three automotive applications were recognized to potentially benefit from an in-car camera. These have been addressed in detail including design, implementation and evaluation in the framework of the proposed application designing methodology.

The first application is seat occupancy detection (SOD) currently realized by weight sensors. An SOD system is an indispensable component to maintain smart airbags that has been required by various automotive safety standards e.g. NHTSA [178] for a long time. The alternative occupancy detection systems have been recently investigated and include for instance: capacitive and electric field sensors, ultrasonic sensors and thermal infrared imaging [251]. Conventional imaging systems are also taken into account. An omni-directional camera can be applied for monitoring all seats simultaneously or several cameras can be used to address seats individually.

The second application is distinguishing between driver and passenger hands during interaction with a center console (UDiS). In particular, the discrimination system is required for commonly used components such as a touch screen or a knob manipulator on the armrest (e.g. BMW iDrive or Audi MMI) when the dual-view display is used. Dual-view display is a very modern technology simultaneously showing different contents to a driver and a front-seat passenger. Currently, Range Rover and Jaguar retrofit luxury series of cars with dual-view displays. However, dual-view touch screens are still not presented in series vehicles, so this technology is still undergoing development. Jaguar makes an effort to distinguish driver and passenger interactions by capacitive technologies [194]. A ceiling camera is proposed that faces the center console region and grasps arms and shoulders.

The third application is car personalization currently realized by memory buttons or a smart-key. Memory buttons are standard in upper class vehicles. They implement one of significant comfort functions namely automatic adjustment of the seat, steering wheel, rear-view mirrors, air conditioning and radio. An alternative technology is an individual smart key with a memory function that is rather seldom, but available on the market. For instance Volvo offers a so-called “Cinderella key” to restrict engine power for young inexperienced drivers [65]. The dashboard camera frontally monitoring driver face can be used for biometric driver recognition, whereby the face modality seems to fit better than other modalities. This application is referred to as facial driver recognition (FDR).

Completely new applications making use of a looking-in-car camera are not proposed, which makes it the issue of future work.

7.1.3 Argumentation of using a camera

An important question is to what extent a camera can improve these applications in terms of reducing costs and increasing safety/security/comfort? Answering this question provides an understanding of how far or how close the proposed applications are to practical implementation in series cars.

There are two factors substantiating optical sensors for SOD: reducing costs and providing additional value towards passenger categorization. One omni-directional camera can replace several weight sensors individually installed in seats significantly reducing costs. Moreover, this camera allows for occupancy detection on the rear seats where conventional weight/pressure sensors still have not reached a high saturation level. Additionally, a camera opens new horizons for occupant classification. If a system knows the type of the occupant, the smart airbag can be deployed so as not to injure a passenger, significantly increasing passenger safety.

UDiS is indispensable for proper interaction with an infotainment system when different contents are simultaneously shown to driver and passenger on the same dual-view display. The interaction components become common so that the correct reaction is not possible without knowledge of who is currently active. The general motivation of retrofitting cars with dual-view devices and universal manipulators is cost reduction (e.g. one display instead of two or elimination of conventional buttons) and the simplification of the cockpit for better usability.

FDR has a fundamental advantage over conventional memory buttons and even over individual smart-keys because the biometric identification of a driver provides a direct link between the registered driver and a person on the driver seat and guarantees that the driver's profile is not compromised. The advantage of FDR over memory buttons is the possibility of completely transparent identification leading to comfort improvement. The advantage of FDR over a smart-key is the impossibility to transfer the biometric modality to another person improving anti-theft protection. The system operates as a biometric immobilizer granting engine ignition only to registered drivers. Additionally, facial occupant categorization can be provided towards the determination of age, gender, emotions, gaze direction, head location and rotation angle. This information allows for safety improvement by warning a driver in the collision avoidance stage and by adapting airbag deployment during a crash.

7.1.4 Possible risks evoked by visual sensing

It is one thing to design a system and another thing to integrate a system into a car. Sometimes there are numerous factors that make even very nice concepts unfeasible. In case of visual systems this factor is uncontrolled outdoor lighting. This can be extreme in an automotive environment. Hence, the utilization of a camera is feasible but not trivial requiring scrupulous adjustment of artificial illumination sources in a cabin and proper selection of the camera setup. Note that recognition performance with an optical system cannot be perfect. There are inevitable errors caused by e.g. overexposed images or dynamic shadows. Therefore, the next important question is what additional risks emerge together with the integration of new camera-based systems.

Surely the risk factor depends on whether the system is safety-, security-, or comfort-oriented. For instance, for SOD, if a camera has missed an occupied seat and the airbag is suppressed, there is a risk of injuries or even fatality for a passenger. Missing an out-of-position situation is even worse because a deployed airbag can kill a passenger. Here, errors must be reduced to zero. So a camera currently cannot replace weight sensors but can greatly support them, improving safety standards. In comfort-oriented systems, errors are not that costly. However, high recognition errors can make the system annoying for a driver, which is an important factor for refusing a car purchase.

Regarding security-related tasks namely anti-theft protection, the mechanisms against fooling the system with photographs are required. These anti-spoofing mechanisms are well-studied [38, 133]. Having a video stream, liveness detection can be provided, for instance, based on blinking [188]. Detailed analysis of this issue with respect to automotive environment should be addressed in future work.

7.1.5 Feasibility of visual monitoring

However, the feasibility of automotive applications making use of visual monitoring can be better examined by answering technical questions such as: What is the operational scenario? Which computer vision algorithms are appropriate? How robust can the implementation be? These questions can be posed more generally. Is it possible to robustly fill the semantic gap between objects in the real world and their representations on a camera image by an intelligent computer vision algorithm under the scenario restrictions?

Answering this question is actually the main research objective of this thesis. Moreover, this question is answered with respect to three proposed applications: SOD, UDiS and FDR. Hence, the research objective is to propose a system design including hardware components, application and operational scenarios, and computer vision algorithms, to implement the proposed recognition systems and to evaluate their recognition performance in a statistically significant way. This objective is successfully completed. The detailed description of design and evaluation can be found in chapters 4 and 5. The results are critically discussed in Chapter 6. To sum up the results, SOD has to consider the current status of a seat (empty/occupied) and indicate changes of the status rather than, as is done in the current approach, provide independent decisions for each frame. The template matching for empty seat detection can be considered a suitable solution especially for the front seats, but the improvement through the generalization of a template representation is probable. Face detection requires adaptation of the camera setup. Regarding UDiS, a credible action-based evaluation setup is implemented, but the applied computer vision algorithms deliver unacceptably high error rates. For FDR, acceptable recognition performance is achieved with SDI when the enrollment is controlled by a human supervisor who filters falsely detected faces and identification is provided in a semi-controlled mode requiring a driver to take a glimpse at the camera. For more details, see Chapter 6. All in all, it can be stated that scenario restrictions have more influence on recognition performance than the fine-tuning of image processing algorithms.

The proposed concept of designing automotive applications provides a basis for comparison of computer vision algorithms or their parts. In fact, recognition performance is quantifiable only if the complete chain representing an automotive application from data acquisition to matching is implemented. Computer vision algorithms can be compared by replacing them in the chain and comparing the resulting error rates of the complete application. Even interchangeable image processing approaches as parts of a computer vision algorithm can be compared in this way. It is important to note that the statement that one approach fits better than another can be made only in the framework of the particular application. General statements are not possible. So, based on recognition rates, it can be asserted that one approach fits better than another, or even more, that an approach suits a certain automotive application. However, it cannot be asserted that there are no other approaches that are more appropriate. Hence, the improvement of computer vision algorithms is always an issue.

One question that remains open, however, is if the recognition performance of the introduced visual recognition systems is sufficient to be immediately integrated into a car. This rather non-scientific question cannot be answered within the framework of the scientific investigation without the link to the industrial project.

7.2 Selected topics from ongoing research

The ongoing investigations are completely dedicated to the enhancement of the proposed face recognition software FaceART towards dynamic template preservation. The ARTMAP network seems to fail to properly adapt its internal structure when new training patterns are utilized to adapt a user model.

Another research direction is the training of attribute classifier as in [138]. The binary sub-classifiers are trained to determine for instance age (rather young vs. rather old), gender or other attributes. If such sub-classifiers are applied sequentially, the resulting classifier can become a very powerful tool for user recognition even in big-scale applications.

The improvement of the software towards feature-based classification is a further point of the ongoing work. The cascades are trained to better localize eyes, mouth and nose providing simplified geometrical model. An effort is made towards robust detection of fiducial points.

Finally, the experiments with a Kinect sensor [172] mounted on a dashboard are carried out to compare if the range images has a potential to increase the recognition rates resulting from the experiments with a CCTV camera.

7.3 Possible directions for future research

Aiming at introducing sustainable visual recognition systems, an effort has always to be done towards improving recognition performance. There are several directions for future investigations:

- Adjust hardware components of the system such as camera setup and illumination sources.
- Improve computer vision algorithms applied for the proposed automotive applications.
- Improve the statistical significance of the experimental results by collecting more data and providing new experiments.
- Strengthen the validity of the experimental results by individually addressing more parameters eventually influencing recognition performance. This can also help to discover drawbacks of the system when marginal parameter values crash the recognition algorithm.

7.3.1 Towards improving proposed automotive applications

With regard to SOD, future efforts have to be devoted to the extension of seat occupancy detection to an occupant classification system with close to errorless recognition of the seat status and determination of the occupation type. The system primarily determines whether the seat is occupied by a person or an object. If a person is detected, the body size has to be categorized as large, medium or small. An optional feature is the determination of gender and an age group of an occupant. Finally, a sub-classifier has to determine eventual out-of-position situations.

Future work regarding UDiS has to be first dedicated to improving computer vision algorithms for reliable detection and tracking of passenger hands. The already provided algorithms need to be fused within an intelligent decision making system in a more sophisticated way. Secondly, the acquisition system introduced in the car simulator can be used to collect more recordings with driver/passenger interactions addressing not only different illumination conditions but also various clothing in a methodological way. After the error rates of the new combined UDiS approach zero, the discrimination system should be migrated from the car simulator to a real car followed by carrying out experiments in a moving car.

The foremost effort regarding FDR has to be devoted to the methodological study of how extreme illumination conditions influence recognition accuracy of SDI. The experimental dataset has to be extended through collecting data in as many automotive environments as possible. The number of drivers can be kept low, but the same data need to be collected in many acquisition sessions. Regarding the algorithmic part of face recognition, the future focus has to be on sophisticated pre-processing algorithms:

- to reliably remove background,
- to determine the degree of head rotation,
- to automatically extract high-quality frontal faces.

One simple solution for determination of frontal faces is described in [157]. Three cameras are applied for simultaneous face observation. Two side cameras detect the moment when a face is posed frontally to the center camera making use of face symmetry. The integration of two additional cameras into a car increases the cost of the system but provides a very robust indicator of frames with frontal faces.

Another important research direction in face recognition is the adaptive template preservation that is critically addressed in [164, 196]. However, it has been shown in experiments that the recognition performance with updated templates is significantly better than without. High interest in this topic is confirmed by recent publications [1, 47].

Anti-spoofing in face recognition is another important issue for future work. One option is to solve this problem with image processing algorithms [38]. Another option is to extend the operational scenario. A driver, for instance, can be asked to provide a certain number of blinks after getting in and presenting his face to the camera.

7.3.2 Alternative vision technologies

Proximity sensing using conventional CCD or CMOS imaging sensors [141] is a next step towards more robust and precise visual recognition systems. Range data collected by sensors can be applied for reconstructing geometry of acquired objects. Three technologies are worth mentioning here: stereo imaging, monocular range imaging based on time-of-flight (ToF) measurement and monocular range imaging based on structured light. All three technologies have been tested for SOD in [251].

Stereo imaging reconstructs the range information based on triangulation and principles of epipolar geometry. A stereo camera is applied for occupant classification in [50]. However, monocular range imaging seems to dominate over stereo cameras because of lower costs and significantly simpler and, therefore, more robust algorithms to extract the range data.

A ToF camera is also referred to as a photonic mixer device (PMD). Ringbeck et al. [209] describe operating principles of PMD cameras and explain how the technology is applied for robust object detection, giving examples for an automotive environment. Schöpp et al. [222] introduce the PMD camera for monitoring car surroundings and a car passenger compartment. Current range imaging devices cannot reproduce exact geometry of the captured objects. Rapp et al. [204] discuss systematic errors and statistical uncertainties of ToF cameras using a camera provided by PMDTechnologies GmbH. For many applications, however, the exact geometry is of less importance compared to the general information of captured objects and their relative positions. Occupant classification systems based on a ToF camera are introduced in [80, 84, 49]. Utilization of ToF cameras for face recognition is studied in [57, 56, 166, 11]. While the first three publications make use of the SwissRanger SR-3000 [168], the latter deals with the PMD CamCube 3.0 [195]. Ebers et al. [56] point out the poor quality of ToF data regarding reconstruction of face geometry. So they focus on denoising data and reconstructing trustworthy geometrical face models. The commonly used ToF cameras include NIR lamps and passive coolers expanding the size of the device. The pocket-sized camera (DS325) from SoftKinetics [228] is an example of a device that can be easily integrated into the dashboard of an automobile for capturing faces.

Structured light for deriving range data is applied in a Kinect sensor [172]. This device is successfully used for motion tracking and can be applied in UDiS to track hands and extract their trajectories. Face modeling with Kinect has recently become a focus in entertainment for creating avatars [259]. An approach for face recognition with Kinect is introduced in [145]. The integration of Kinect into a car is impractical because of its dimensions. However, the original manufacturer of the technology behind Kinect, the Israeli developer PrimeSense, offers significantly more compact solutions even for short-range sensing [201].

Despite relatively low recognition rates provided by current systems, range sensing is slowly gaining market positions and will dominate the scene soon because of the plausibility of the measuring technique and the availability of sensors.

7.4 Conclusion

By now, biometric driver recognition is an exclusive option in modern luxury cars. Nonetheless, to the best of my knowledge, face recognition has not yet been implemented in mass-produced cars. The idea of facial driver recognition in a car has been discussed since the late 1990s, but concrete implementation concepts have still not been developed and surely not patented. In the meantime, other camera-based applications such as lane departure warning or driver drowsiness detection have already been developed, patented in the USA as well as in Europe, and installed in series cars. Clearly, the integration of biometric driver recognition into a car faces numerous barriers connected to the inherent properties of biometric applications such as handling the sensitive biometric data and inevitable authentication errors. However, recent trends let us assume the prompt development of automotive-related biometric systems in the near future. Eventually, high recognition rates can be gained with more sophisticated face recognition algorithms using range data along with conventional images. However, a real-time 3D-model fitting is required to assure fast authentication. In this regard, wide horizons are opened by new range data sensing devices e.g. Time-of-Flight cameras or the Kinect sensor.

Bibliography

[1] A. J. Abboud and S. A. Jassim. Biometric templates selection and update using quality measures. In *Proc. SPIE 8406*, pages 840609–9, 2012.

[2] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[3] F. Althoff, R. Lindl, and L. Walchshäusl. Robust multimodal hand- and head gesture recognition for controlling automotive infotainment systems. In *VDI-Tagung - Der Fahrer im 21. Jahrhundert, VDI-Berichte Nr. 1919*, pages 187–205, 2005.

[4] A. Altman and M. Tennenholtz. Ranking systems: The PageRank axioms. In *Proc. 6th ACM Conference on Electronic Commerce*, pages 1–8, 2005.

[5] S. Askar, Y. Kondratyuk, K. Elazouzi, P. Kauff, and O. Schreer. Vision-based skin-colour segmentation of moving hands for real-time applications. In *Proc. 1st European Conference on Visual Media Production*, pages 79–85, 2004.

[6] AudiWorld. Driver-oriented personalisation: one-touch memory. September 8, 2002, `http://www.audiworld.com/news/02/a8launch/content6.shtml`. [accessed 17-11-2013].

[7] A. Augst, S. Durach, M. Fuchs, and S. Weidhaas. Anforderungen an ein Bildverarbeitungssystem für Innenraumkamera in Premiumfahrzeugen. In *Optische Technologien in der Fahrzeugtechnik, VDI-Berichte Nr. 1731*, pages 223–232, 2003.

[8] F. Aurenhammer. Voronoi Diagrams – A survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.

[9] M. Barry and E. Granger. Comparison of ARTMAP neural networks for classification for face recognition from video. In *Proc. 4th International Conference on Image Analysis and Recognition (ICIAR)*, pages 794–805, 2007.

[10] BAST. Zahl der Getöteten im Straßenverkehr sinkt wieder. 18.12.2012, Nr.: 30/2012, `http://www.bast.de/cln_033/nn_42254/DE/Presse/2012/presse-30-2012.html`. [accessed 17-11-2013].

[11] S. Bauer, J. Wasza, K. Müller, and J. Hornegger. 4D Photogeometric face recognition with time-of-flight sensors. In *Proc. IEEE Workshop on Applications of Computer Vision (WACV)*, pages 196–203, 2011.

[12] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[13] R. E. Bellman. *Dynamic programming.* Princeton University Press, 1957.

[14] K.S. Benli, R. Düzagac, and M. T. Eskil. Driver recognition using gaussian mixture models and decision fusion techniques. In *Proc. 3rd International Symposium on Advances in Computation and Intelligence (ISICA), LNCS 5370*, pages 803–811, 2008.

[15] K. P. Bennett and C. Campbell. Support vector machines: Hype or hallelujah? *SIGKDD Explorations*, 2(2):1–13, 2000.

[16] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez. Real-time system for monitoring driver vigilance. *IEEE Trans. on Intelligent Transportation Systems*, 7(1):63–77, 2006.

[17] M. Bertozzi, A. Broggi, and A. Fascioli. Vision-based intelligent vehicles: State of the art and perspectives. *Robotics and Autonomous Systems*, 32(1):1–16, 2000.

[18] R. Bhatt, G. Carpenter, and S. Grossberg. Texture segregation by visual cortex: perceptual grouping, attention, and learning. *Vision Research*, 47(25):3173–3211, 2007.

[19] BiometricTechnologyToday. Iris recognition option for car occupants. *Biometric Technology Today*, 15(11-12):12, November–December 2007.

[20] U. Büker and R. Schmidt. Biometrische Fahreridentifikation. In *Automotive Security, VDI-Berichte Nr. 2016*, pages 95–110, 2007.

[21] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999.

[22] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to Biometrics.* Springer Verlag, 2003.

[23] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. The relation between the ROC curve and the CMC. In *Proc. 4th IEEE Workshop on Automatic Identification Advanced Technologies (AUTOID)*, pages 15–20, 2005.

[24] G. Bradski, G. Carpenter, and S. Grossberg. Working memory networks for learning temporal order with application to 3-D visual object recognition. *Neural Computation*, 4:270–286, 1992.

[25] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees.* Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984.

[27] J. Brodkin. Facial recognition software used for national security. 2008, `http://www.itbusiness.ca/news/facial-recognition-software-used-for-national-security/1712`. [accessed 17-11-2013].

[28] C. Burges. Simplified support vector decision rules. In *Proc. International Conference on Machine Learning*, pages 71–77, 1996.

[29] B. N. Campbell, J. D. Smith, and W. G. Najm. Examination of crash contributing factors using national crash databases. Technical report, John A. Volpe National Transportation Systems Center, April 2002.

[30] J. A. Canny. Computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[31] A. Carleson, C. Cumby, J. Rosen, and D. Roth. The SNoW learning architecture. Technical Report UIUCDCS-R-99-2101, UIUC Computer Science Department, 1999.

[32] G. Carpenter and S. Grossberg. Adaptive resonance theory: self-organizing networks for stable learning, recognition, and prediction. In E. Fiesler and R. Beale, editors, *Handbook of Neural Computation*. IOP Publishing Ltd and Oxford University Press, 1997.

[33] G. Carpenter, S. Grossberg, N. Markuzon, and D. B. Rosen. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *Neural Networks*, 3(5):698–713, 1992.

[34] R. Chen, M. F. She, X. Sun, L. Kong, and Y. Wu. Driver recognition based on dynamic handgrip pattern on steering wheel. In *Proc. 12th ACIS Int. Conf. on Software Eng., Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 107–112, 2011.

[35] T. Chen, D. Breed, and K. Xu. Development of an optical occupant position sensor system to improvement frontal crash detection. In *Proc. 18th Int. Tech. Conf. on the Enhanced Safety*, pages 426–433, 2003.

[36] Y. Chen, J. Liu, C. Tsai, and C. Chen. Anti-counterfeiting system of drunk driving using driver's facial image identification. In *SAE Technical Paper 2011-01-0210*, 2011.

[37] S. Y. Cheng and M. M. Trivedi. Vision-based infotainment user determination by hand recognition for driver assistance. *IEEE Trans. on Intelligent Transportation Systems*, 11(3):759–764, 2010.

[38] I. Chingovska and other. The 2nd competition on counter measures to 2D face spoofing attacks. In *Proc. 6th International Conference of Biometrics (ICB)*, pages 1–6, 2013.

[39] K. Choudhary and N. Goel. A review on face recognition techniques. In *Proc. SPIE 8760*, pages 87601E–10, 2013.

[40] C. K. Chow and T. Kaneko. Boundary detection of radiographic images by threshold method. In S. Watanabe, editor, *Proc. International Conference on Frontiers of Pattern Recognition*, pages 61–82, 1972.

[41] C. K. Chui. *An Introduction to Wavelets*. Academic Press, San Diego, 1992.

[42] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[43] C. Cortes and V. N. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[44] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[45] DaimlerAG. Attention Assist: Drowsiness-detection system warns drivers to prevent them falling asleep momentarily. `http://media.daimler.com/dcmedia/0-921-614216-1-1147698-1-0-1-999999-0-0-12637-614216-0-1-0-0-0-0-0.html`. [accessed 17-11-2013].

[46] J. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, 1988.

[47] M. De-la Torre, E. Granger, P.V.W. Radtke, R. Sabourin, and D.O. Gorodnichy. Incremental update of biometric models in face-based video surveillance. In *Proc. Int. Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2012.

[48] B. DeCann and A. Ross. Can a 'poor' verification system be a 'good' identification system? A preliminary study. In *Proc. IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 31–36, 2012.

[49] P. R. Devarakota, M. Castillo-Franco, R. Ginhoux, B. Mirbach, and B. Ottersten. Occupant classification using range images. *IEEE Trans. on Vehicular Technology*, 56(4):1983–1993, 2007.

[50] P. R. Devarakota, B. Mirbach, M. Castillo-Franco, and B. E. Ottersten. 3-D vision technology for occupant detection and classification. In *Proc. International Conference on 3-D Digital Imaging and Modeling (3DIM)*, pages 72–79, 2005.

[51] L. Devroye, L. Gyorfi, and G Lugosi. *A probabilistic theory of pattern recognition.* Springer, 1996.

[52] P. Dietz and D. Leigh. Diamondtouch: a multi-user touch technology. In *Proc. 14th Annual ACM Symposium on User Interface Software and Technology*, pages 219–226, 2001.

[53] D. F. Dinges, M. M. Mallis, G. Maislin, and J. W. Powell. Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. Technical report, National Highway Traffic Safety Administration, 1998.

[54] M. Duckwitz. Best-fit plane removal. Technical notes, January 24, 2007, `http://hkn.colorado.edu/resources/latex/plane-removal/plane-removal.pdf`. [accessed 17-11-2013].

[55] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley-Interscience, 2nd edition, 2000.

[56] O. Ebers, T. Ebers, M. Plaue, T. Spiridonidou, G. Bärwolff, and H. Schwandt. Study on three-dimensional face recognition with continuous-wave time-of-flight range cameras. *Optical Engineering*, 50(6):063201, 2011.

[57] O. Ebers, T. Ebers, T. Spiridonidou, M. Plaue, P. Beckmann, G. Bärwolff, and H. Schwandt. Towards robust 3D face recognition from noisy range images with low resolution. Technical report, Technische Universität Berlin, 2008.

[58] D. W. Eby and L. P. Kostyniuk. Safety vehicles using adaptive interface technology (task 1): Distracted-driving scenarios: A synthesis of literature, 2001 Crashworthiness Data System (CDS) data, and expert feedback. Technical report, University of Michigan, Transportation Research Institute, 2004.

[59] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[60] M. L. Eichner and T. P. Breckon. Integrated speed limit detection and recognition from real-time video. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 626–631, 2008.

[61] A. Eleyan and H. Demirel. PCA and LDA based neural networks for human face recognition. In K. Delac and M. Grgic, editors, *Face Recognition*, chapter 6, pages 93–106. I-Tech, 2007.

[62] EPFL. Local Normalization. `http://bigwww.epfl.ch/demo/jlocalnormalization/index.html`. [accessed 17-11-2013].

[63] H. Erdogan, A. Ercil, H. K. Ekenel, S. Y. Bilgin, I. Eden, M. Kirisci, and H. Abut. Multimodal person recognition for vehicular applications. In N. C. Oza and other, editors, *Proc. 6th Int. Conf. on Multiple Classifier Systems (MCS), LNCS 3541*, pages 366–375, 2005.

[64] C. W. Erwin. Studies of drowsiness. Technical report, The National Driving Center, Durham, NC, 1976.

[65] A. Eugensson. Volvo Vision 2020. `http://www.unece.org/fileadmin/DAM/trans/roadsafe/unda/Sweden_Volvo_Vision_2020.pdf`. [accessed 17-11-2013].

[66] EuropeanParliament. Answer given by Mr. Tajani on behalf of the Commission, 24 January 2012. `http://www.europarl.europa.eu/sides/getAllAnswers.do?reference=E-2011-011477&language=EN`. [accessed 17-11-2013].

[67] C. Evers and K. Auerbach. Behaviour-related causes of severe truck accidents. Berichte der Bundesanstalt für Straßenwesen, Reihe M: Mensch und Sicherheit, volume 174, 2011, `http://bast.opus.hbz-nrw.de/frontdoor.php?source_opus=211&la=en`. [accessed 17-11-2013].

[68] M. E. Farmer and A. K. Jain. Smart automotive airbags: Occupant classification and tracking. *IEEE Trans. on Vehicular Technology*, 56(1):60–80, 2007.

[69] D. Fell. Safety update: problem definition and countermeasure summary: Fatigue. Technical report, New South Wales Road Safety Bureau, Australia, 1994.

[70] W. Feller. *An Introduction to Probability Theory and Its Applications (Volume 1), Section VII.3*. Wiley, 1968.

[71] R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[72] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[73] W. J. Fleming. Overview of automotive sensors. *IEEE Sensors Journal*, 1(4):296–308, 2001.

[74] L. Fletcher and A. Zelinsky. Driver inattention detection based on eye gaze - Road event correlation. *Int. Journal of Robotics Research*, 28(6):774–801, 2009.

[75] FordMotorCompany. Lane Departure Warning and Lane Keeping Aid. `http://www.ford.ie/Technology/LaneDepartureAndLaneKeeping`. [accessed 17-11-2013].

[76] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. 2nd European Conference on Computational Learning Theory (EuroCOLT)*, pages 23–37, 1995.

[77] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*, pages 148–156, 1996.

[78] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[79] Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[80] M. Fritzsche, C. Prestele, G. Becker, M. Castillo-Franco, and B. Mirbach. Vehicle occupancy monitoring with optical range-sensors. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 90–94, 2004.

[81] FRVT. Face Vendor Recognition Test. `http://www.nist.gov/itl/iad/ig/frvt-home.cfm`. [accessed 17-11-2013].

[82] M. Fukumi. Driver face monitoring using a near-infrared camera. In M.W. Marcellin, editor, *Proc. IASTED Int. Conf. on Signal and Image Processing (SIP)*, pages 156–160, 2005.

[83] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[84] S. B. Goktuk and A. Rafii. An occupant classification system eigen shapes or knowledge-based features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPRW)*, pages 57–64, 2005.

[85] M. Golz, D. Sommer, M. Chen, U. Trutschel, and D. Mandic. Feature fusion for the detection of microsleep events. *Signal Processing Systems for Signal, Image, and Video Technology, Special Issue: Data Fusion for Medical, Industrial, and Environmental Applications*, 49(2):329–342, 2007.

[86] R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* Pearson Education, 2nd edition, 2002.

[87] R. C. Gonzalez and R. E. Woods. Histogram equalization. In *Digital Image Processing*, pages 91–93. Pearson Education, 2nd edition, 2002.

[88] R. C. Gonzalez and R. E. Woods. Morphological image processing. In *Digital Image Processing*, pages 519–566. Pearson Education, 2nd edition, 2002.

[89] R. C. Gonzalez and R. E. Woods. Thresholding. In *Digital Image Processing*, pages 595–611. Pearson Education, 2nd edition, 2002.

[90] P. Grünwald. A tutorial introduction to the minimum description length principle. In *Advances in Minimum Description Length: Theory and Applications.* MIT Press, 2005.

[91] S. Grossberg. Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23(3):121–134, 1976.

[92] S. Grossberg. Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, 23(3):187–202, 1976.

[93] P. Grother and P. J. Phillips. Models of large population recognition performance. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 68–75, 2004.

[94] M. Gu, J.-Z. Zhou, and J.-Z. Li. Online face recognition algorithm based on fuzzy ART. In *Proc. Int. Conf. on Machine Learning and Cybernetics*, volume 1, pages 556–560, 2008.

[95] S. R. Gunn. On the discrete representation of the Laplacian of Gaussian. *Pattern Recognition*, 32(8):1463–1472, 1999.

[96] K. L. Gwet. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment, No. 1*, 2002.

[97] K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

[98] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910.

[99] R. Hammoud, G. Witt, R. Dufour, A. Wilhelm, and other. On driver eye closure recognition for commercial vehicles. *SAE Int. Journal of Commercial Vehicles*, 1(1):454–463, 2009.

[100] R. M. Haralick and K. Shanmugam. Textural features for image classification systems. *IEEE Trans. on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973.

[101] V. Hargutt, S. Hoffmann, M. Vollrath, and H.-P. Krüger. Compensation for drowsiness and fatigue. In T. Rothengatter and R. D. Huguenin, editors, *Traffic and Transport Psychology - Theory and Application*, volume 1, pages 257–266. Elsevier, Amsterdam, 2004.

[102] K. Hayashi, K. Ishihara, H. Hashimoto, and K. Oguri. Individualized drowsiness detection during driving by pulse wave analysis with neural network. In *Proc. IEEE Intelligent Transportation Systems*, pages 901–906, 2005.

[103] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, 2nd edition, 1998.

[104] M. Hazewinkel. Inner product. In *Encyclopedia of Mathematics.* Springer, 2001.

[105] G. Heitink. *Practical Theology: History, Theory, Action Domains: Manual for Practical Theology.* Grand Rapids, MI: Wm. B. Eerdmans Publishing, 1999.

[106] K. Hellwig. Viewing Assistance für den Durchblick. *E&E 04.2009, Abenteuer Anwendung*, pages 48–50, 2009.

[107] E. Herrmann, A. Makrushin, J. Dittmann, and C. Vielhauer. Driver/passenger discrimination for the interaction with the dual-view touch screen integrated to the automobile centre console. In *Proc. SPIE 8295*, page 82950W, 2012.

[108] E. Herrmann, A. Makrushin, J. Dittmann, C. Vielhauer, M. Langnickel, and C. Kraetzer. Hand-movement-based in-vehicle driver/front-seat passenger discrimination for centre console controls. In *Proc. SPIE 7532*, pages 75320U–9, 2010.

[109] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll. Bimodal fusion of emotional data in an automotive environment. In *Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 1085–1088, 2005.

[110] J. A. Horne and L. A. Reyner. Sleep-related vehicle accidents. *British Medical Journal*, 310(6979):565–567, 1995.

[111] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Trans. on Information Theory*, 8(2):179–187, 1962.

[112] J. Huppertz, R. Hauschild, B. J. Hosticka, T. Kneip, S. Müller, and M. Schwarz. Fast CMOS imaging with high dynamic range. In *Proc. IEEE Workshop Charge Coupled Devices and Advanced Image Sensors*, pages R7/1–R7/4, 1997.

[113] A. Hussain, B. Bais, S. Samad, and S. Farshad Hendi. Novel data fusion approach for drowsiness detection. *Information Technology Journal*, 7(1):48–55, 2008.

[114] K. Igarashi, K. Takeda, F. Itakura, and H. Abut. Biometric identification using driving behavioral signals. In *DSP for In-Vehicle and Mobile Systems, Chapter 17*. Springer Science, New York, 2005.

[115] Intel. OpenCV programming library. `http://opencv.org/`. [accessed 17-11-2013].

[116] K. A. Ishak, S. A. Samad, and A. Hussain. A face detection and recognition system for intelligent vehicles. *Information Technology Journal*, 5(3):507–515, 2006.

[117] ISO. Standard ISO/IEC 2382-37:2012 Information technology – Vocabulary – Part 37: Biometrics, 2012.

[118] R. Jafri and H. R. Arabnia. A survey of face recognition techniques. *Journal of Information Processing Systems*, 5(2):41–68, 2009.

[119] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 14(1):1–29, 2004.

[120] A.K. Jain, S.C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Proc. Int. Conf. on Biometric Authentication, LNCS 3072*, pages 731–738, 2004.

[121] J. Janesick. Dueling detectors. CMOS or CCD? *SPIE's OE Magazine*, pages 30–33, February 2002.

[122] Q. Ji and X. Yang. Real-time eye, gaze and face pose tracking for monitoring driver vigilance. *Real-Time Imaging*, 8(5):357–377, 2002.

[123] H. Jobling. Range Rover announces 'dual-view' touchscreen. April 8, 2009, `http://www.trustedreviews.com/news/Range-Rover-Announces--Dual-View--Touchscreen`. [accessed 17-11-2013].

[124] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proc. 11th International Conference on Machine Learning*, pages 121–129, 1994.

[125] T. Kasuba. Simplified fuzzy ARTMAP. *AI Expert*, pages 18–25, 1993.

[126] N. Kawaguchi and other. Construction and analysis of the multi-layered in-car spoken dialogue corpus. In H. Abut, J. H. L. Hansen, and K. Takeda, editors, *DSP in Vehicular and Mobile Systems*, pages 1–18. Springer, 2005.

[127] W. Kienzle, G. Bakir, M. Franz, and B. Scholkopf. Face detection - efficient and rank deficient. In *Proc. Advances in Neural Information Processing Systems*, volume 17, pages 673–680, 2005.

[128] R. Kimmel and J. A. Sethian. Optimal algorithm for shape from shading and path planning. *Journal of Mathematical Imaging and Vision*, 14(3):237–244, 2001.

[129] K. Kira and L. A. Rendell. The feature selection problem: traditional methods and a new algorithm. In *Proc. 10th National Conf. on Artificial Intelligence*, pages 129–134, 1992.

[130] S. G. Klauer, T. A. Dingus, V. L. Neele, J. D. Sudweeks, and D. J. Rasey. The impact of driver inattention on nearcrash/crash risk: An analysis using the 100-car naturalistic driving study data. Technical report, Virginia Tech Transportation Institute, 2006.

[131] P. R. Knipling and S. S. Wang. Revised estimates of the US drowsy driver crash problem size based on general estimates system case reviews. In *Proc. 39th Annual Meeting of Association for the Advancement of Automotive Medicine*, pages 451–466, 1995.

[132] C. Koch, T. J. Ellis, and A. Georgiadis. Real-time occupant classification in high dynamic range environments. In *Proc. IEEE Intelligent Vehicle Symposium*, volume 2, pages 284–291, 2002.

[133] J. Komulainen, A. Hadid, and M. Pietikainen. Face spoofing detection using dynamic texture. In *ACCV 2012 Workshops, Part I (LBP 2012), LNCS 7728*, pages 146–157, 2013.

[134] K. Kozak, J. Pohl, W. Birk, J. Greenberg, B. Artz, M. Blommer, and other. Evaluation of lane departure warnings for drowsy drivers. In *Proc. Human factors and ergonomics society 50th annual meeting*, pages 2400–2404, 2006.

[135] S. J. Krotosky, S. Y. Cheng, and M. M. Trivedi. Real-time stereo-based head detection using size, shape and disparity constraints. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 550–556, 2005.

[136] J. Krumm and G. Kirk. Video occupant detection for airbag deployment. In *Proc. 4th IEEE Workshop on Applications of Computer Vision (WACV)*, pages 30–35, 1998.

[137] T. Kruscha, M. Langnickel, and S. Tuscheerer. DE Patent DE 10 2010 044 449 A1 2011.07.07, Erkennen des Grades der Fahrfähigkeit des Fahrers eines Kraftfahrzeugs, 2011.

[138] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. IEEE 12th Int. Conf. on Computer Vision*, pages 365–372, 2009.

[139] M. Kutila. *Methods for Machine Vision Based Driver Monitoring Applications.* PhD thesis, Tampere University of Technology, 2006.

[140] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[141] R. Lange. *3D Time-of-Flight distance measurement with custom solid-state image sensors in CMOS/CCD technology.* PhD thesis, University of Siegen, 2000.

[142] M. Langnickel and K. Seifert. Anforderungen und Möglichkeiten - Biometrie im Kontext automobiler Applikationen. In *Automotive Security, VDI-Berichte Nr. 2016*, pages 111–129, 2007.

[143] M. Langnickel, S. Tuchscheerer, and K. Seifert. Besondere Anforderungen an biometrische Verfahren im Fahrzeugkontext. In *Proc. Special Interest Group on Biometrics and Electronic Signatures (BIOSIG), LNI Volume P-108*, pages 47–60, 2007.

[144] K. I. Laws. Rapid texture identification. In *Proc. SPIE Conference on Image Processing for Missile Guidance*, pages 376–380, 1980.

[145] B. Y. L. Li, A. S. Mian, Wanquan Liu, and A. Krishna. Using Kinect for face recognition under varying poses, expressions, illumination and disguise. In *Proc. IEEE Workshop on Applications of Computer Vision (WACV)*, pages 186–192, 2013.

[146] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Proc. DAGM 25th Pattern Recognition Symposium*, pages 297–304, 2003.

[147] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In *Proc. IEEE International Conference on Image Processing*, volume 1, pages 900–903, 2002.

[148] P. Lilly. Google patents facial recognition for Android security. 2013, `http://hothardware.com/News/-Google-Patents-Facial-Recognition-For-Android-Security/`. [accessed 17-11-2013].

[149] C. Lin, R. Wu, S. Liang, W. Chao, Y. Chen, and T. Jung. EEG-based drowsiness estimation for safety driving using independent component analysis. *IEEE Trans. on Circuits and Systems*, 52(12):2726–2738, 2005.

[150] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

[151] D. Litwiller. CCD vs. CMOS: Facts and Fiction. *Photonics Spectra, Laurin Publishing Co. Inc.*, January 2001.

[152] D. Litwiller. CMOS vs. CCD: Maturing technology, maturing markets. *Photonics Spectra, Laurin Publishing*, August 2005.

[153] X. Lu. Image analysis for face recognition. Personal notes, Michigan State University, 2003, `http://www.face-rec.org/interesting-papers/general/imana4facrcg_lu.pdf`. [accessed 17-11-2013].

[154] J. Luettin and G. Maître. Evaluation protocol for the the XM2VTS database (Lausanne protocol). Technical report, IDIAP, 1998.

[155] LuxandInc. Luxand FaceSDK. `http://www.luxand.com/facesdk/`. [accessed 17-11-2013].

[156] P. C. Mahalanobis. On the generalised distance in statistics. *Proc. National Institute of Sciences of India*, 2(1):49–55, 1936.

[157] A. Makrushin. Biometrische Gesichtserkennung mit multiplen Infrarot-Sensoren. Master's thesis, Otto-von-Guericke Universität Magdeburg, 2007.

[158] A. Makrushin, J. Dittmann, C. Vielhauer, M. Langnickel, and C. Kraetzer. The feasibility test of state-of-the-art face detection algorithms for vehicle occupant detection. In *Proc. SPIE 7532*, pages 75320V–10, 2010.

[159] A. Makrushin, J. Dittmann, C. Vielhauer, and M. Leich. User discrimination in automotive systems. In *Proc. SPIE 7870*, page 78700J, 2011.

[160] A. Makrushin, M. Langnickel, M. Schott, C. Vielhauer, J. Dittmann, and K. Seifert. Car-seat occupancy detection using a monocular 360° NIR camera and advanced template matching. In *Proc. 16th Int. Conf. on Digital Signal Processing*, pages 1–6, 2009.

[161] A. Makrushin, C. Vielhauer, and J. Dittmann. The impact of ARTMAP to appearance-based face verification. In S. Craver and J. Dittmann, editors, *Proc. 12th ACM Workshop on Multimedia and Security (MM&Sec '10)*, pages 89–94, 2010.

[162] S. Mann and R. W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proc. 48th Annual Conference Society for Imaging Science and Technology (IS&T)*, pages 442–448, 1995.

[163] C. Manning, P. Raghavan, and H. Schütze. Vector space classification. In *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[164] G. L. Marcialis, L. Didaci, A. Pisano, E. Granger, and F. Roli. Why template self-update should work in biometric authentication systems? In *Proc. 11th Int. Conf. on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1086–1091, 2012.

[165] J. Matas and other. Comparison of face verification results on the XM2VTS database. In *Proc. 15th Int. Conf. on Pattern Recognition*, volume 4, pages 858–863, 2000.

[166] S. Meers and K. Ward. Face recognition using a Time-of-Flight camera. In *Proc. 6th Int. Conf. on Computer Graphics, Imaging and Visualization (CGIV)*, pages 377–382, 2009.

[167] J. P. Mello. Facial recognition technology: Facebook photo matching is just the start. PCWorld, 2011, `http://www.techhive.com/article/240363/facial_recognition_technology_facebook_photo_matching_is_just_the_start.html`. [accessed 17-11-2013].

[168] MesaImaging. Mesa Imaging SwissRanger SR-3000. `http://www.mesa-imaging.ch/prodviews.php`. [accessed 17-11-2013].

[169] K. Messer, J. Kittler, J. Short, G. Heusch, F. Cardinaux, S. Marcel, Y. Rodriguez, S. Shan, Y. Su, and X. Chen. Performance characterisation of face recognition algorithms and their sensitivity to severe illumination changes. In D. Zhang and A. K. Jain, editors, *Advances in Biometrics, LNCS 3832*, pages 1–11, 2006.

[170] K. Messer, J. Matas, J. Kittler, and K. Jonsson. XM2VTSDB: The Extended M2VTS Database. In *Proc. 2nd Int. Conf. on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[171] K. Messer and other. Face verification competition on the XM2VTS database. In *Proc. 4th Int. Conf. Audio and Video Based Biometric Person Authentication*, pages 964–974, 2003.

[172] MicrosoftCorporation. Kinect for Xbox 360. `http://www.xbox.com/en-US/kinect`. [accessed 17-11-2013].

[173] B. Miller. Vital signs of identity. *IEEE Spectrum*, 31(2):22–30, 1994.

[174] H. Moon and K. Lee. Biometric driver authentication based on 3D face recognition for telematics applications. In C. Stephanidis, editor, *Proc. 4th Int. Conf. on Universal Access in Human-Computer Interaction (UAHCI), Part I, LNCS 4554*, pages 473–480, 2007.

[175] B. Moore. ART 1 and pattern clustering. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proc. Connectionist Models Summer School*, pages 174–185, 1988.

[176] T. Nakano, K. Sugiyama, M. Mizuno, and S. Yamamoto. Blink measurement by image processing and application to warning of driver's drowsiness in automobiles. In *Proc. IEEE International Conference on Intelligent Vehicles*, pages 285–290, 1998.

[177] NHTSA. National Highway Traffic Safety Administration, Air Bag Fatalities, Statistical Breakdown of Air Bag Fatalities. 26.02.2008, `http://web.archive.org/web/20080226234316/http://www.nsc.org/partners/status3.htm`. [accessed 17-11-2013].

[178] NHTSA. National Highway Traffic Safety Administration, Federal Motor Vehicle Safety Standard (FMVSS) No. 208, Occupant Crash Protection, 2001.

[179] M. Nilsson, M. Dahl, and I. Claesson. The successive mean quantization transform. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 429–432, 2005.

[180] M. Nilsson, J. Nordberg, and I. Claesson. Face detection using local SMQT features and split up SNoW classifier. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 589–592, 2007.

[181] NIST. National Science and Technology Council (NSTC): Biometrics Glossary. 2006, `http://biometrics.gov/Documents/Glossary.pdf`. [accessed 17-11-2013].

[182] M. Nixon and A. S. Aguado. *Feature Extraction and Image Processing.* Academic Press, 2nd edition, 2008.

[183] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer-Verlag, Berlin, New York, 2nd edition, 2006.

[184] T. Ogawa and K. Takagi. Lane recognition using on-vehicle LIDAR. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 540–545, 2006.

[185] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[186] C. A. Opperman and G. P. Hancke. A generic NFC-enabled measurement system for remote monitoring and control of client-side equipment. In *Proc. 3rd Int. Workshop on Near Field Communication (NFC)*, pages 44–49, 2011.

[187] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[188] G. Pan, Z. Wu, and L. Sun. Liveness detection for face recognition. In K. Delac, M. Grgic, and M. S. Bartlett, editors, *Recent Advances in Face Recognition*, page 236–252. InTech, 2008.

[189] C. Papadelis, C. Kourtidou-Papadeli, P. Bamidis, I. Chouvarda, D. Koufogiannis, E. Bekiaris, and N. Maglaveras. Indicators of sleepiness in an ambulatory EEG study of night driving. In *Proc. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6201–6204, 2006.

[190] S.-B. Park. *Optische Kfz-Innenraumüberwachung.* PhD thesis, University of Duisburg, 2000.

[191] J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving.* Addison-Wesley, New York, 1983.

[192] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[193] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[194] C. A. Pickering, R. Thorp, and K. J. Burnham. Automotive human machine interface user discrimination with low frequency electric field sensors. In *Proc. 28th International Conference on Systems Engineering*, pages 359–364, 2006.

[195] PMDTechnologies. pmd[vision] CamCube 3.0. `http://www.pmdtec.com/news_media/video/camcube.php`. [accessed 17-11-2013].

[196] N. Poh, A. Rattani, and F. Roli. Critical analysis of adaptive biometric systems. *IET Biometrics*, 1(4):179–187, 2012.

[197] R. Polikar. Ensemble learning. *Scholarpedia*, 4(1):2776, 2009.

[198] B. Popa. How Volvo's pedestrian protection system works. 2009, `http://www.autoevolution.com/news/how-volvo-s-pedestrian-protection-system-works-11554.html`. [accessed 17-11-2013].

[199] J. E. Porter. On the '30 error' criterion. In *National Biometric Test Center - Collected Works - 1997-2000*, pages 51–56. San Jose State University, 1977.

[200] W. K. Pratt. Morphological image processing. In *Digital Image Processing*, pages 401–442. John Wiley and Sons, 2nd edition, 2001.

[201] PrimeSenseLTD. PrimeSense 3D Sensors. `http://www.primesense.com/solutions/3d-sensor/`. [accessed 17-11-2013].

[202] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[203] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[204] H. Rapp, M. Frank, F. A. Hamprecht, and B. Jahne. A theoretical and experimental investigation of the systematic errors and statistical uncertainties of time-of-flight cameras. *Int. Journal of Intelligent Systems Technologies and Applications*, 5(3/4):402–413, 2008.

[205] P. S. Rau. Drowsy driver detection and warning system for commercial vehicle drivers: Field operational test design, data analysis, and progress. In *Proc. 19th Int. Tech. Conference on the Enhanced Safety of Vehicles*, Washington DC, 2005.

[206] K. Reif. *Automobilelektronik. Eine Einführung für Ingenieure.* Vieweg Verlag, Berlin, 2006.

[207] A. Reiner. Gestural interaction in vehicular applications. *Computer*, 45(4):42–47, 2012.

[208] A. Riener and A. Ferscha. Supporting implicit human-to-vehicle interaction: Driver identification from sitting postures. In *Proc. 1st Annual Int. Symposium on Vehicular Computing Systems (ISVCS)*, 2008.

[209] T. Ringbeck and B. Hagebeuker. A 3D Time of Flight camera of object detection. In *Optical 3-D Measurement Techniques 09-12.07.2007 ETH Zürich*, 2007.

[210] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[211] RobertBoschGmbH. Touchscreen-Bildschirm von Bosch. 2010, `http://www.bosch-presse.de/presseforum/details.htm?txtID=4360`. [accessed 17-11-2013].

[212] A. Rosenfeld and A. C. Kak. *Digital Picture Processing.* Academic Press, Inc. Orlando, FL, USA, 2nd edition, 1982.

[213] L. Sachs. *Applied Statistics: A Handbook of Techniques.* Springer, 2nd edition, 2012.

[214] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[215] W. S. Sarle. Why statisticians should not FART. SAS Institute, Cary, NC, USA, Aug. 1, 1995, `http://medusa.sdsu.edu/Robotics/Neuromuscular/Articles/ATM_articles/fart.txt`. [accessed 17-11-2013].

[216] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[217] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1686, 1998.

[218] H. Scharr. *Optimal Operators in Digital Image Processing.* PhD thesis, Universität Heidelberg, 2000.

[219] T. Scheidat, M. Biermann, J. Dittmann, C. Vielhauer, and K. Kümmel. Multi-biometric fusion for driver authentication on the example of speech and face. In J. Fierrez-Aguilar and other, editors, *Proc. BioID MultiCom., LNCS 5707*, pages 220–227, 2009.

[220] B. Schneider, H. Fischer, S. Benthien, and other. TFA image sensors: from the one transistor cell to a locally adaptive high dynamic range sensor. In *Technical Digest of International Electronic Devices Meeting*, pages 209–212, 1997.

[221] M. Schneider. Automotive Radar – Status and Trends. In *Proc. German Microwave Conference*, pages 144–147, 2005.

[222] H. Schöpp, A. Stiegler, T. May, M. Paintner, J. Massanell, and B. Buxbaum. 3D-PMD Kamerasysteme zur Erfassung des Fahrzeugumfelds und zur Überwachung des Fahrzeug-Innenraums. In *VDI Tagung Elektronik im Kraftfahrzeug, VDI-Berichte Nr. 2000*, 2007.

[223] W. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325, 1955.

[224] R. Shaw. Minkowski space. In *Linear Algebra and Group Representations*, pages 221–242. Academic Press, 1982.

[225] A. Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–43, 2001.

[226] L. I. Smith. A tutorial on Principal Components Analysis. Technical notes, February 26, 2002, `http://nyx-www.informatik.uni-bremen.de/664/1/smith_tr_02.pdf`. [accessed 17-11-2013].

[227] P. Smith, M. Shah, and N. da Vitoria Lobo. Determining driver visual attention with one camera. *IEEE Transactions on Intelligent Transportation Systems*, 4(4):205–218, 2003.

[228] SoftKinetic. SoftKinetic DepthSense Cameras. `http://www.softkinetic.com/en-us/products/depthsensecameras.aspx`. [accessed 17-11-2013].

[229] J. Stallkamp, H. K. Ekenel, H. Erdogan, R. Stiefelhagen, and A. Ercil. Video-based driver identification using local appearance face recognition. In *Proc. Workshop on DSP in Mobile and Vehicular Systems*, 2007.

[230] T. Stevens. Toyota's new eyelid-monitoring system wakes up sleepy drivers. January 24, 2008, `http://www.switched.com/2008/01/24/upcoming-toyotas-will-watch-your-eyes/`. [accessed 17-11-2013].

[231] A. A. Thomas and M. Wilscy. Face recognition using Simplified Fuzzy ARTMAP. *Int. Journal on Signal and Image Processing (SIPIJ)*, 1(2):134–146, 2010.

[232] C. Tinto. Toyota's approach toward the realization of sustainable mobility. The 2008 Toyota Sustainable Mobility Seminar, `http://www.fuelsandenergy.com/presentations/TintoSustainableMobilitySeminar.pdf`. [accessed 17-11-2013].

[233] M. M. Trivedi, S. Y. Cheng, E. Childers, and S. Krotosky. Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. *IEEE Trans. Veh. Technol., Special Issue on In-Vehicle Vision Systems*, 53(6):1698–1712, 2004.

[234] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[235] H. Ueno, M. Kaneda, and M. Tsukino. Development of drowsiness detection system. In *Proc. Vehicle Navigation and Information Systems Conference*, pages 15–20, 1994.

[236] UniversityOfMassachusetts. Labeled Faces in the Wild. `http://vis-www.cs.umass.edu/lfw/results.html`. [accessed 17-11-2013].

[237] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[238] W. Vanlaar, H. Simpson, D. Mayhew, and R. Robertson. Fatigued and drowsy driving: Attitudes, concern and practices of Ontario drivers. Technical report, Traffic Injury Research Foundation, 2007.

[239] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[240] V. Vapnik. *Statistical Learning Theory.* John Wiley, New York, 1998.

[241] A. Verri. Machine learning techniques in biometrics. In *Int. Summer School for Advanced Studies on Biometrics for Secure Authentication, New Technologies and Embedded System*, Alghero, Italy, June 11-15 2007.

[242] P. Viola and M. J. Jones. Robust real-time face detection. *Int. Journal of Computer Vision*, 57(2):137–154, 2004.

[243] VolkswagenAG. Biometric driver identification: Greater safety and comfort, better anti-theft protection. `http://www.volkswagenag.com/content/vwcorp/content/en/innovation/communication_and_networking/Biometric.html`. [accessed 17-11-2013].

[244] Volvo. Volvo Driver Alert Control and Lane Departure Warning. August 28, 2007, `http://www.zercustoms.com/news/Volvo-Driver-Alert-Control-and-Lane-Departure-Warning.html`. [accessed 17-11-2013].

[245] T. von Jan, T. Karnahl, K. Seifert, J. Hilgenstock, and R. Zobel. Don't sleep and drive - VW's fatigue detection technology. In *Proc. 19th International Technical Conference on the Enhanced Safety of Vehicles*, 2005.

[246] J. L. Wayman. Fundamentals of biometric authentication technologies. In *National Biometric Test Center - Collected Works - 1997-2000*, pages 1–19. San Jose State University, 2000.

[247] J. A. Webb. Global image processing operations on parallel architectures. In *Proc. SPIE 1295, Real-Time Image Processing II*, pages 176–187, 1990.

[248] Weka. Weka 3: Data mining software in Java. `http://www.cs.waikato.ac.nz/ml/weka/`. [accessed 17-11-2013].

[249] S. Wender and O. Loehlein. Multiple classifier cascades for vehicle occupant monitoring using an omni-directional camera. In *Proc. 3rd Workshop on Self-Organization of Adaptive behavior*, pages 37–46, 2004.

[250] E. S. Wentzel. *Probability Theory.* High School, Moscow, 10th edition, 2006. [in russian].

[251] J. Wikander. Automated vehicle occupancy technologies study: Synthesis report. Technical Report FHWA-HOP-07-134, Prepared for the HOV Pooled-Fund Study of the U.S. Department of Transportation Federal Highway Administration by Texas Transportation Institute, 2007.

[252] J. R. Williamson. Gaussian ARTMAP: A neural network for fast incremental learning of noisy multidimensional maps. Technical Report CAS/CNS-95-003, Boston University, 1995.

[253] J. Wu and M. M. Trivedi. Robust facial landmark detection for intelligent vehicle system. In *Proc. IEEE Int. Workshop on Analysis and Modelling of Faces and Gestures*, pages 213–228, 2005.

[254] X. Wu and other. Top 10 algorithms in data mining, knowledge and information systems. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[255] W.Wang and H. Qin. A FPGA based driver drowsiness detecting system. In *Proc. IEEE Int. Conference on Vehicular Electronics and Safety*, pages 358–363, 2005.

[256] M.-H. Yang, D. Roth, and N. Ahuja. A SNoW-based face detector. In *Advances in Neural Information Processing Systems 12 (NIPS)*, pages 855–861. MIT Press, 2000.

[257] W. Zhao and R. Chellappa. Symmetric shape-from-shading using self-ratio image. *Int. Journal of Computer Vision*, 45(1):55–75, 2001.

[258] D. Zittlau, S. Boverie, and P. Mengel. Kameraanwendung im Fahrzeuginnenraum. In *Elektronik im Kraftfahrzeug, VDI-Berichte Nr. 1547*, pages 1065–1082, 2000.

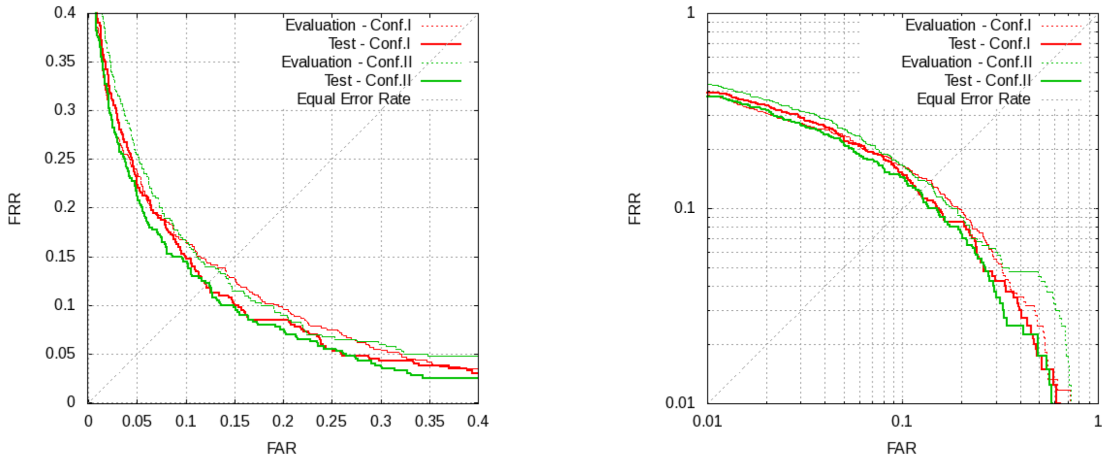
[259] M. Zollhöfer, M. Martinek, G. Greiner, M. Stamminger, and J. Süßmuth. Automatic reconstruction of personalized avatars from 3D face scans. *Computer Animation and Virtual Worlds*, 22(2-3):195–202, 2011.

Appendix A

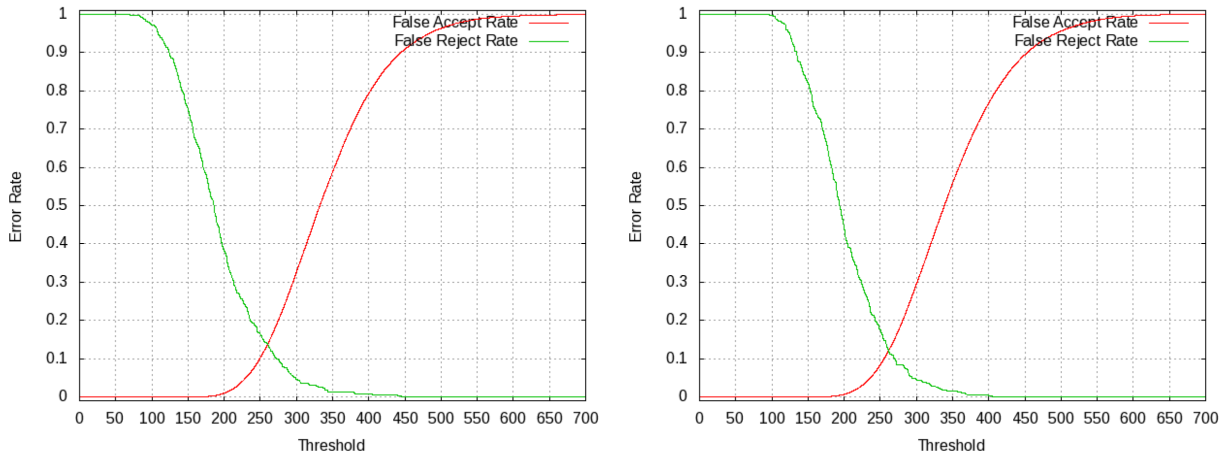
Appendix A: Experiments with XM2VTS

A.1 FaceART

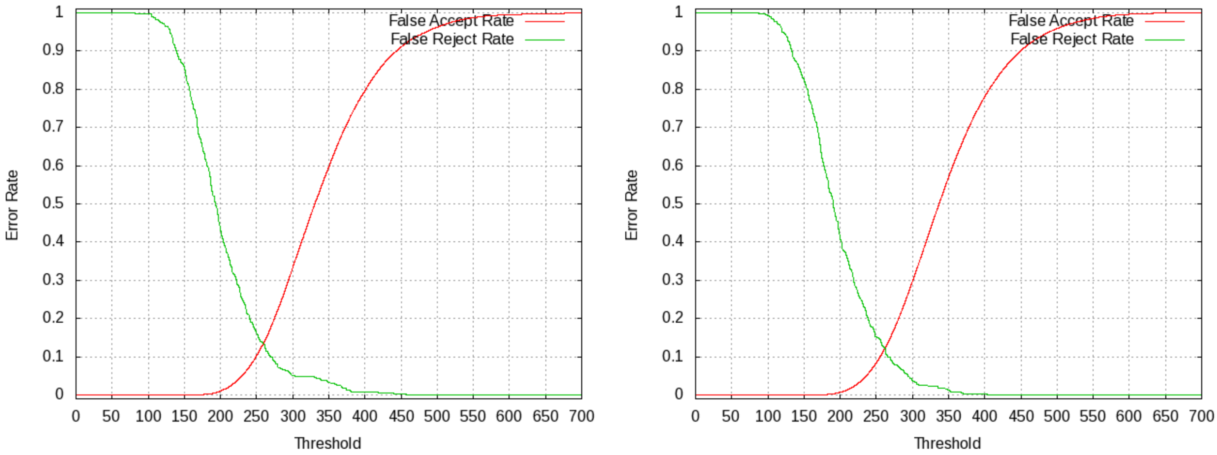
FaceART, Mode 1, ROC (left) and logarithmic ROC (right) curves



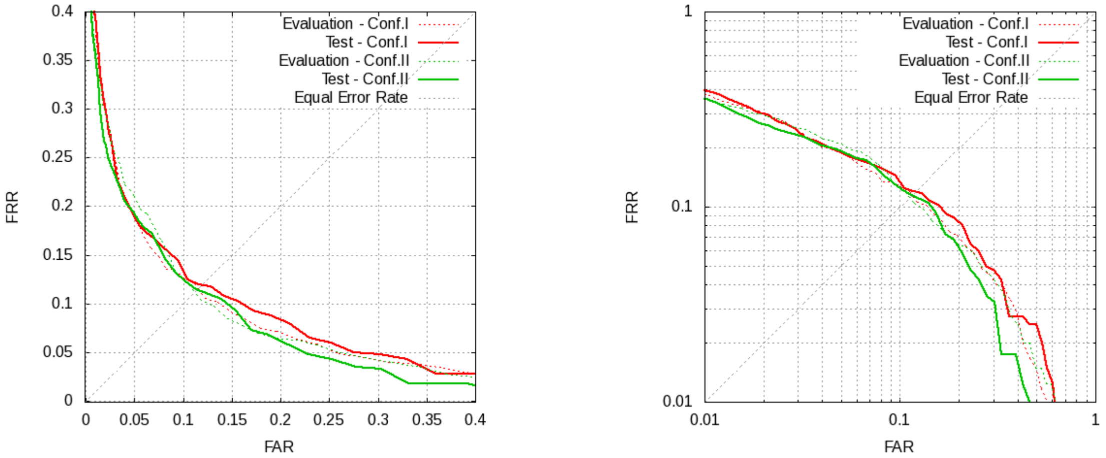
FaceART, Mode 1, 1^{st} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes



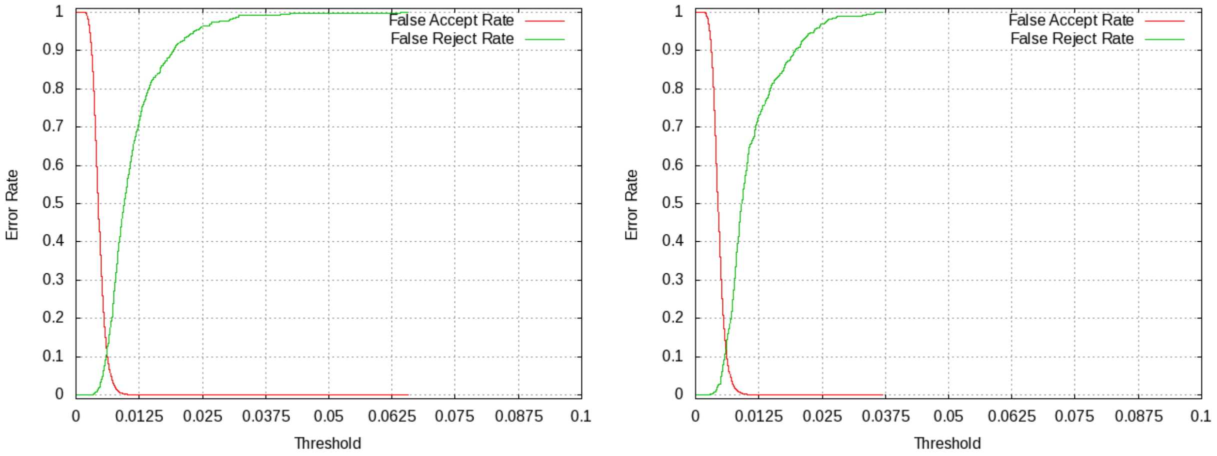
FaceART, Mode 1, 2^{nd} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes



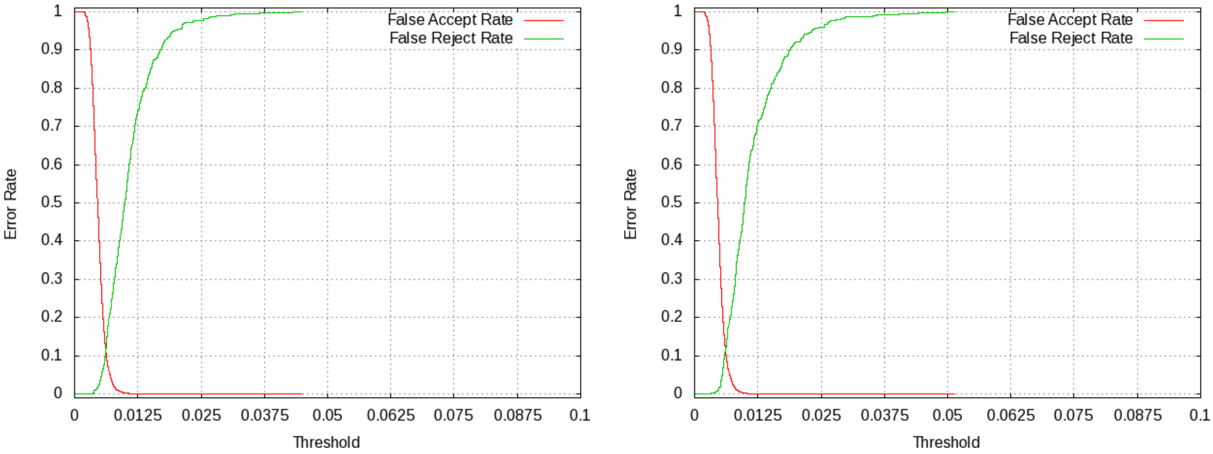
FaceART, Mode 2, ROC (left) and logarithmic ROC (right) curves



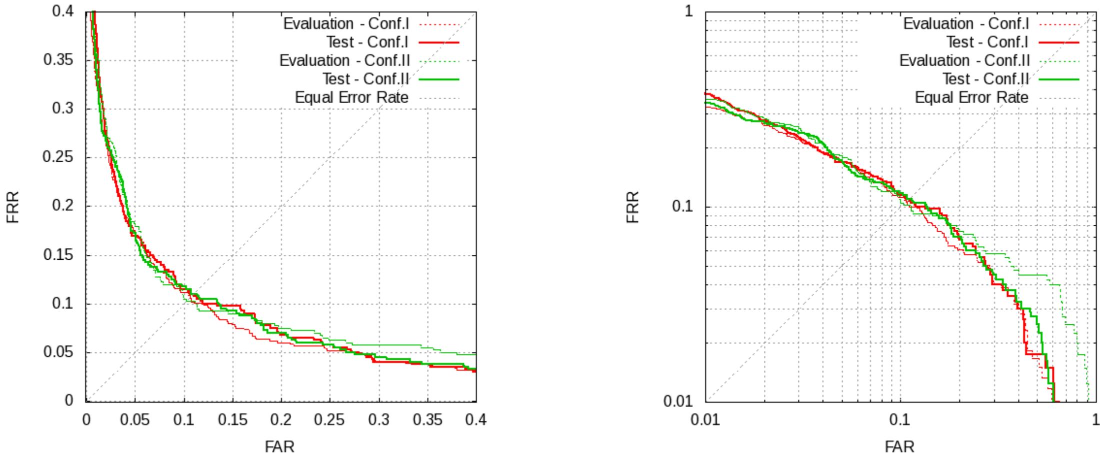
FaceART, Mode 2, 1^{st} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes



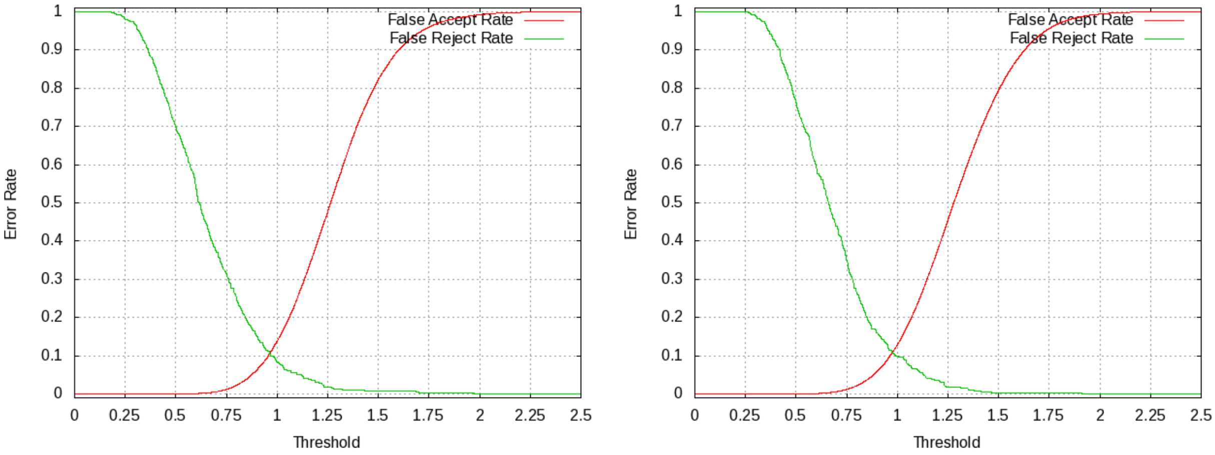
FaceART, Mode 2, 2^{nd} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes



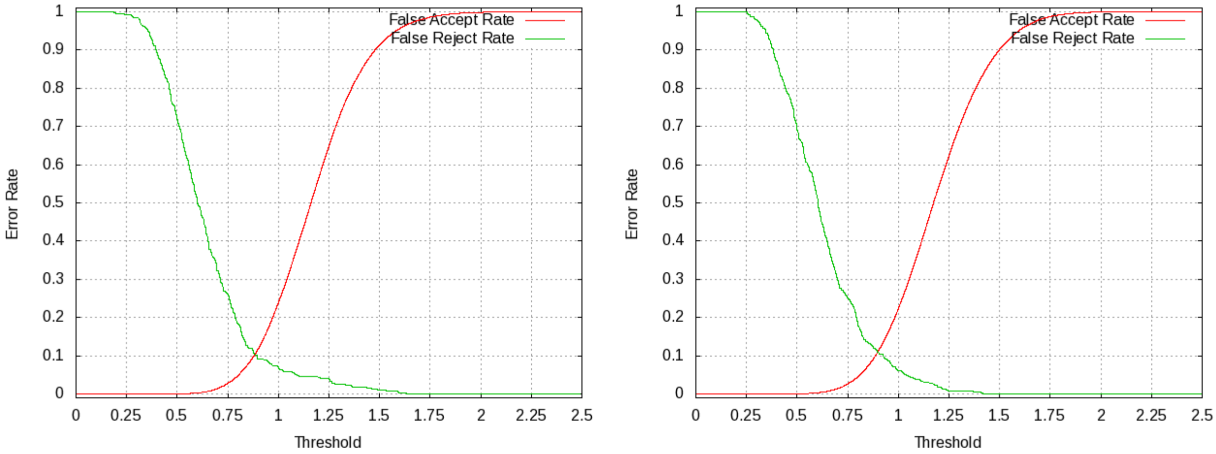
FaceART, Mode 3, ROC (left) and logarithmic ROC (right) curves



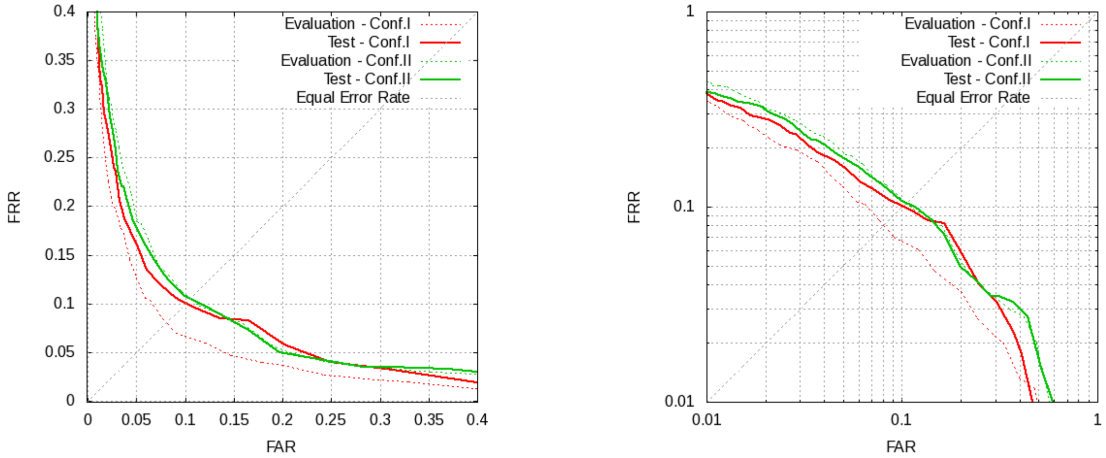
FaceART, Mode 3, 1^{st} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes



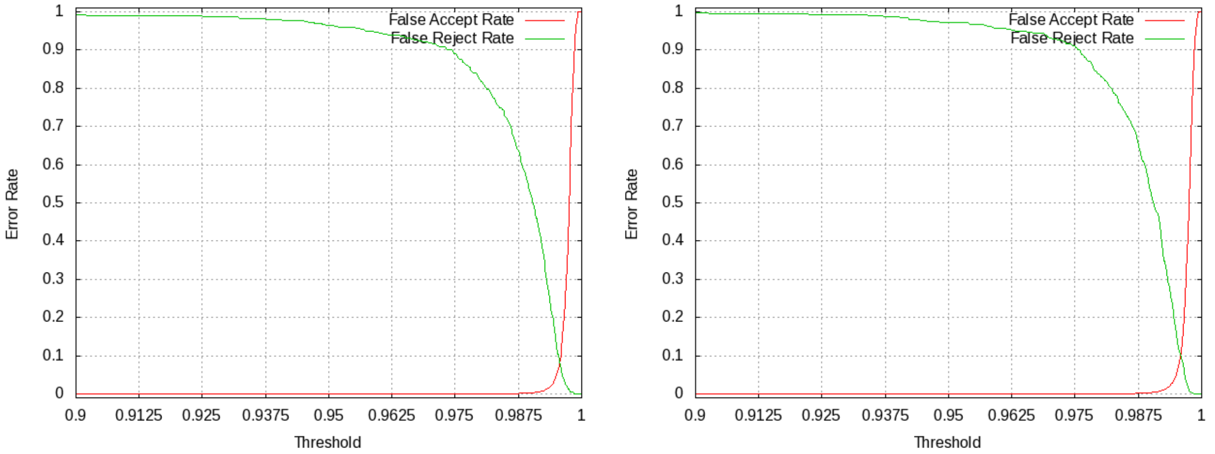
FaceART, Mode 3, 2^{nd} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes



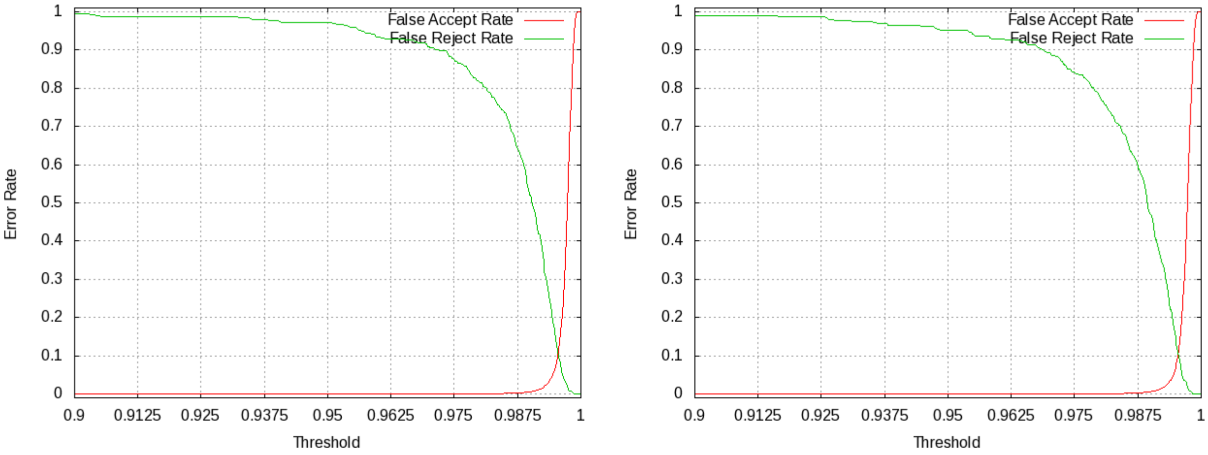
FaceART, Mode 4, ROC (left) and logarithmic ROC (right) curves



FaceART, Mode 4, 1^{st} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes

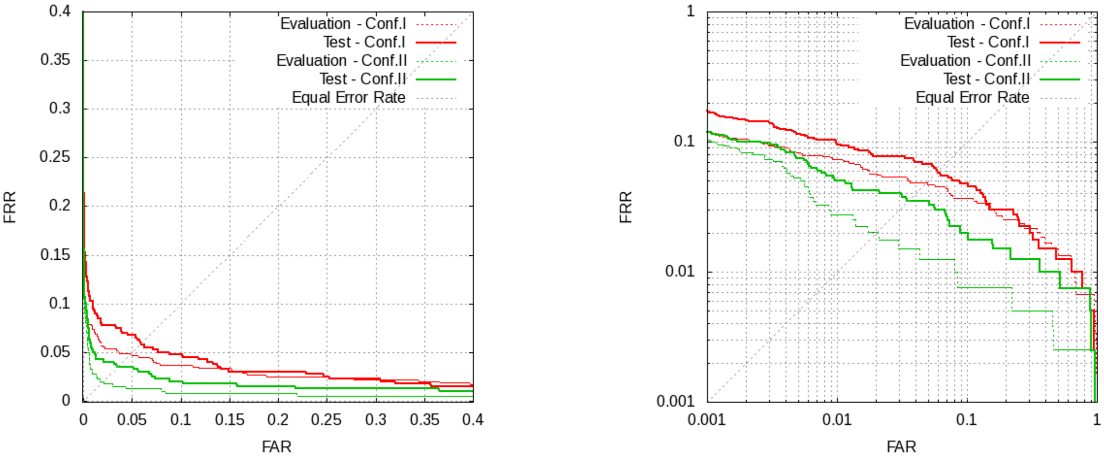


FaceART, Mode 4, 2^{nd} configuration of the Lausanne protocol, FAR/FRR curves in evaluation (left) and testing (right) modes

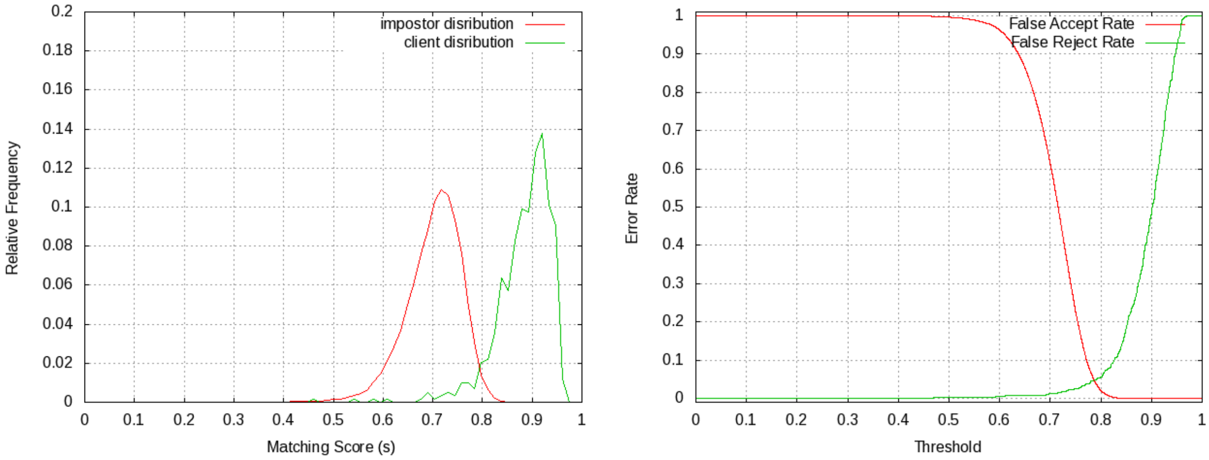


A.2 Luxand FaceSDK

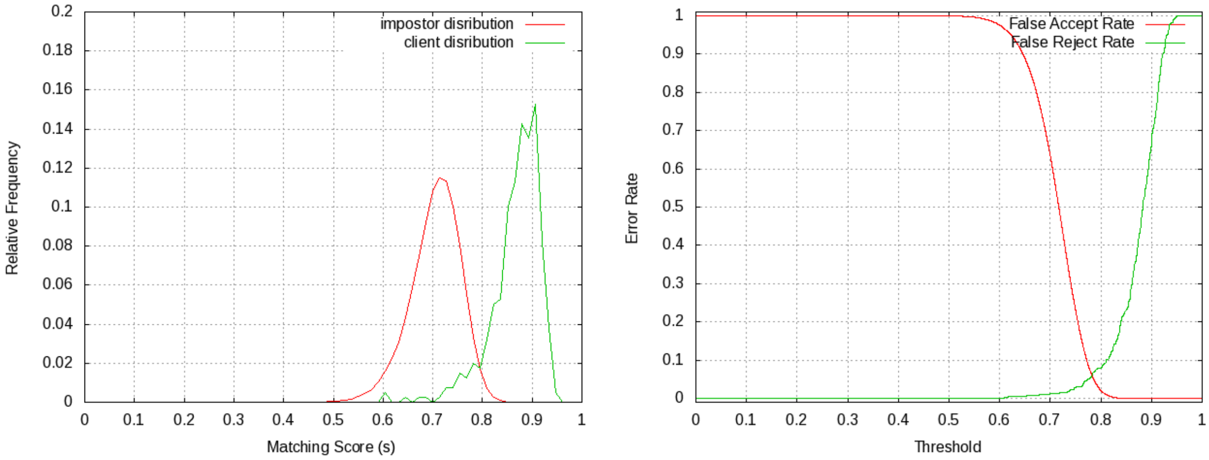
Luxand FaceSDK, ROC (left) and logarithmic ROC (right) curves



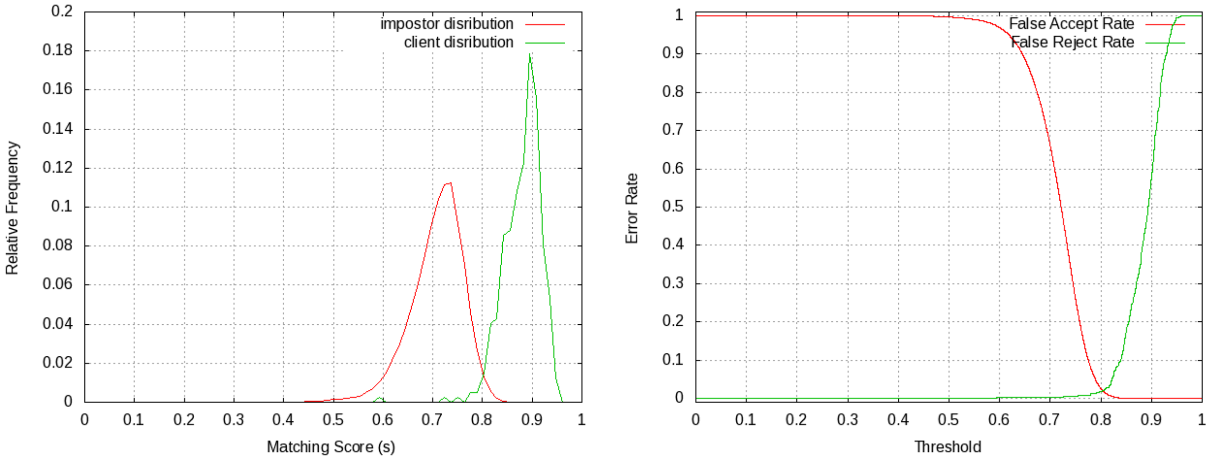
Luxand FaceSDK, 1^{st} configuration of the Lausanne protocol, evaluation mode, distributions of genuine and impostor scores (left) along with FAR/FRR curves (right)



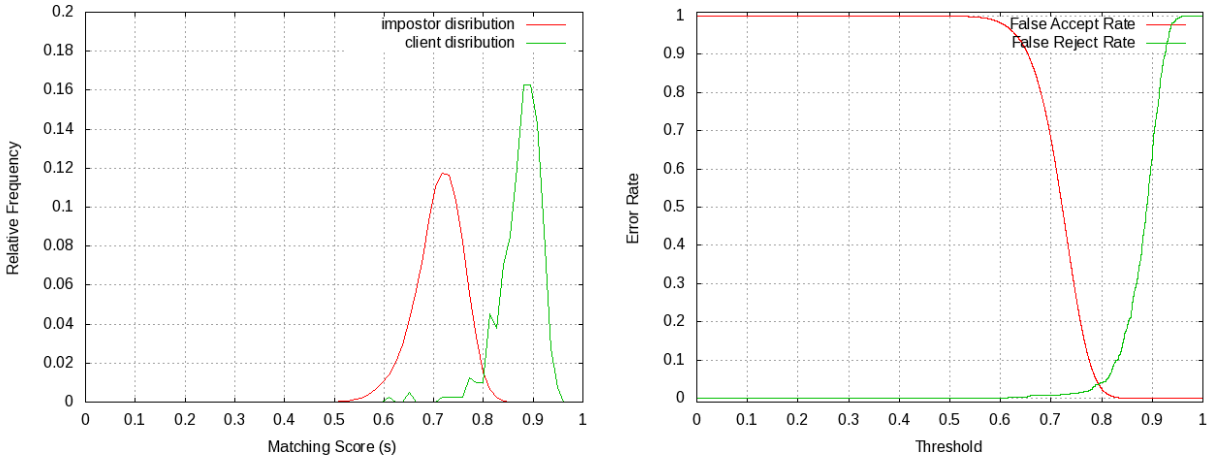
Luxand FaceSDK, 1^{st} configuration of the Lausanne protocol, testing mode, distributions of genuine and impostor scores (left) along with FAR/FRR curves (right)



Luxand FaceSDK, 2^{nd} configuration of the Lausanne protocol, evaluation mode, distributions of genuine and impostor scores (left) along with FAR/FRR curves (right)



Luxand FaceSDK, 2^{nd} configuration of the Lausanne protocol, testing mode, distributions of genuine and impostor scores (left) along with FAR/FRR curves (right)

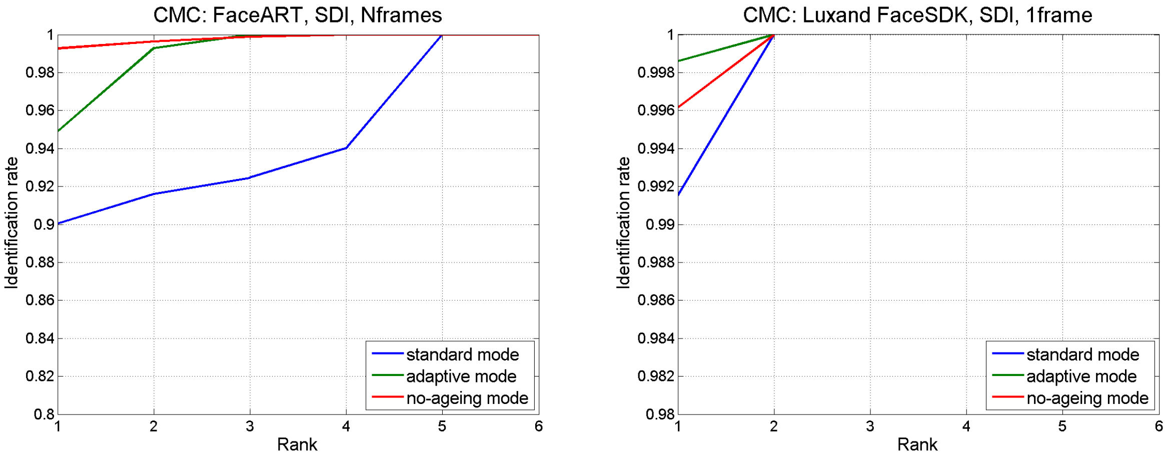


Appendix B

B Appendix B: Experiments with Opel Databases

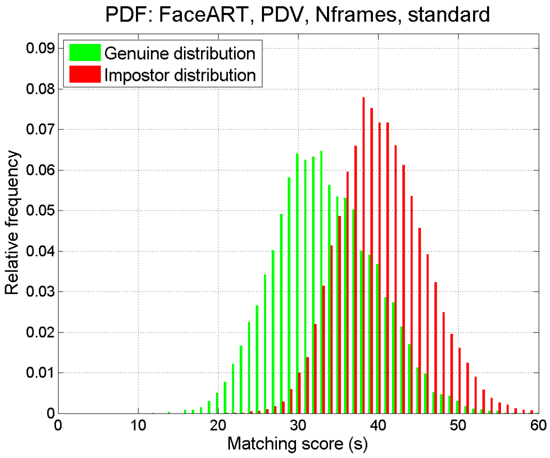
B.1 Part 1: OpelDB2

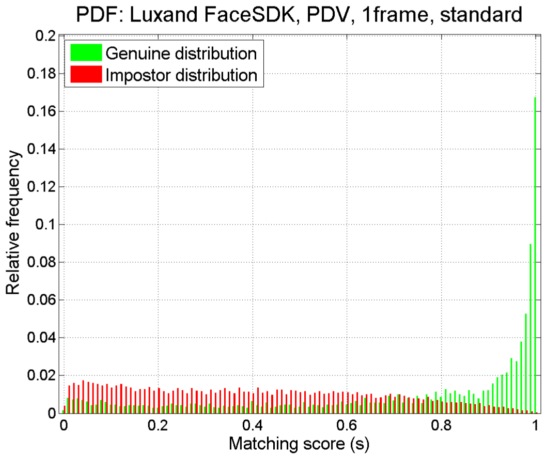
CMC curves resulting from Single Driver Identification (SDI) with FaceART (left) and Luxand FaceSDK (right)



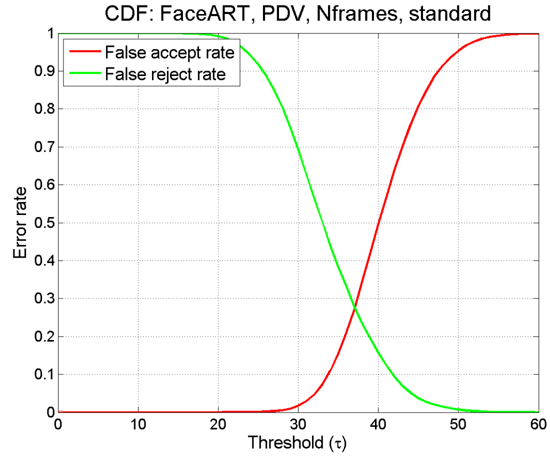
Permanent Driver Verification (PDV) with FaceART (left) and Luxand FaceSDK (right)

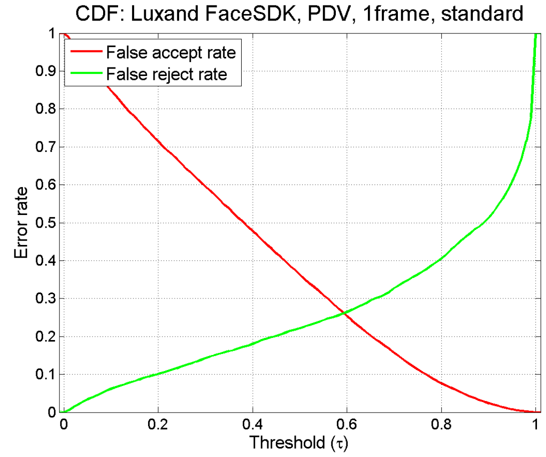
Standard template preservation mode: PDF



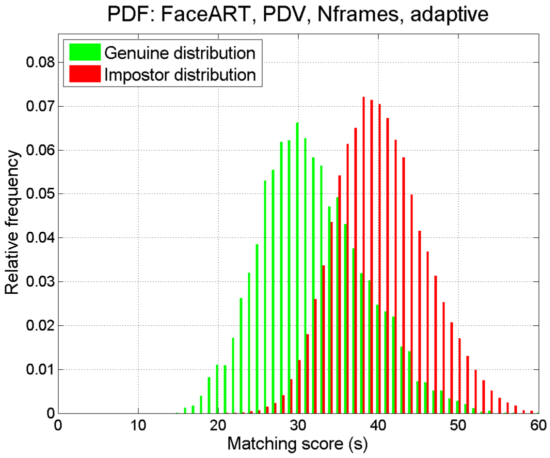


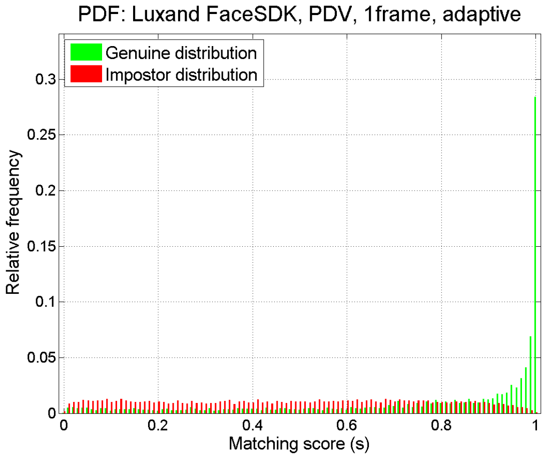
Standard template preservation mode: CDF



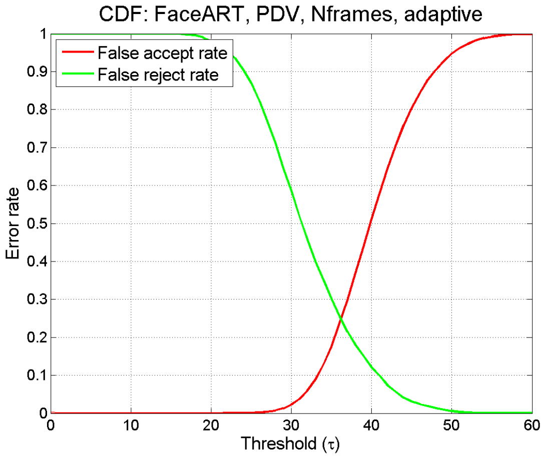


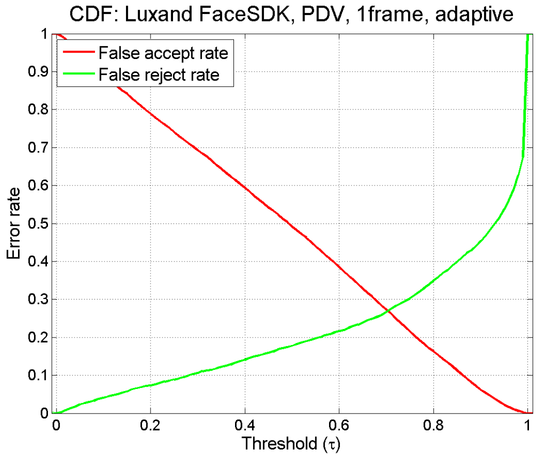
Adaptive template preservation mode: PDF



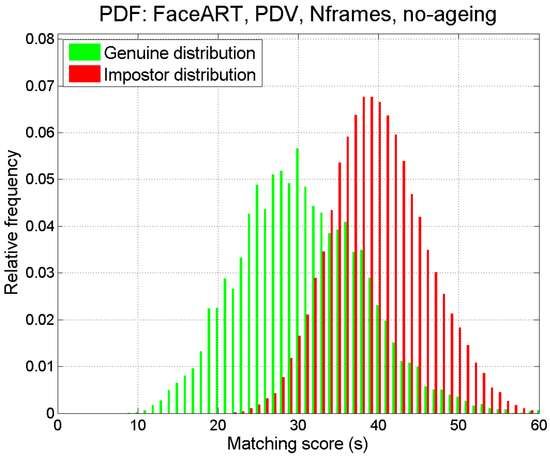


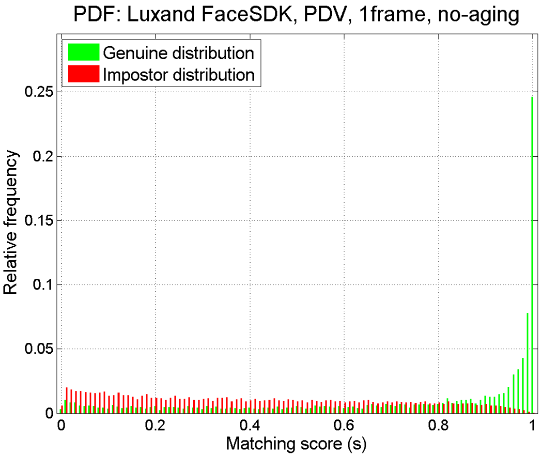
Adaptive template preservation mode: CDF



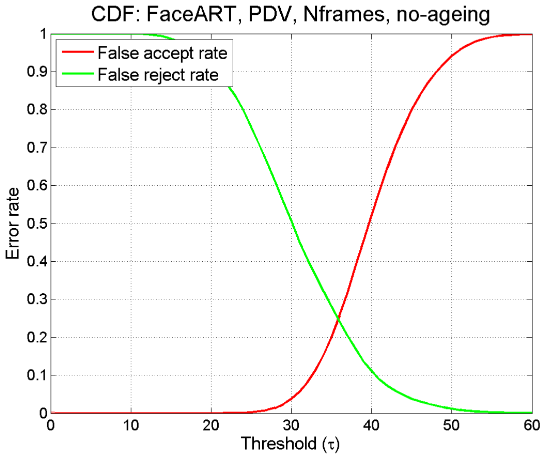


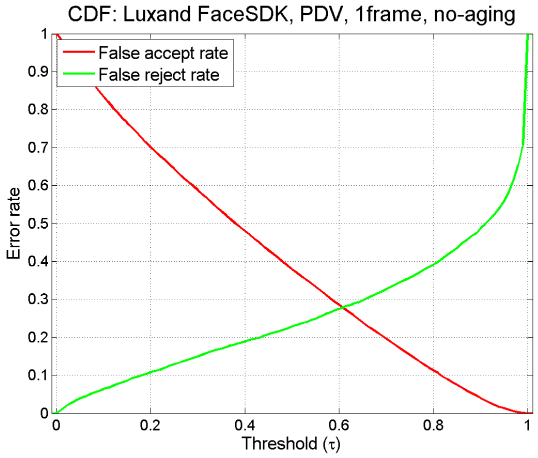
No-aging template preservation mode: PDF





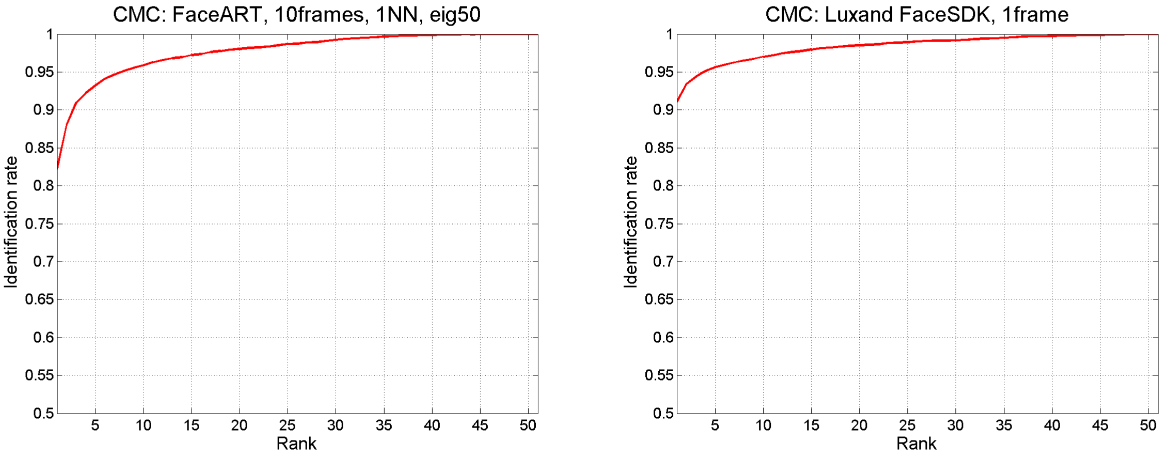
No-aging template preservation mode: CDF





B.2 Part 2: OpelDB3

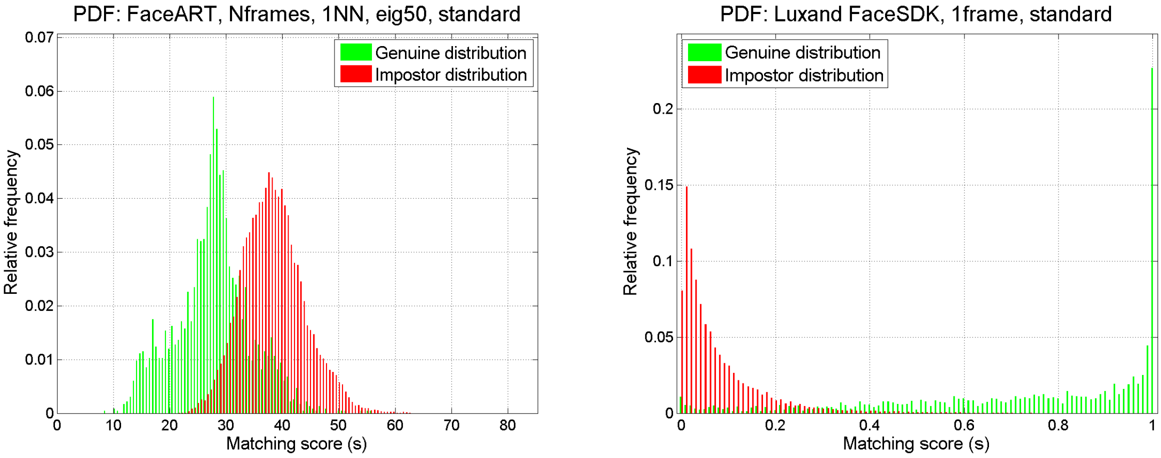
CMC curves resulting from Single Driver Identification (SDI) with FaceART (left) and Luxand FaceSDK (right)



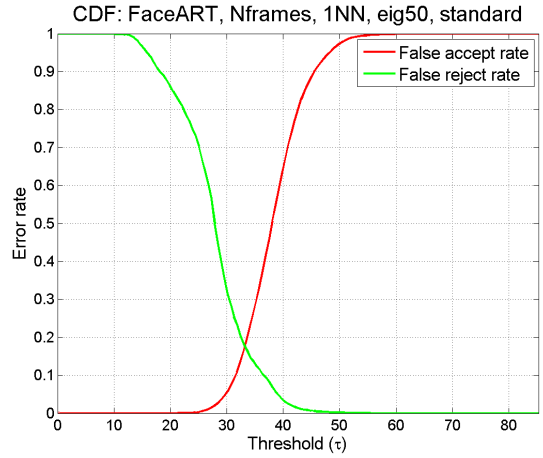
B.3 Part 3: OpelDB2+3

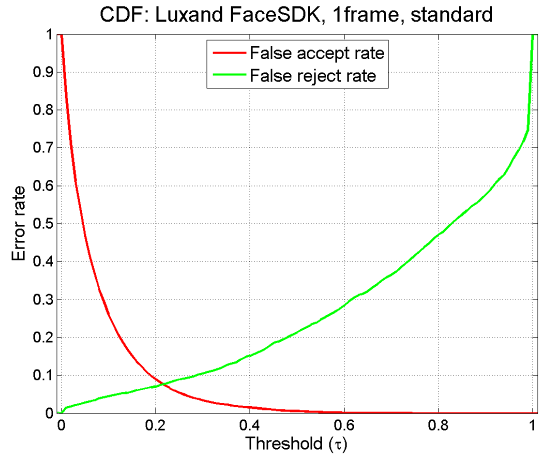
PDF and CDF curves are used for threshold estimation to support Single Driver Identification (SDI). Be advised that an identification trial is reckoned to be successful when the true identity is on the first position in the rank-list and the similarity/dissimilarity score is higher/lower than the threshold. Left - FaceART, right - Luxand FaceSDK.

Standard template preservation mode: PDF

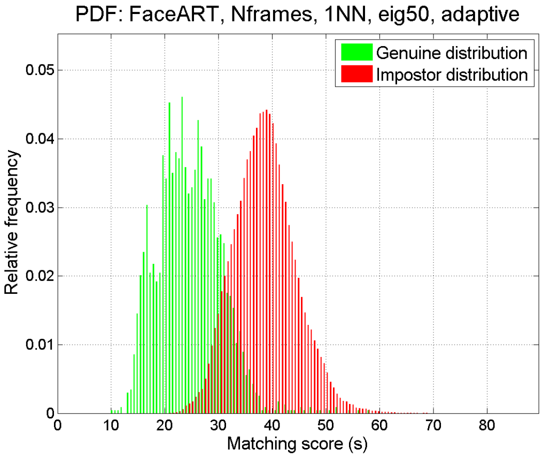


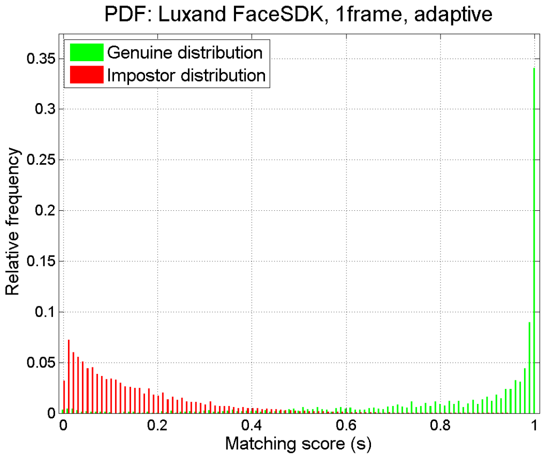
Standard template preservation mode: CDF



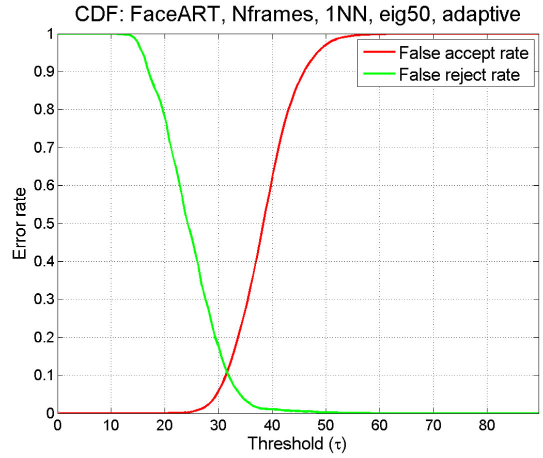


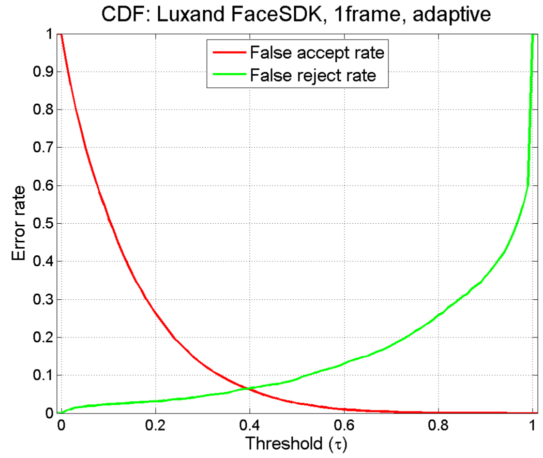
Adaptive template preservation mode: PDF



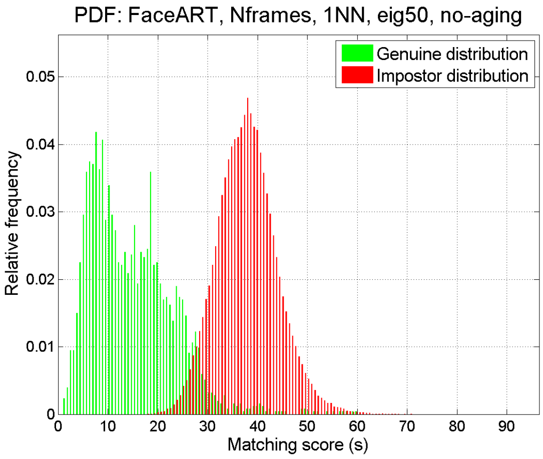


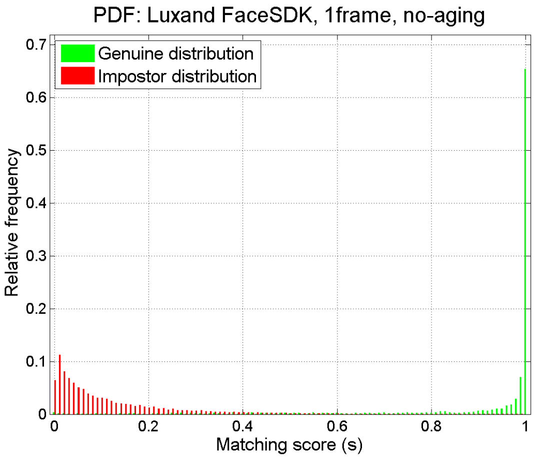
Adaptive template preservation mode: CDF



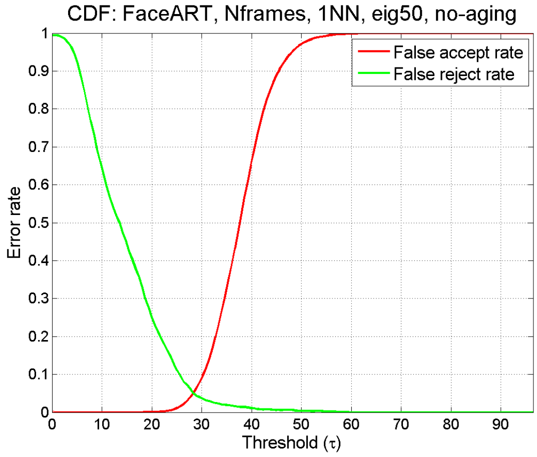


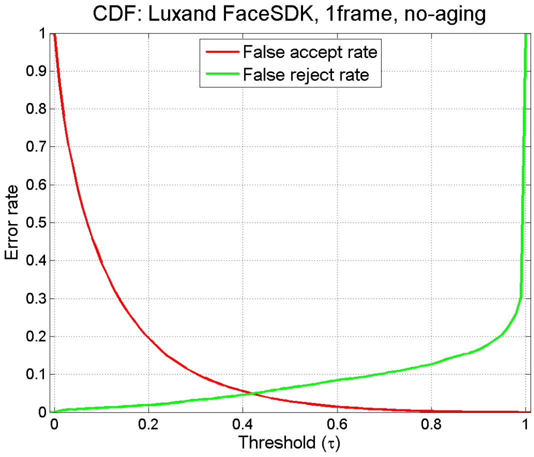
No-aging template preservation mode: PDF





No-aging template preservation mode: CDF


CDF: FaceART, Nframes, 1NN, eig50, no-aging
False accept rate
False reject rate
Error rate
Threshold (τ)


CDF: Luxand FaceSDK, 1frame, no-aging
False accept rate
False reject rate
Error rate
Threshold (τ)