FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Thermal imaging for vehicle occupant monitoring

Gustavo Rocha da Silva

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Professor Luís Corte-Real

Second Supervisor: Professor Pedro Carvalho

Industry Supervisor: André Ferreira

2018-07-25

Thermal imaging for vehicle occupant monitoring

Gustavo Rocha da Silva

Mestrado Integrado em Engenharia Informática e Computação

2018-07-25

Abstract

The automotive industry is pursuing the path of automation in their vehicles, and perceiving the environment of a vehicle requires the ability to detect, recognize and identify objects and events. This includes not only the analysis of the vehicle surroundings, but also of the inside of the cabin. When captured, this information is relevant for defining intelligent responses to events occurring in vehicle surroundings and interior.

Previous works have applied thermal imaging to detect and perform a series of analysis on the human face including landmark detection, emotion recognition, face recognition, respiratory rate measurement, stress levels, heart rate estimation and drunkenness detection, yet there is not much research on how those algorithms behave for vehicle interior monitoring purposes.

This work focuses on the analysis of the interior of vehicles using thermal cameras combined with RGB cameras. Such analysis includes the definition, implementation and comparison of different algorithms. Using far-infrared imagery for this purpose has some advantages against RGB images, which require sun or artificial lighting and are prone to light variations. However, it also presents some disadvantages, such as the lack of high-frequency information and the higher cost of the cameras, so this work studies the applicability of infrared imagery for monitoring vehicle occupants.

More specifically, the challenges tackled in this work are focused in face detection, smoking detection, respiratory rate estimation and emotion recognition. The results are evaluated by using public databases and a custom-built dataset specific for this work. Our face detection algorithm scores 99.60% in AP_{50} and 79.47% in AP in our database. Smoking activities are detected using temporal information on the location of hot spots. Respiratory rate estimation is accomplished by measuring the mean temperature of the philtrum region through time. For emotion recognition, we created a convolutional neural network for valence prediction in thermal images that has an accuracy of 80.1% in NVIE spontaneous database and of 67.8% in our database. We also created an ensemble using RGB and thermal images and compared the results.

Resumo

A indústria automóvel está a caminhar no sentido de aumentar a automação nos seus veículos, e compreender o ambiente de um veículo implica a capacidade de detetar, reconhecer e identificar objetos e eventos. Isto inclui não só uma análise do exterior do veículo, mas também do seu interior. Ao ser capturada, esta informação é importante para definir respostas inteligentes aos eventos que aconteçam no exterior e interior do veículo.

Trabalhos anteriores aplicaram imagens térmicas para detetar e efetuar diferentes tipos de análise da face humana, incluindo deteção de pontos fiduciais, reconhecimento de emoções, identificação de pessoas, medição da frequência respiratória, níveis de stresse, estimação do batimento cardíaco e deteção de estado de embriaguez, no entanto não existe muita investigação em como esses algoritmos se comportam num cenário de monitorização do interior de veículos.

Este trabalho foca-se na análise do interior de veículos usando câmaras térmicas combinadas com câmaras RGB. Esta análise inclui a definição, implementação e comparação de diferentes algoritmos. A utilização de imagens térmicas com este propósito tem algumas vantagens em relação às imagens RGB, que requerem luz solar ou artificial e estão sujeitas a variações de luz. No entanto, também apresentam desvantagens, como a falta de informação de alta-frequência e o custo mais elevado das câmaras, portanto este trabalho estuda a aplicabilidade de imagens infravermelhas para a monitorização de ocupantes de um veículo.

Mais especificamente, os desafios tratados neste trabalho estão focados na deteção de faces, deteção de fumadores, estimação da frequência respiratória e reconhecimento de emoções. Os resultados são avaliados usando bases de dados públicas e uma base de dados construída especificamente para este trabalho. O nosso algoritmo de deteção facial atinge 99.60% em AP_{50} e 79.47% em AP na nossa base de dados. Atividades de fumar são detetadas usando informação temporal sobre a localização de pontos quentes. A estimação da frequência respiratória é realizada medindo a temperatura média da região entre o nariz e a boca ao longo do tempo. Para o reconhecimento de emoções, criámos uma rede neuronal convolucional para previsão da valência emocional com imagens térmicas atingindo uma precisão de 80.1% em expressões espontâneas da base de dados NVIE e de 67.8% na nossa base de dados. Desenvolvemos também um algoritmo que junta a informação RGB com a térmica e comparámos os resultados.

Acknowledgements

I am deeply grateful to my thesis supervisors, Professor Luís Corte-Real and Professor Pedro Carvalho, who helped me with their guidance to complete my dissertation.

I would also like to thank my colleagues during my internship at Bosch for their help, namely the other interns, Rui, Guilherme, João and Mariana. I would like to single out Filipa Ramos, for always being ready to help me both professionally and personally. I am also grateful for the help of other collegues, especially Joaquim Fonseca for his valuable support and my institutional coordinator, André Ferreira, for his excellent guidance and for always ensuring I had all conditions necessary to finish my work.

Finally, I thank my friends and my family, who helped me throughout my life.

Gustavo Silva

“When you can’t make them see the light, make them feel the heat.“

Ronald Reagan

Contents

List of Figures

List of Tables

Abbreviations

3D	Three-dimensional
A/C	Air conditioning
AP	Average Precision (same as mAP)
AUC	Area Under the Curve
bpp	Bits per pixel
CNN	Convolutional Neural Network
COCO	Common Objects in COntext
CPU	Central Processing Unit
EER	Equal Error Rate
EVM	Eulerian Video Magnification
FDDB	Face Detection Data set and Benchmark
FER	Facial Expression Recognition
FFC	Flat-field correction
FFT	Fast-Fourier Transform
FIR	Far infrared
FOV	Field-of-view
FPS	Frames-per-second
GDPR	General Data Protection Regulation
GPU	Graphical Processing Unit
HFOV	Horizontal field-of-view
HOG	Histogram of Gradients
IoU	Intersection over Union
IR	Infrared
KCF	Kernelized Correlation Filters
KTFE	Kotani Thermal Facial Emotion
LBP	Local Binary Patterns
LSTM	Long short-term memory
LWIR	Long-wavelength infrared
mAP	Mean Average Precision (same as AP)
MIT	Massachusetts Institute of Technology
MSE	Mean Squared Error
MWIR	Mid-wavelength infrared
NIR	Near-infrared
NVIE	Natural Visible and Infrared facial Expression Database
PPM	Parts-per-million
PSD	Power Spectral Density
R-CNN	Regions with Convolutional Neural Networks features
R-FCN	Region-based Fully Convolutional Networks

ABBREVIATIONS

RGB	Red-green-blue
RGB-D-T	Red-green-blue, Depth and Thermal
ROC	Receiver operating characteristic
RoI	Region of Interest
RVS	Respiration variability spectrogram
SSD	Single Shot Detector
SVM	Support vector machine
SWIR	Short-wavelength infrared
T	Thermal
TNR	True Negative Rate
TPR	True Positive Rate
UAV	Unmanned aerial vehicle
USB	Universal Serial Bus
VFOV	Vertical field-of-view
VOC	Visual Object Classes
YOLO	You Only Look Once

Chapter 1

Introduction

1.1 Context and Motivation

The automotive industry is pursing the path to achieve fully autonomous vehicles. This can happen at multiple levels, namely driving automation, passenger monitoring, entertainment and safety improvement.

In the context of self-driving vehicles, passenger transportation companies have the opportunity to provide a transportation service to their clients. Unlike traditional transportation services, there is no driver on-board to take the passengers to their destination and to act as a representative of the company during the service. Such driver is also responsible for monitoring the behavior of the clients and ensuring they are satisfied with the ride, feeling well and complying with the rules of the service, for example, in situations where passengers are smoking or feeling angry. For this reason, there is the need of coming up with alternative ways of monitoring passengers using automation, to assure users are being monitored while keeping the monetary advantage for transportation companies of not requiring a driver in their service.

Most of the automation in vehicles takes advantage of a combination of sensorial data that feeds the on-board computer with essential information to achieve that automation. The sensors used for this purpose can also be optical cameras. RGB cameras are a common choice, because they capture visible light, usually have high resolution and low cost. However, cameras that operate in different wavebands can also help automation. In fact, with the decreasing cost of infrared cameras in the last years, the idea of using them in the context of inner-vehicle monitoring becomes attractive. As further described in Chapter 2, different types of cameras have advantages and disadvantages when applied to the problem of this dissertation and it is possible to combine information from those different types to collect more information of the scene.

The work in this dissertation is part of the efforts made by Bosch Car Multimedia Portugal, S. A. in researching innovative ways of applying infrared cameras in their products. Bosch Car Multimedia is a division of Bosch responsible for developing intelligent embedded solutions for entertainment, navigation, telematics and driving assistance.

This dissertation is divided in two parts. First, an overview of applications based on thermal cameras for monitoring the interior of vehicles is presented together with information on previous work that tackled similar challenges. This study is critical to narrow down the scope of the proposed solution to some of the most relevant use cases. In the second part of this dissertation, the proposed solution is implemented and the results evaluated and compared. This includes understanding how the algorithms behave in our use case and how they can be adapted for robustness in the context of monitoring of a vehicle interior.

1.2 Objectives

The main driver of this dissertation is to explore algorithms to analyze the interior of vehicles using images captured by different types of cameras, with a special focus on the ones that operate in the far-infrared region of the electromagnetic spectrum.

In this context, a list of objectives were defined:

- Face detection
- Smoking detection
- Respiratory rate estimation
- Emotion recognition

Combining images of different modalities may increase the ability of sensing the scene, so this possibility is also subject to study.

1.3 Contributions

This dissertation has the following main contributions:

- Data capturing guide — a database containing video sequences from 38 subjects was created, and the capturing guide that details the process is part of this document.
- Face detection — an existing algorithm for general object detection in visible images is adapted for face detection in far-infrared images captured inside vehicles. High detection accuracies are reached and thermal imaging is validated as a reliable method for face detection that has the advantage of not being dependent on light illumination conditions.
- Facial landmark detection — an algorithm is developed to detect certain key points in thermal images of faces.
- Glasses detection — a new method is proposed for detecting if a person is wearing glasses, using a thermal image of the face.

- Smoking activity detection — a new algorithm for detecting smokers is presented, taking advantage of the heat emitted by smoking devices and common movement patterns during smoking. Its performance is evaluated with two different types of devices: cigarettes and heated tobacco.

- Respiratory rate estimation — this dissertation analyzes how existing algorithms for respiratory rate estimation can be adapted for vehicle occupant monitoring and how they behave in such scenario.

- Facial expression recognition — new algorithms are proposed that improve on previous work on valence prediction in thermal images. Furthermore, this work proposes an ensemble that combines visible with thermal information and compares the results between a public database and our custom-made dataset.

1.4 Structure of the Dissertation

After this introductory chapter, this document is structured in 4 more chapters.

Chapter 2 describes the state of the art and mentions previous related work, including a brief introduction on thermal imaging and available equipment. For some of the possible use cases, an analysis is made on existing solutions and how they can be improved and combined to meet the requirements. This analysis includes understanding the camera setup (location of the camera(s)), the algorithms involved and the eventual process of gathering data for training and testing.

Chapter 3 details the data acquisition process that was followed during the construction of a database specific for this dissertation.

Chapter 4 exposes and explains the implementation details of our solution.

Finally, chapter 5 presents the conclusions extracted from the work developed and suggestions for future work that could improve the results.

Chapter 2

State of the Art

This chapter summarizes infrared imagery and describes the previous work related to this dissertation. It informs the reader with what has been done previously to address the problem of monitoring of the inside of vehicles taking special advantage of cameras that work in the far-infrared region of the spectrum of light.

2.1 Thermal radiation

The infrared band can be divided into a number of regions as displayed in Fig. 2.1.

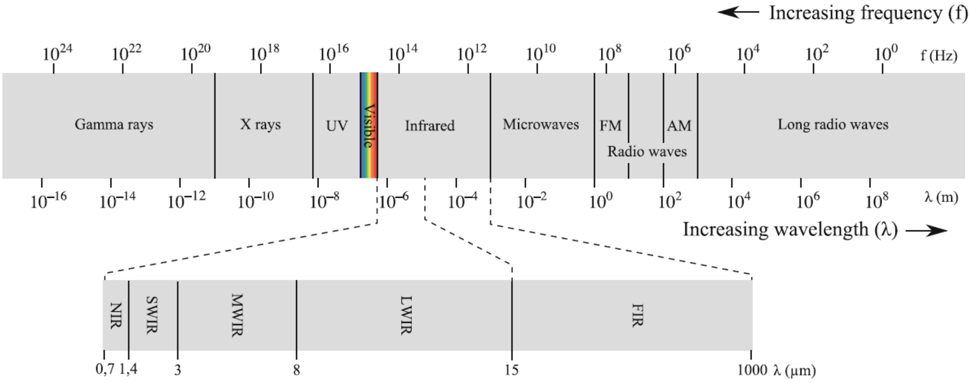

Increasing frequency (f)
10^{24} 10^{22} 10^{20} 10^{18} 10^{16} 10^{14} 10^{12} 10^{10} 10^{8} 10^{6} 10^{4} 10^{2} 10^{0} f (Hz)
Gamma rays
X rays
UV
Visible
Infrared
Microwaves
FM
AM
Radio waves
Long radio waves
10^{-16} 10^{-14} 10^{-12} 10^{-10} 10^{-8} 10^{-6} 10^{-4} 10^{-2} 10^{0} 10^{2} 10^{4} 10^{6} 10^{8} λ (m)
Increasing wavelength (λ)
NIR
SWIR
MWIR
LWIR
FIR
0,7 1,4 3 8 15 1000 λ (μm)

Figure 2.1: The electromagnetic spectrum with sub-divided infrared band. (Extracted from [GM14])

Thermal radiation is emitted in the largest wavelengths of the infrared spectrum by all objects with a temperature above absolute zero ($0K$). The radiation emitted by an object at temperature T is defined by Plank's wavelength distribution function [VKP+17]:

$$M_\lambda(T)d\lambda = \frac{2\pi hc^2}{\lambda^5}\frac{1}{\frac{hc}{e^{\lambda kT}} - 1} \tag{2.1}$$

where λ is the wavelength, T is the temperature of the object, $h = 6.626x10^{-34}J$ is the Planck's constant and $c = 2.998x10^{8}ms^{-1}$ is the speed of light in vacuum.

However, no real object can emit this maximum thermal radiation at a given temperature. Every object has a property, the emissivity ε, of value between 0 and 1, that should be multiplied by the blackbody[1] radiation to calculate the amount of radiation actually emitted from the surface. Most surfaces do not have an emissivity very close to 1, but human skin is an exception, where the value of e is around 98% and race-independent [JP02].

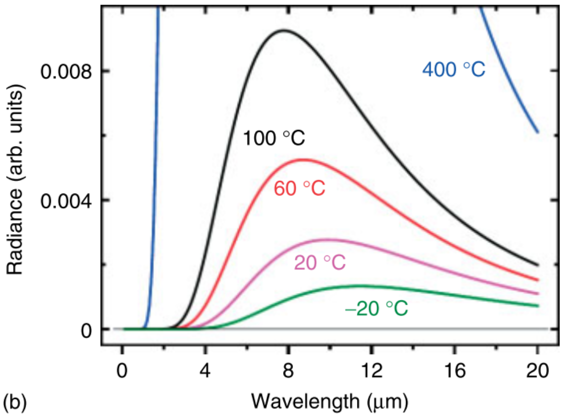


Figure 2.2: Radiance emitted by a blackbody at different temperatures. (Extracted from [VKP^{+}17])

2.2 Equipment analysis

Table 2.1 contains a non-exhaustive list of thermal cameras available in the market and their specifications. All the commercially-available cameras mentioned throughout this document are detailed in this table.

From the list, only the FLIR ONE models and the NEC R300 include both a thermal and a visible camera. The advantage of having both sensors in a single device instead of using two devices (one for each modality) is the synchronization, as this is handled by the hardware of the device. Another aspect to notice is the low resolution of those thermal cameras when compared to the common resolutions of modern RGB cameras. The same can be observed from the low FOV (Field-of-view), which can be more than two times wider in RGB cameras. The FLIR Lepton model has the special characteristic of being compact (8.5x8.5x5.6mm for the smallest model) so it is the technology behind some other cameras such as the FLIR ONE models. These camera models from FLIR are also one example of the decreasing cost of thermal cameras in general [Sou12] [Koc15] [Fre15] [GPVPW14]: the FLIR ONE Pro LT camera is priced at 299.99$ and the FLIR ONE Pro costs 399.99$.

[1] An idealized physical body that behaves as a perfect absorber and emitter.

Model	Res.	Waveband	Temp. Range	FPS	FOV	RGB res.
Argus P7225	320x240	8 – 14μm	-40^{o}C – 200^{o}C	60	24^{o}x18^o	-
AXIS Q1922	640x480	8 – 14μm	unspecified	30	57^{o}x44^o	-
FLIR A10	160x128	7.5 – 13.5μm	0 – 400^{o}C	30	40^{o}x30^o	-
FLIR A20M	320x240	7.5 – 13μm	-20^{o}C – 250^{o}C	60	25^{o}x19^o	-
FLIR A40M	320x240	7.5 – 13μm	-40^{o}C – 2000^{o}C	60	24^{o}x18^o	-
FLIR Lepton 25^o	80x60	8 – 14μm	0^{o}C – 120^{o}C	8.6	25^{o}x19^o	-
FLIR Lepton 50^o	80x60	8 – 14μm	0^{o}C – 120^{o}C	8.6	51^{o}x38^o	-
FLIR Lepton 3	160x120	8 – 14μm	0^{o}C – 450^{o}C	8.8	56^{o}x42^o	-
FLIR ONE Gen 2	160x120	8 – 14μm	-20^{o}C – 120^{o}C	8.7	46^{o}x35^o	640x480
FLIR ONE Gen 3	80x60	8 – 14μm	-20^{o}C – 120^{o}C	8.7	50^{o}x38^o	1440x1080
FLIR ONE Pro LT	80x60	8 – 14μm	-20^{o}C – 120^{o}C	8.7	55^{o}x43^o	1440x1080
FLIR ONE Pro	160x120	8 – 14μm	-20^{o}C – 400^{o}C	8.7	55^{o}x43^o	1440x1080
FLIR SC640	640x480	7.5 – 13μm	-40^{o}C – 1500^{o}C	30	45^{o}x34^o	-
NEC R300	320x240	8 – 14μm	-40^{o}C – 2000^{o}C	60	22^{o}x17^o	2048x1536
TESTO 880-3	160x120	8 – 14μm	-20^{o}C – 350^{o}C	9	32^{o}x24^o	-
VarioCAM hr head	1280x960	7.5 – 14μm	-40^{o}C – 2000^{o}C	60	90^{o}x74^o	-

Table 2.1: Non-exhaustive list of commercially-available thermal cameras. When multiple options are available for the same camera model, the best possible value for each characteristic has been chosen. In terms of FOV, the values reported are for the widest lenses.

2.3 Analyzing the vehicle

There has been previous research on how thermography can be useful to detect defects in certain vehicle components [BLA+13] [VKP+17]. As an example, Fig. 2.3 contains a thermal image of a rear-window defogger that is visibly malfunctioning, as the second highest wire and the three lower ones are not heating. In a similar way, it is also possible to analyze the effectiveness of front-window heaters, as well as seat heating systems, to understand if they are working as expected.

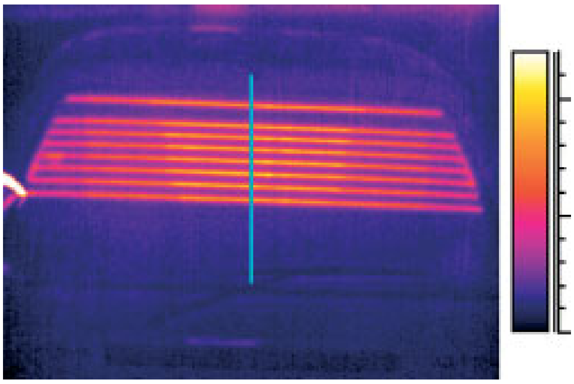


Figure 2.3: Defective defogger captured with a thermal camera. (Extracted from [VKP+17])

Other types of malfunctions can also be detected. For example, a car manufacturer has used thermal images to detect air leaks in vehicle cabins, by heating the air inside and then checking for the areas of lower temperature, from where the air is escaping [BLA+13]. It is also possible to detect initial stages of fires inside the vehicle, to allow for a quick reaction to the incident.

Thermal imaging is also a very effective method of measuring the temperatures inside a vehicle cabin to improve passenger comfort. In some previous works, the air fields inside the cabin have been measured with the help of a solid object placed inside it, but with high porosity, so as to reduce the impact on the air flow [BHP93] [Peš13]. However, this solution is very intrusive, so it cannot be used for the purpose of this dissertation, where we are looking for a passive approach. Instead, we will focus on the work that has been done to measure temperatures in the surfaces of the vehicle cabin, which are an indicator of the air temperature near them. One of these studies was conducted by Korukçu et al. [KK09], where thermal imaging was used to measure surface temperatures in conjunction with thermocouples that served as ground truth data for comparison. Their conclusion is that thermography produces identical temperature measurements. They have studied how the surface temperatures change over time during heating and cooling (Fig. 2.4), as well as the facial temperatures of vehicle occupant. Interestingly, during heating from 14ºC to 30ºC, there was an increase of 31ºC to 35ºC in facial temperature, but, on the other hand, there was almost no variation during cabin temperature cooling from 35ºC to 31.5ºC, although the temperature variation of the cooling test is smaller. The work performed by Korukçu et al. proves that thermal imaging can be a reliable asset for the analysis of temperatures inside a vehicle. This work was later extended in [KK12] to measure not only facial temperature, but also hand temperature, as those are the two parts of the body most commonly exposed during winter conditions.

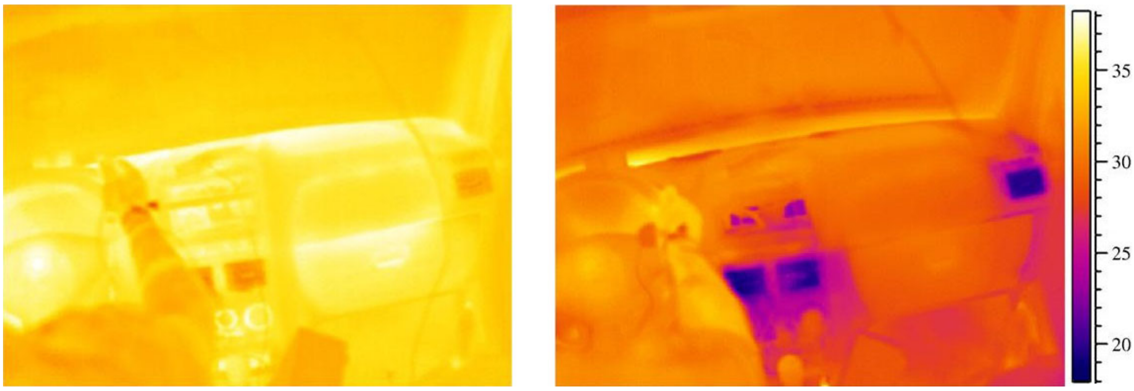


Figure 2.4: Analysis of the vehicle cabin temperatures during a cooling period of 30 minutes, as the temperature lowers from around 35ºC on the left image to 31.5ºC on the right one. (Adapted from [KK09]).

Another aspect that one might attempt to detect using thermal cameras is liquid spilling inside a vehicle. There has been previous successful research in using infrared cameras with a light source to detect and identify spilled liquids in a scene [HBD+08] [BES+11]. However, to the best of our knowledge, no prior work has been done in this topic using thermal imaging. There is successful research in detecting liquids being poured/spilled (still in the air, without having contacted the landing surface) [SF16] [YA16], using RGB cameras only. These studies have proven that using sequences of frames instead of single images results in a considerably better detection accuracy. In [SF16], experiments are conducted using a simple Convolutional Neural Network (CNN), a multi-frame CNN (fixed number of frames) and a CNN with a LSTM (Long short-term memory). Due to the difficulty and time-expensiveness of the task of labeling each pixel in video sequences of liquids being poured, data generation was achieved with a 3D physically accurate liquid simulator, pouring a liquid into a bowl while a video is playing behind the objects to avoid having a static background. The results show that a CNN+LSTM approach provides the best accuracy in liquid pouring detection, although no experiments have been conducted with a moving viewpoint. This approach, however, is not prepared to handle movement of the handler and does not detect viscous liquids. To tackle those cases, a new solution was developed that uses the Lucas-Kanade optical flow method [ZZF04] together with spacial and temporal filters [YA16].

More than monitoring the vehicle habitable itself, it is interesting to monitor the people inside it. Understanding who are the occupants of a vehicle and how they feel is a complex task. In order to do a proper analysis, usually there are two initial mandatory steps: face detection (locating faces in an image) and feature extraction (either manually or automatically with machine learning). The latter may require face registration beforehand, which is the process of locating fiducial points, such as the nose, mouth corners, eyes, cheek and other relevant points.

2.4 Multimodal sensor fusion

The solutions presented in this chapter until now use either RGB or thermal imaging. In order to improve on the accuracy obtained using a single-modality camera, some approaches use a multimodal fusion strategy. These can be mainly categorized as early or late fusion methods.

Early fusion combines the different types of imagery before passing the data to a predictor. This can be performed in multiple ways, such as a weighted fusion of each pair of corresponding pixels [JLL$^+$17] [GM14] or merging one of the images with the gradient magnitude of the other [GM14]. In some contexts, some colors are not common in the RGB images and therefore the thermal image can occupy that colorspace in the fused image. One example of that is wilderness search, where the RGB image can be enhanced with thermal information painted in magenta [RMGE09]. Studies on the way reptiles combine IR with RGB have shaped more complex fusion algorithms that provide more entropy on the result image, but are not necessarily a better input for a classifier [SLO11]. However, the weights can be assigned per-pixel instead of per-image by using an adaptive pixel weight that is proportional to their importance, a property which can estimated using some cues, such as determining if the pixel belongs to an edge (RGB), considering high space and temporal variations (RGB) or assuming that the highest and lowest temperatures are more relevant than warmer ones (Thermal). Other fusion techniques include combining Curvelet with Discrete Wavelet Transforms, or using Discrete Wavelet Packet Transforms [SMD10]. In a different approach, some researchers applied Covariance Intersection with Expectation Maximization [SL09] [GCLL10].

Before performing early fusion, it is important to guarantee that the two images are properly registered. Most works that rely on multimodal information perform registration with a simple affine transformation. However, this method is not perfect when the scenes are not planar, for example in images of faces. Ma et al. [MZMT15] have developed a more complex registration technique specifically for faces, which converts the image into edge maps and registers them via a non-rigid transformation.

Late fusion (or decision fusion) means merging the final results of multiple separate classifiers into a single stronger classifier. In [SCN$^+$16], the authors propose a solution to recognize faces in a RGB, thermal and depth facial database, combining the output of six different classifiers. Although the recognition problem is very different from detection, part of the high-level architecture of the solution can be adapted to face detectors. The system extracts hand-crafted features for each type of imagery, combines the results obtained with each of those types into a single feature vector that is fed to a weighted nearest neighbor classifier. At the same time, each modality is processed using a deep learning approach with Convolutional Neural Networks which is then also combined into a single classification and is finally merged with the manual feature extraction approach to obtain a final classification.

The work of Simon et al. uses the RGB-D-T Facial Database [NNGM14] from 2014 which is open to the research community and contains facial images from 51 subjects with varying poses, expression and lighting. The thermal images of this database were captured with a AXIS Q1922 camera, along with the synchronized RGB and depth images that were captured with a Microsoft Kinect of resolution 640x480.

Ensemble methods are a way of increasing the accuracy of a machine learning system by combining the output of different algorithms to improve the accuracy. They can be considered late fusion techniques and tend to provide better results, mainly if the merged models have significant differences [CSCG16]. Some of these methods are bagging, stacking, voting and boosting [OM99].

- **Bagging** divides the training data into small sample populations, each of them being fed to a classifier (always the same algorithm, but different input data). After that, the outputs are combined into a single result by aggregating the predictions of each classifier into a single one. The generation of each training set relies on generating bootstrap samples, which means randomly collecting samples of the original dataset, so each new dataset will contain repeated elements and some of the original ones will now not be present, ultimately reducing the variance.

- **Stacking** ensembles comprise different models where each one makes their prediction and the results are combined with another machine learning algorithm that is stacked on top of the output of the initial models.

- In a **voting** ensemble, each model makes its prediction, here called vote, and the final output for each instance is either the one that receives the majority of the votes or an weighted average of them with predefined weights.

- The core principal of **boosting** relies on combining multiple weak models (perform only slightly better than a random one) to obtain a much stronger model. Instead of being a parallel method as the others mentioned in this section, boosting is run in multiple iterations and in each of them a certain weight is applied to each training sample according to how hard that sample is to predict. Difficult samples will be assigned a larger weight than easier ones, so the models will tend to focus on the hardest instances. In the end, the output of each model is combined into a single one. This principle of boosting is applied by the well-known AdaBoost algorithm [FS95].

2.5 Occupancy and face detection

Occupancy detection means being able to understand how many people are inside a vehicle and where they are seated. Usually this is performed using face detection algorithms, but this is not always the case. An exception can be found in [MLS+09], where the authors rely on template matching to understand if someone or something is occupying a seat, using a system composed of a single 360º near-infrared camera. The camera is placed under the rooftop, in the middle of the vehicle cabin, in a position that allows it to see all passengers of a 5-seat car. The document compares the error rate obtained using three different template selection strategies and combining edge detection, local normalization and temporal fusion. The different strategies were tested with a dataset composed of 53928 video frames of children, adults, objects and empty seats, with variations in illumination. In order to evaluate the system, the authors measured the Equal Error Rate (EER) of each strategy, which is obtained when thresholds are adjusted so that the frequency in which empty seats are detected as occupied equals the frequency in which an occupied seat is detected as being empty. In non-uniform illumination conditions, which is the type of conditions this dissertation wants to tackle, the authors report EERs of 1.68%, 0.69%, 4.01%, 11.07% and 1.84%, respectively for the front left seat, front right seat, rear left seat, rear center seat and rear right seat.

Considering a face detection approach for occupancy detection, the output of a face localization algorithm can either be a pixel-based segmentation or a bounding box or shape around the facial area. Multiple methods of face detection in RGB images have been developed. An algorithm that works in real-time is Viola-Jones framework [VJ04], which was the state of the art in real time face detection for many years. It acts as a binary classifier and detects objects by partitioning the original image in multiple rectangular patches, each being submitted to a cascade of weak classifiers that are combined to create a stronger classifier according to a learning process. More recently, new object detection techniques based on deep learning have been proven to achieve great performance at a low inference time.

In the context of face detection in thermal imaging, the fact that the human body stays at a fairly constant temperature of 37°C eases the detection of a body using a thermal camera. In that sense, to detect faces in thermal images, some works simply apply image segmentation according to the temperature registered at each pixel [KYNT09] [TOHH05], an approach which may not provide good and robust results in a vehicle environment, where people seat behind each other or even where the car itself may be hot and be confused with a face by the algorithm.

Viola-Jones algorithm can also be applied to thermal imaging and it is possible to optimize it to work better with this type of imagery. Basbrain et al. [BGC17] propose preprocessing the image using the gradient magnitude method and a method proposed by Otsu of automatic threshold selection for picture segmentation, which maximizes the separability of the resultant classes in gray levels [Ots79]. Also, instead of utilizing Haar-like features as in the original Viola-Jones algorithm, they propose using LBP features and HOG features.

A different solution using thermal information was proposed in [CYN14]. It applies a binary threshold filter to the image and counts the number of white pixels in each line to estimate the position of the face.

A database aimed at face recognition, named Carl Database, consists of thermal, NIR and RGB pictures of 41 subjects and a total of 7380 pictures under different lighting conditions, captured with a TESTO 880-3. The dataset contains expression variations and people with and without glasses [FZMFA14]. Alongside this database, a new and simple face segmentation algorithm was conceived for thermal imagery [MEDFZ10]. This algorithm starts by determining the vertical and part of the horizontal projections of the binarized image, similarly to the approach proposed in [CYN14], except that the latter does not determine the vertical projection and uses the full horizontal one. Then, the top of the face is estimated from the vertical projection. Since the subject is always at the center of the image, by halfing the distance between the top of the face and the bottom of the picture, the algorithm calculates a center point. From that point to the top of the face, the horizontal projection is determined, with the purpose of estimating the right and left limits of the bounding box. Finally, the lower limit is guessed according to the most common rectangular aspect ratio of the human face. It is important to note that both these algorithms based in projections have very limited applicability when the background is not clean.

More low-level image processing approaches are proposed in [RFN17], all of them preceded of a binary threshold filter. The authors experiment with Template Matching, Face Countours (using an edge detection algorithm and then extracting the highest point and the ones that correspond to the largest facial width, similarly to [MEDFZ10] and [CYN14]) and taking advantage of Chamfer Matching. There is no information on the accuracies obtained, but the processing times were respectively 69ms, 68ms and 257ms.

Another solution proposed in [WHD$^+$12] claims to detect faces in thermal images with an accuracy of 91.4% using curves with a head-like shape, although only one person can be facing the camera at a time and there should exist no other heat-emitting objects in the image, a condition which is not always guaranteed in the context of monitoring of the vehicle interior that this dissertation tackles. The dataset used contains images captured by a FLIR Thermo Vision A20M, which are the input of a three-phase algorithm for face detection. Their images do not contain absolute temperature information, instead they are converted by the camera to RGB using a color palette.As a first step of their solution, since the images contain three color channels, the images are preprocessed, resulting in the extraction of only the red component and a conversion of this component to grayscale, discarding the other colors. This grayscale image is then converted into black and white according to a predefined threshold. After that, a disk element with a certain radius is used to perform morphological closing, followed by the filling of the holes in the closed image, resulting in the object's boundary and terminating the preprocessing phase. Secondly, some fiducial points are searched using specific algorithms for each (Fig. 2.5), more specifically to calculate the position of the top of the head (a), left most (b) and right most (c) points, left (f) and right (e) neck points and the left (f) and right (g) points of the lower part of the neck. Finally, the left most point and right most point are used to calculate the center of the face, from which a curve is generated and used to mask the image, keeping only the facial area.

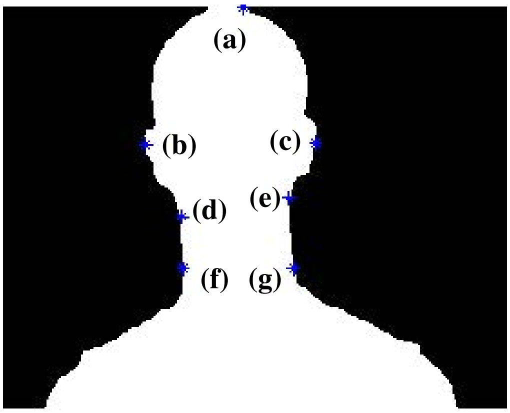


Figure 2.5: Identification of some fiducial points for further analysis using head curve geometry. (a) Head Point (b) Left Most Point (c) Right Most Point (d) Left Neck Point (e) Right Neck Point (f) Left Mean (g) Right Mean. (Adapted from [WHD$^+$12])

In the context of monitoring the interior of vehicles, a solution has been proposed to detect occupants in a vehicle as well as their position and physique [ISH$^+$09]. This is achieved using a custom-made thermal camera positioned in front of each seat and aiming to the expected face location. The camera developed to tackle this problem had a resolution of 320x240 pixels, a FOV of 52°x40° and captures light in the wavelength range $8 - 12\mu$m. The system detected the occupants and classifies them as adult or child and in terms of position (left, center or right). Some of the challenges considered were the elimination of the reflection caused by the side windows (glass reflects FIR light) and the background in case its temperatures are high. The authors claim to obtain satisfiable accuracy, although no details are given on the experiment methodology and validation results.

Recent advances in deep machine learning caused the surge of more solutions for object detection. Since 2013, multiple object detection frameworks based in deep learning have been proposed, vastly improving on the state of the art. These include R-CNN [GDDM14], Fast-R-CNN [Gir15], Faster-R-CNN [RHGS17], SSD [LAE$^+$16], R-FCN [HCE$^+$17], YOLO [RDGF15], YOLOv2 [RF17], YOLOv3 [RF18] and RetinaNet[LGG$^+$17]. The latter beats the competitors in AP (39.1%) and AP_{50} (59.1%)[2] in the COCO [LMB$^+$14] dataset, being able to predict at 5fps. The three YOLO variants are the fastest in the list of deep learning based detectors while maintaining a competitive accuracy. YOLOv2 is a real-time object detection system, processing images at 40

[2] Average precision (AP) is a metric used for object detection. It is further explained in section 4.2.1.

to 90FPS on a Titan X and with a AP_{50} on VOC 2007 of 78.6% and an AP_{50} on COCO test-dev of 48.1%. Most applications of deep learning object detection frameworks use pretrained weights on ImageNet and retrain them to the intended task with the necessary image classes. However, one problem with following this approach in the context of this dissertation is the lack of publicly available pretrained weights for this network on thermal images. Despite that, it is possible to apply deep learning to thermal imaging. In [JLL+17], the authors use YOLOv2 to detect vehicles in images taken from an UAV, captured with a RGB and a thermal camera. Different fusion techniques are applied. One of them is weighted early fusion, where the RGB image is merged with the thermal image, with a weight of 80% for the first and 20% for the second, resulting in a single RGB image (Fig. 2.8). Another fusion technique that is experimented is band combination, where the neural network is fed an image of 4 channels, 3 for the RGB and 1 for the thermal component. The last experiment is performed using late decision fusion, in which YOLOv2 is run on each type of imagery and the detection boxes are later combined according to their confidence scores. After training in images of 1000 vehicles and testing in 673, the conclusion is that all three methods work significantly better than detection on solely RGB or thermal, but with lower error ratios for weighted fusion (9.59%) and decision fusion (10.04%) methods than for band combination (14.46%).

To understand which object detection framework might perform best in the task of face detection, Nguyen-meidine et al. [NMGKBM17] compiled a comparative empirical analysis on the accuracy of different frameworks applied to the face detection problem. According to the study, Viola-Jones has the lowest accuracies with a mAP of 67.16% and 37.68%, respectively on the FDDB [JLM10] and Casablanca [Ren08] datasets but is the fastest, hitting 60fps in an Intel Xeon CPU E3-1270 (3.60GHz) with a Nvidia M4000 GPU. R-FCN obtains the best accuracy with a mAP of 92.92% at 6fps. The authors also claim that YOLOv2 is the second fastest, running at 19fps, but with a considerably bad accuracy on those datasets, although there is no information provided accuracy obtained with this framework. Therefore, it is hard to extract conclusions on how YOLOv2 might compare with the other object detection frameworks in those face datasets. According to the authors of YOLOv2, its performance on COCO dataset is on pair with the best alternatives in the previous primary challenge metric, AP_{50}, although it performs poorly in the new metric, AP. However, they claim better results with the 3rd version of the framework, as described in section 2.5.1.

2.5.1 YOLOv3

The high-level architecture of YOLOv3 is documented in Fig. 2.9. Its fully-convolutional architecture comprises a feature extractor followed by three more sets of CNNs, in a similar way to feature pyramid networks.

The feature extractor, called Darknet-53, contains 53 convolutional layers. For pretraining the classifier on 1000 classes from images of ImageNet [JWS+09], the author stacks a global average pooling and a fully connected layer on top with the softmax activation for prediction. Those extra layers are then removed but the remaining layers are used as feature extractors for the full object

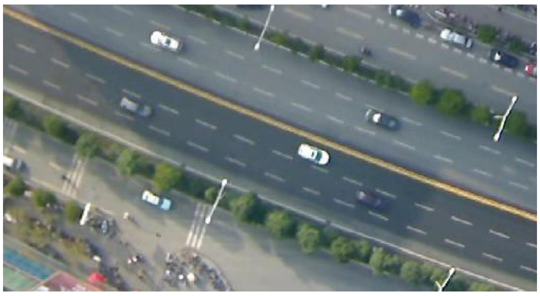


Figure 2.6: Aerial RGB image taken from an UAV. (Extracted from [JLL+17])

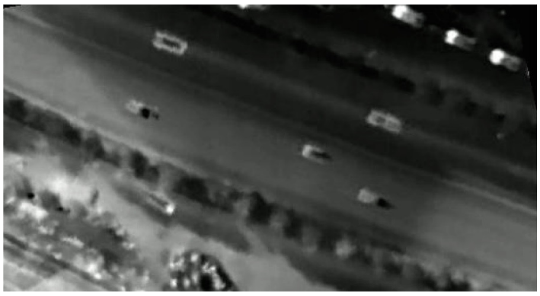


Figure 2.7: Aerial thermal image taken from an UAV. (Extracted from [JLL+17])

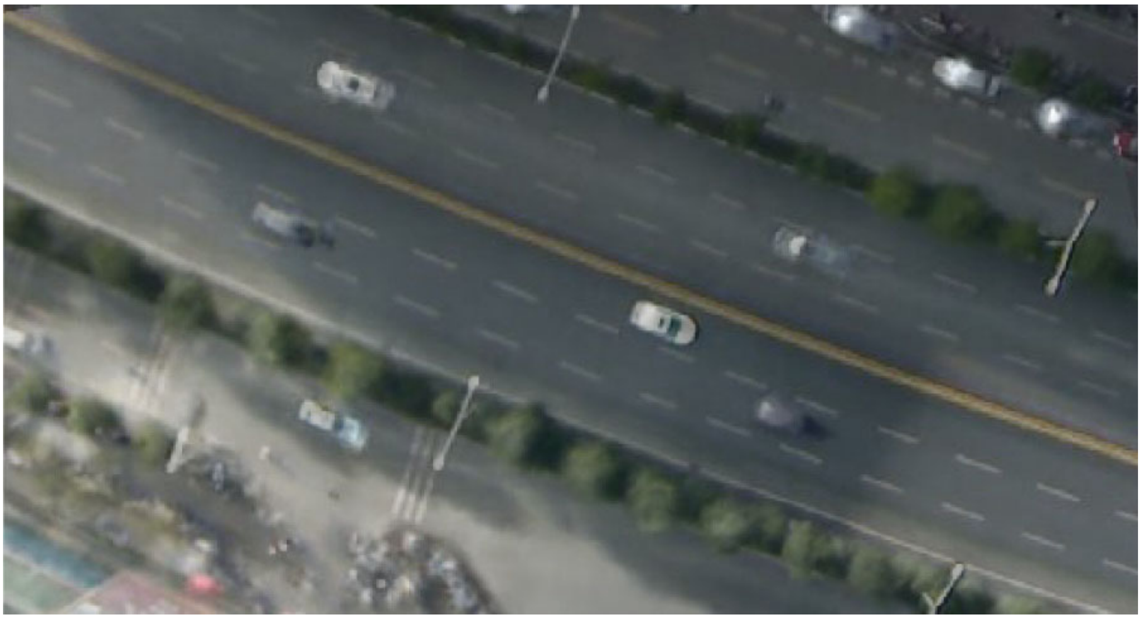


Figure 2.8: Fusion between RGB and thermal image (weight = 0.5). Notice the brighter spots corresponding to the heat emitted by the vehicle engines, mainly visible in the top right corner. (Adapted from [JLL+17])

detector, which is then trained with the COCO dataset to predict bounding boxes, confidences and classify the objects.

The fact that the feature extractor is followed by a feature pyramid network, means that the YOLOv3 object detection system returns, for a given input, three different sets of bounding boxes, with the objective of predicting well at different scales. In fact, one of the major improvements when compared to its previous version is the performance on small objects, AP_S, reaching 18.3% in the 608x608 version against YOLOv2's 5.0% in the COCO dataset.

Another change brought by the new version was a slightly different loss function (2.2). Here, L is the number of outputs. S and B are, respectively, the size of the grid used for prediction, where each cell predicts B bounding boxes. Each bounding box is defined by 5 values corresponding to its center (x and y), its dimensions (w and h) and a confidence level C. 1_i^{obj} is either 1 or 0, depending on whether there is an object in cell i. 1_{ij}^{obj} is 1 if cell i contains an object and the confidence of the j th bounding box predictor is higher than the other predictors of the same cell,

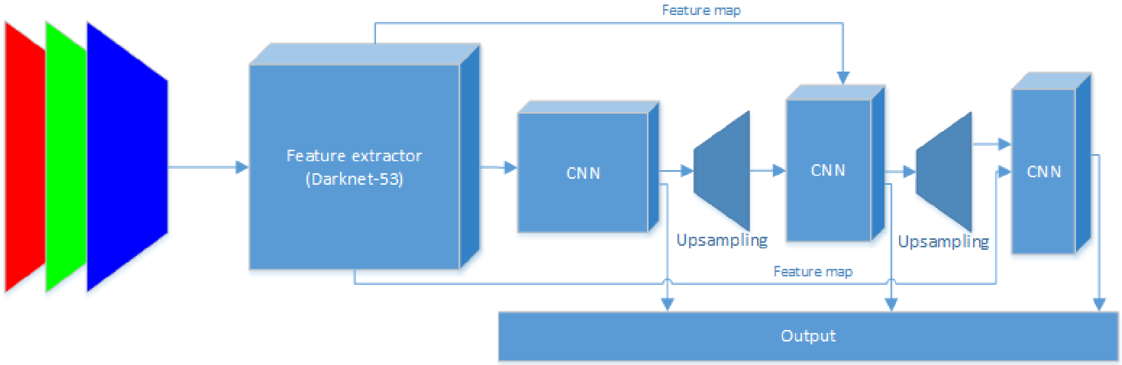


Figure 2.9: High-level architecture of YOLOv3.

otherwise it takes value 0. 1_{ij}^{noobj} works similarly but takes value 1 when no objects are in cell i.

$$
\begin{aligned}
&\sum_{l=0}^{L}(\\
&\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2 \\
&+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2] \\
&+\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}-\hat{C}_i log(p_i(C_i)) \\
&+\lambda_{noobj}\sum_{j=0}^{B}1_{ij}^{noobj}-\hat{C}_i log(p_i(C_i)) \\
&+\sum_{i=0}^{S^2}1_{ij}^{obj}\sum_{c\in classes}-c_i log(p_i(c)) \\
&)
\end{aligned}
\tag{2.2}
$$

As in previous versions, the loss function takes into account the correctness of the center and dimensions of the predicted bounding boxes, the confidence given to objects, the confidence when there are no objects and the correctness of classification when there is an object. The changes are in the way the last three components of the loss function are calculated, using logistic regression instead of the previous squared difference, and the enclosing sum ($\sum_{l=0}^{L}$) that corresponds to the output of each layer of the feature pyramid network.

2.6 Facial component localization

Many algorithms require the analysis of specific areas of the face. Therefore, after performing face detection, it is possible to locate certain fiducial points (marks used as reference, usually the most distinct and easier to identify features of a set of images). A comprehensive survey on the detection of fiducial points in RGB images is available in [WGT+18]. Some algorithms focus instead on the detection of specific regions corresponding to facial features such as the nose, lips, eyes, eyebrows, forehead and chin.

Thermal images present some drawbacks for locating facial features: they lose relevant high-frequency texture information and when a person is wearing glasses it is not possible to see the eyes. Despite that, previous research has been conducted on face registration in thermal images.

One approach focuses on detecting the eyes and mouth by determining a set of interest points on the face and then clustering those points by location in order to obtain centroids that correspond to the three facial features [TOHH05].

In [MBP10], the authors locate the eyes, nostrils and mouth of a person using. Haar features are used to find the eyes and nostrils, an information which is then used to estimate the position of the mouth. Their experiments contained 78 images of 22 subjects and the algorithm reaches a detection accuracy of 80% for the eyes and nostrils and an accuracy of 65% and 73%, respectively for closed-mouth and open-mouth situations.

A limitation of the work by Martinez et al. is that they do not consider people wearing glasses. A solution that contemplates both situations was developed by Wang et al. [WLSJ13] comprising a pipeline that starts with an eyeglass detector. After that, for faces with glasses, they look for a valley in the horizontal and vertical projection of the facial image and consider that to be the position of the eyes. For faces without glasses, the process requires face detection, extraction of Haar-like features and classification with a SVM.

In order to detect the eyes of a person in a thermal video, it is also possible to take advantage of the blinking action, which is a spatio-temporal characteristic [YHG+14]. After estimating the eye position, Yue et al. determine a facial ROI for blush detection, by extracting a region slightly below the eyes. For the frames where there is no blinking information, the authors apply the Lucas-Kanade tracker.

In terms of RGB images, multiple research has been conducted on facial landmark detection with different characteristics mainly in terms of accuracy and speed. An analysis of such algorithms is out of scope for this work so the reader is referred to a survey in [WJ18].

2.7 Person state characterization

Characterizing the state of a person includes monitoring their activities, overseeing their physical condition and understanding their emotions. The more information a system can collect on the state of its users, the better it will be able to satisfy them.

2.7.1 Emotion recognition

Multiple research has been done in terms of identification of human emotions, using and combining RGB, depth and thermal information [CSCG16]. The output of such an algorithm can either be a classification of the facial expression from a defined set of possibilities, a measure of the arousal[3] or an estimation of valence[4].

Although RGB images tend to give more high frequency information than thermal images when capturing faces, combining both types of imagery has been empirically demonstrated to improve the accuracy of facial expression recognition [NKCL14]. In fact, thermal images detect variations in the blood flow produced by some emotions. For example: fear decreases the blood flow in the nose, forehead and maxillary; anxiety increases it in the periorbital, supraorbital and forehead regions; joy decreases it in the nose region; pain decreases it in the forehead and maxillary regions and the feeling of guilt decreases blood flow in the nose and maxillary [IGM14]. These changes in blood flow cause temperature variations that can be captured by thermal cameras and thus can be applied to a vehicle occupant monitoring system [Fli16]

Multiple natural databases for facial expression analysis have been created and their study is out of scope for this dissertation. Instead, the reader is referred to a surveyed list of RGB facial expression image databases in [CSCG16].

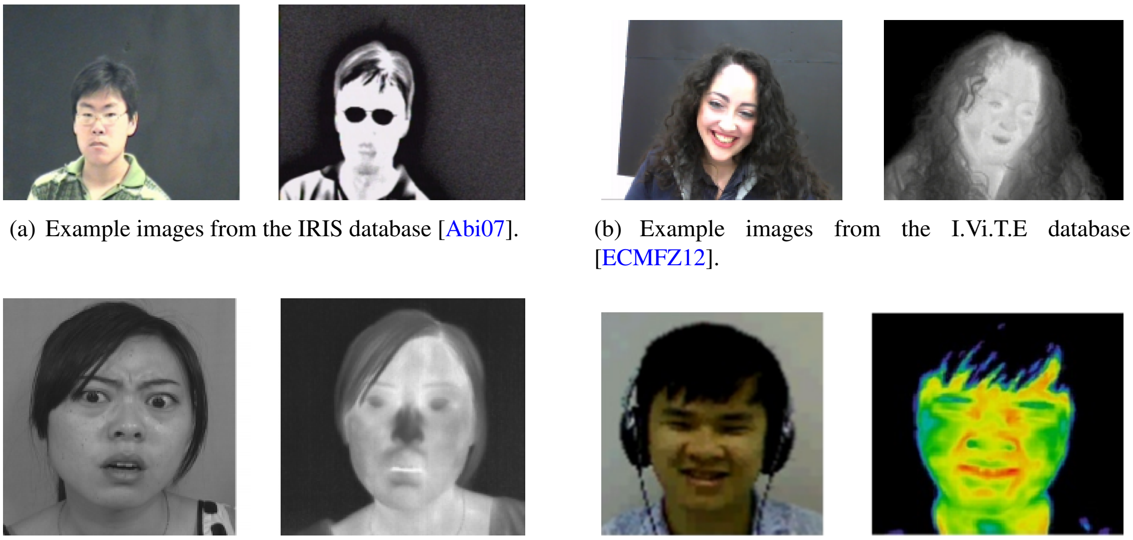
There are some datasets of facial expressions captured with a thermal camera which also include the corresponding RGB image. IRIS Thermal/Visible Face Database [Abi07] consists of simultaneously acquired unregistered thermal and visible face images under variable illuminations, expressions, and poses. Another database containing RGB and IR images of facial expressions is I.Vi.T.E [ECMFZ12]. It contains pictures from 49 subjects and the expressions are triggered through audiovisual media (happy, sad, disgusted, feared and neutral) and captured with a Testo 880-3. The MAHNOB Laughter database [PMP13] is focused on laughter, but also includes images of speech to help discriminating between both. A collection of images of 22 subjects expressing laughter, smiling and speaking was obtained with a VarioCAM hr head at a resolution of 384x288, together with the visual camera JVC GR-D23E Mini-DV of resolution 720x576.

The possibly largest currently available dataset with thermal images of people expressing different emotions is the Natural Visible and Infrared facial Expression Database (NVIE) [WLL$^+$10]. The database includes both RGB and thermal images of 215 subjects expressing six different emotions: happiness, sadness, fear, surprise, anger and disgust, although not all of them are present for each subject. The images of this dataset include variations in terms of lighting and usage of glasses and the emotions are both posed and spontaneous, the latter being triggered through audiovisual media.

The low amount of waiting time between each emotion clip is one of the criticism to the previous datasets mentioned in [NKCL14], leading to the creation of the Kotani Thermal Facial Emotion (KTFE) dataset, where a smaller number of subjects also display posed and spontaneous motions, triggered through audiovisual media and paying special attention to the time lag phenomenon. Images from 26 subjects were obtained and the camera used in this dataset for visible and thermal image collection was a NEC R300.

[3] Intensity of an emotion.

[4] Measure of attractiveness or averseness of a particular situation, usually a way of categorizing emotions according to their degree of positiveness. For example, the emotion of happiness has a high valence whereas anger and sadness have negative valence.



(a) Example images from the IRIS database [Abi07].

(b) Example images from the I.Vi.T.E database [ECMFZ12].

(c) Example images from the NVIE database [WLL$^+$10].

(d) Example images from the KTFE database [NKCL14].

Figure 2.10: Example images from different RGB and thermal facial expression databases.

Fig. 2.10 contains examples of images from these facial expression databases.

In terms of algorithms for facial expression recognition, there are multiple possibilities. Researchers have automated the process of emotion recognition on visible images and a survey on those algorithms can be found in [CSCG16].

Considering thermal images only, one of the previous works applied deep Boltzmann machines [NRM09] to estimate the level of valence [WHG$^+$14]. Their accuracy is of 62.9% in 32 subjects from the NVIE dataset, but it increases to 68.2% when mixing that data with images from MAHNOB Laughter database and NIST/Equinox. The latter is a dataset created by Equinox Corporation that has been used in multiple studies but is no longer available to public. In a different work, researchers divided the face in 6 regions and the corresponding moment invariants were fed to a SVM multi-classifier, together with the mean pixel value and standard deviation of each RoI [BRSD15]. The authors tested the algorithm in the KTFE database and, despite the lack of temperature information (only color-mapped images were available), they were able to obtain an accuracy of 87.5% predicting between 4 facial expressions and "neutral".

2.7.1.1 Blush

The detection of blushing in faces opens the possibility of identification of some feelings such as anxiety and embarrassment. It can also be an indicator of deception, which can be particularly useful in the context of vehicle interior analysis.

Thermal imaging has been used to detect blushing successfully [YHG$^+$14]. S. Yue et al. claim to achieve recognition rates of about 77% TPR and 60% TNR, by analyzing images of 51 subjects captured with an Argus P7225. Their method for ROI locating uses the eyes as points of reference to obtain a small portion of the skin below them. Tracking is performed using the Lucas-Kanade method, employing a template large enough to guarantee robustness during blinking. In the task of binary classification of blush, their classifier hits an accuracy of 67%, but the authors note that the algorithm is highly susceptible to pitch and yaw head movements.

A combination of thermal imaging and a RGB camera was experimented by Ioannou et al. [IMB+17] and the subjects participated in a dialog with the experimenter, being asked questions to enforce social, serious and compliment conditions, in varying order. The main difference of this study when compared to previous works is that it focuses on a positive reinforcement rather than negative In the visible images, only the red channel was considered and in the thermal ones the temperature of certain ROIs was extracted.

2.7.2 Breath

There are multiple ways of detecting the breath patterns, the most commonly used being physical sensors placed on the body of the analyzed person. This type of solution is intrusive, so a solution using only thermal imaging would be helpful in multiple situations. Such experiments have been performed in previous works [PYC+15] [MP06] [WMR14] [FP10] [AHJ+11], but none of those studies were conducted in the context of mobile computing with low resolution images. However, recent research has demonstrated that it is possible to detect the breath patterns of a person with a smart-phone based thermal camera.

In [ANTV16], the authors propose a new method for respiration rate estimation using a mobile device, claiming an accuracy of 94% by estimating the respiratory pattern from thermal video of the philtrum region captured by a FLIR Lepton 50^{o} at a distance between 30 and 40cm. This work proved that a low-resolution camera can recognize respiratory patterns with relatively high accuracy. Face detection was performed with the Viola-Jones framework and the Kanade–Lucas–Tomasi feature tracker [JT94] later applied to address the problem of subject motion. Since MATLAB's implementation of Viola-Jones is focused on RGB images, the authors built their own dataset of thermal images of faces (as positives) and other heat-emitting objects (as negatives). After successful face detection, the ROI is extracted and its values are averaged. Finally, a Fast-Fourier Transform (FFT) of the data is obtained to estimate the breath rate from the strongest frequency.

A different approach for the estimation of breath patterns is proposed in [CJMBB17], where a thermal camera is pointed at the nostril area, as in other previous works, but improving the robustness of the system in three main challenges:

- Scenes with varying temperatures - although our body temperature distribution is internally controlled, the corresponding pattern can be affected by the ambient temperature [KK09];
- Capability of tracking the nostril region of interest in most conditions
- Improvement of the quality of the breath signal.

To tackle these challenges, a respiration tracking algorithm is proposed, with a processing pipeline of three main components.

To solve the first challenge, an optimal quantization is calculated, mapping temperature values to digitalized color-mapped pixels. This is done by carefully tweaking the quantization algorithms usually applied in thermal processing for mapping from the temperature provided by a pixel's value to a color equivalent according to a palette of colors.

Secondly, the thermal gradient magnitude matrix is calculated using a Median Flow algorithm [KMM10] on points automatically selected from the nostril on the gradient magnitude map of the thermal image. Sometimes the point features are totally lost and, to handle that, a threshold is applied on the number of tracked points, so that if it falls below a certain amount, the ROI is reset and new point features are found.

As a final step in the respiratory rate tracking process, the algorithm proposed by Y. Cho et al. improves on other solutions, which simply calculate the average temperature of the ROI. The problem with the previous approaches is that they are highly vulnerable to small changes in the bounding box of the tracked ROI and to windy situations, as well as when the orientation of the head changes. Instead, the new solution consists in building a 3D Thermal Voxel-based feature in order to increase the quality of the respiratory signals. The 2D thermal image is converted into a 3D surface where the z-axis corresponds to the temperature and, after that, the nostril area is integrated to obtain a volume corresponding to the relative z-depth of the nostril in relation to the rest of the facial area. The algorithm is tested on a custom database composed of three different sets of images in different conditions: constrained movement and varying ambient temperature, sedentary activity under controlled room temperature and mobile context with a dynamic scene temperature. The MSE for each dataset was, respectively, 0.677, 1.015 and 1.57, values which are better than the traditional algorithm of pixel value averaging, which results in MSE values of 0.996, 1.453 and 1.609. Based on this solution, DeepBreath [CBBJ18] was proposed to automatically recognize people's psychological stress level (mental overload) based on their respiratory patterns. This solution contains three main points. First, a respiration variability spectrogram is generated from the respiratory data by applying the Power Spectral Density function for the duration of the recording, resulting in a single-channel 2D image where the horizontal dimension is the time, the vertical one is the frequency and the pixel values are determined by the magnitude of the corresponding frequency as calculated using Power Spectral Density (PSD). Then, this data is augmented employing a sliding square window over it of size 120x120 that is repeated for every second of the original spectrogram. Finally, the augmented data is fed to a CNN composed of two convolutional-pooling layers and a fully-connected layer before the output. The result of the system is a binary prediction or based in three possible levels: no-stress, low-level or high-level. The solution achieves an accuracy of 84.59% in the binary predictions and an accuracy of 56.52% distinguishing between three levels.

As an experimental application of the work in [CJMBB17] and [CBBJ18], a smartphone-based breathing sensing platform named ThermSense was developed [CBBJM18]. It is composed by three main modules: thermal image extractor, breathing estimator and respiration variability spectrogram (RVS) generator. The thermal image extractor obtains the raw absolute temperature information of each pixel from the thermal camera. Next, the image is passed to the breathing estimator. In the reported version of ThermSense, to minimize computation resources, the ROI selection is done manually but, as seen before, the process can be automated. After the identification of the ROI, the remaining steps proposed in the DeepBreath solution are executed in order to produce one-dimensional breathing signals, to extract the respiration variability spectrogram and to estimate the mental stress level.

One aspect that was not addressed in the studies mentioned in this section is the influence of mustaches when measuring the temperature of the philtrum area. It is possible that facial hair causes the occlusion of the skin and the ROI becomes less vulnerable to temperature changes caused by breath, therefore increasing the difficulty of the task of respiratory pattern estimation.

2.7.3 Heart rate

Massachusetts Institute of Technology (MIT) introduced a technique entitled "Eulerian video magnification" (EVM) [WRS^{+}12] to amplify color or small motions in videos that are invisible to the naked human eye. The EVM framework is able to amplify those motions to make them easier to see such as sound from vibrating objects, the motion of hot air or the human pulse. Therefore, the heart rate of an human can be extracted from a video by preprocessing it with EVM. Multiple studies and experiments have been carried to detect the heart rate using a preprocessed video with EVM, but only a few use thermal images to achieve that or as an improvement in a combined RGB and thermal solution. The first research conducted in this sense can be found in [GSMP07], where the authors obtain an accuracy of about 90% in their own dataset composed of thermal videos from 34 subjects recorded with an high-resolution camera built by Indigo Systems (no longer commercialized), that captures images at 640x480 in the MWIR region at 30fps. The heart rate is calculated from three different facial regions. The main issue of this solution is the fact that it relies on high-resolution imagery, but in some cases it might be useful to be able to measure the heart rate from lower quality images. Another similar work implied identifying blood vessels of the forehead region and estimating the heart rate from their color variations [GF13], resulting in accuracies above 85% in a dataset with 32 subjects. However, this also requires a high-resolution thermal camera in order to extract the necessary information.

More recently, researchers attempted to measure the heart rate using lower resolution thermal cameras, calculating the mean temperature of the ROI after the video's color changes are enhanced with EVM. In [BGK16], the authors are able to achieve this with a thermal camera of resolution 640x480 capturing at 60Hz. The chosen ROI was the chest because it was closer to ground truth than the other considered ROIs: entire face, right cheek, forehead and arm. The algorithm was tested in a 10 second video of a still subject and the results show a strong correlation when evaluated visually and quantitatively. This work was later improved by the same authors through experimentation and using different ROIs [BHGK17]. In this second study, two of those chosen ROIs were facial, more specifically the forehead and jugular artery, as visible in Fig. 2.11. Of the two, the forehead ROI provided results better correlated with the electrocardiogram ground truth data. A limitation of this work is the fact that it does not address the problem of subject movement, which is a big barrier for heart rate estimation, but corresponds to the type of situations expected in the context of this dissertation when monitoring vehicle occupants.

Therefore, to measure the heart rate when there is subject movement, tracking algorithms have to be applied on the images, as proposed in [HBM16], where before applying a spatio-temporal filter (not EVM) the ROI is tracked using the framework proposed in [KMM12], achieving an accuracy of 91.04% in videos from 22 subjects. Still, however, each subject is required to stay still and only subtle motion is considered.

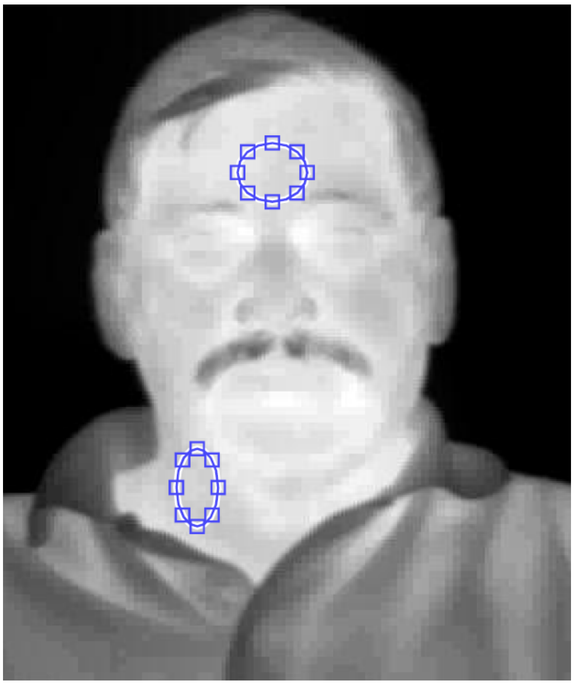


Figure 2.11: Facial ROIs for heart rate measurement. (Extracted from [BHGK17])

There is still space for further research in the field, namely by attempting heart rate recognition with thermal cameras of lower resolution and applying new methods for using EVM in videos with large motions [EHDF15]. The latter is particularly important in a vehicular system, where subjects may change their position or the orientation of their face and body.

2.7.4 Drunkenness

Drunkenness is a physiological condition that can be observed with thermal images, because, when a person is drunk, the facial circulatory system increases activity as alcohol is consumed [KA15].

Extensive work in this topic has been done by Koukiou and Anastassopoulos [KA12] [KA13] [KA15]. In [KA15] they use images of 41 subjects, captured with a FLIR A10 camera. For each person, they record 50 pictures sober and 50 images 90 minutes after the subject ingests 62.4ml of alcohol, equivalent to four glasses of wine. They train multiple neural networks where each analyzes a specific small region of the images, in order to understand which parts of the face are better to discriminate between sober and drunk. Their conclusion from that experiment is that the forehead and the nose are the areas that contain more vital information for this type of classification. Then, they show that the accuracy is higher when considering only the forehead than the whole face. Using a feed-forward neural network with one hidden layer trained in images from a single person and tested on the others, they obtain an accuracy of 93.1% for the sober state and of 84.2% for the drunk state.

In [NC17], the authors employed a processing pipeline where the image is fed to a Pulse-Coupled Neural Network for segmentation, then subject to the Principal Component Analysis algorithm and finally to a SVM classifier. In their database of 10 subjects and 400 images, they claim an overall accuracy of 97.75%. This value is significant higher than the results obtained by Koukiou and Anastassopoulos, but it is hard to perform a perfect algorithm comparison because the datasets are different.

2.7.5 Smoking

Smoking while driving is a cause of distraction, and therefore considered risky driving [MP07]. Wearable equipment can be used by the monitored occupant to detect if he his smoking [SLMT13] [RGST17], but that approach is very intrusive. Using information obtained from optical cameras, it is possible to detect the action of smoking in a passive manner.

In terms of visible light, multiple previous works have shown the possibility of detecting smoking with RGB cameras [WC11] [TCTW13]. The first step of the detection process is face detection. Then it can be followed by locating the mouth, since when the cigarette is found near that area, it is highly likely that the person is smoking [WC11]. Another possibility is to locate the hands of the subject and, if they are close to the face, look for a small handheld object (cigarette), otherwise try to detect the smoke itself by clustering the optical flow vectors of two consecutive frames obtained with Lucas-Kanade method, according to their length [TCTW13].

We have been unable to find any attempted work on the detection of people smoking (cigarettes or cigars) taking advantage of infrared cameras. However, it is a fact that the tip of a cigarette burns at considerably high temperatures. In fact, the peripheral of the coal burns at around 500°C [Bak74], therefore using thermal imaging vastly eases the process of detecting the act of smoking, since there will be an easily detectable high temperature spot (Fig 2.12). Nevertheless, there are challenges. One of them is the avoidance of false positives, since smoking is not the only reason for high temperatures in a small area to be captured by a thermal camera. Another problem is the occurence of occlusions, which may happen very frequently while smoking, specially if the cigar/cigarette is lowered or moved outside the vehicle through a window.

2.8 Discussion

There are numerous applications of thermal imaging in the context of inner-vehicle analysis, both in terms of the habitable and in terms of perceiving the state of its occupants. However, it is important to understand that some of them require specific cameras and camera positioning, which means that multiple cameras would be required to detect everything mentioned in this section.

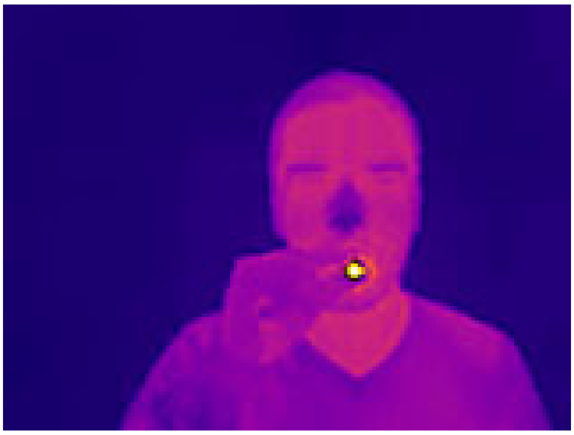


Figure 2.12: Person smoking a cigarette, easily detectable by a thermal camera.

Moreover, some of the solutions here presented are not well prepared to perform in situations of high camera and person movement, which are expected in a driving scenario. Therefore, there still exists the challenge of developing methods for the techniques described in this section with images that suffer from high trepidation, subject movement, light variations and ambient temperature changes.

Chapter 3

Data acquisition

In order to have training and test data for the algorithms described in this document, a dataset was created by capturing data inside a vehicle. This dataset is property of Bosch Car Multimedia Portugal, S.A. and is not limited to data important for the purposes of this dissertation, but also includes subjects performing activities that are relevant for the development of other monitoring algorithms. A predefined process was used for capturing data from each subject after a brief explanation and the signing of a written consent.

The setup consisted of a camera operating both in the infrared thermal and in the visible light spectrum. For this purpose, the FLIR ONE Pro camera was chosen, mainly due to the fact that it combines both modalities in a small device and ensures calibration between frames. This camera has a thermal resolution of 160x120 and a RGB resolution of 1440x1080, capturing frames at a rate of 8.7 per second with a FOV of $55^\circ/43^\circ \pm 1$. The thermal sensor operates in the $8 - 14\mu$m waveband, measuring temperatures between -20°C and 400°C with a thermal sensitivity of 0.15°C.

The participants were asked to perform some specific actions and activities, namely head movements in multiple axis, simulating facial expressions, simulating fatigue, wearing glasses, smoking, entering the vehicle and leaving the vehicle.

3.1 Equipment

The collection process described in this section requires the equipment described in Table 3.1. Not all of the items are required, as some parts of the guide may be skipped (such as smoking). Due to temporal constraints, in our dataset, no respiratory measurement system was available to be used as ground truth data.

Equipment	Quantity	Model	Purpose
Respiratory measurement system	1	Any	Ground truth data
Heart rate measurement system	1	Quirumed portable pulse oximeter [Por]	Ground truth data
RGB + Thermal camera	1	FLIR ONE Pro	Record synchronized thermal and visible video
Cigarettes	1+ per subject	Any	Simulate smoking activities
Lighter	1	Any	Light cigarettes
Linux-running device with available USB port	1	Any	Obtain frames from the camera

Table 3.1: Required equipment for the capturing process.

3.2 Capturing software

Since the FLIR ONE Pro camera is designed to be used while coupled to a smartphone, no official driver exists for directly acquiring its frames from platforms other than Android or iOS. Therefore, a custom-built driver was used in order to extract synchronized frames from the camera connected via USB to a Linux-running machine. During our recordings, we connected the camera to a NVIDIA Jetson TX2 via a USB (Universal Serial Bus) interface and established the physical connection between the camera's USB-C male port to the Jetson's USB-A female port using an adapter that, despite not being part of the USB standard [USB17], is available in the market.

The driver we developed is based on previous reverse engineering efforts done by the community of users of the FLIR ONE Pro camera to be able to interface with the camera from a Linux machine [Que]. The driver extracts the frames from the USB endpoint 0x85, using a bulk transfer. The data transmitted by the device at every interrupt contains both frames and metadata. The first sixteen bytes are the magic bytes (0xEFBE0000) and are followed by integers containing the size of the remaining fields in the transmitted data. Then, follows the raw thermal image and the compressed visible image, finally followed by meta information, such as the state of the shutter of the thermal sensor (to ensure the thermal image is valid, that is, not captured during a calibration). When the shutter is in FFC state (flat-field correction) we discard the corresponding frames.

3.3 Session setup

A capturing session with a subject should always be preceded by the following steps:

1. Gather the required equipment.

2. Obtain written consent from the subjects.

3. Explain the process to the subjects.

4. Choose a route or multiple routes that will be taken. The route should take slightly more than 7min and 10s to traverse, though this is not a requirement. We don't need every subject to perform the route, some subjects can be recorded when the car is not moving. We just need to check for unexpected issues that may occur while the car is moving on the road.

5. Connect the camera to the on-board computer.

6. Secure the camera in front of the right front seat.

A picture of the full setup is presented in Fig. 3.1.

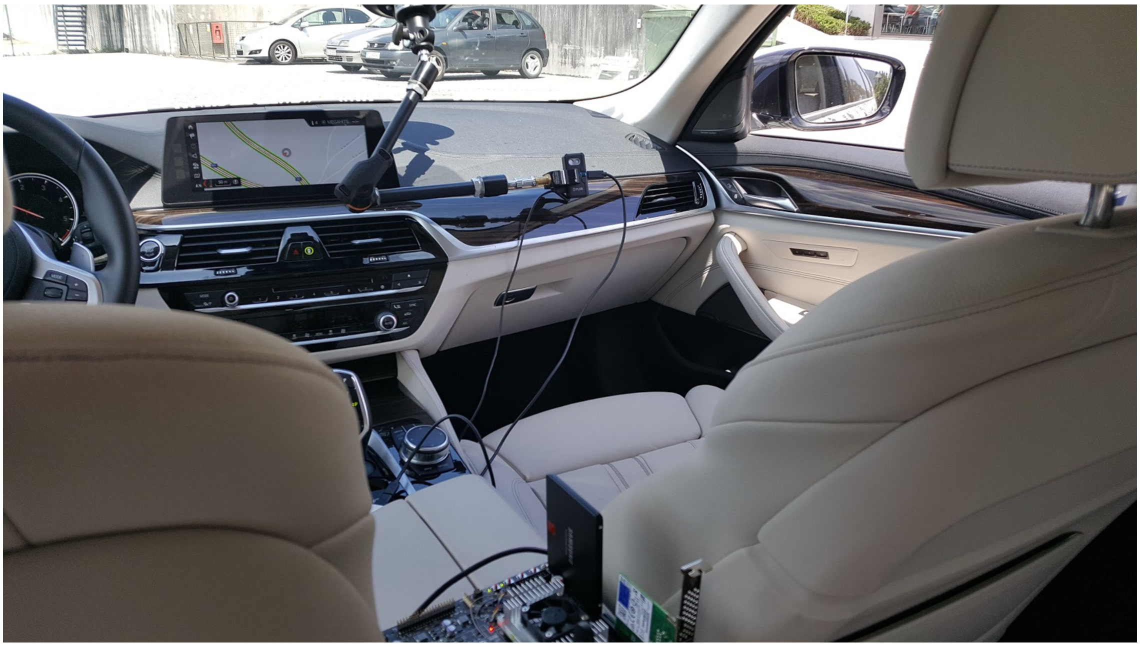


Figure 3.1: Example capturing setup inside a test vehicle. The camera is placed in front of the right front seat and is firmed in position with a folding arm.

3.4 Collection process

During recording, and to minimize waiting times, the temperature should be changed once for each subject, in order to capture images in hot and cold environments for everyone. If the recording sessions follow each other without interruptions, having a single point of temperature change brings the advantage of requiring less waiting time and decreases the cost of the heating and cooling process.

If at any "wait" instruction the required time has already passed, the process should be continued anyway without waiting.

When a step requires to "change vehicle temperature" the temperature should be set to the opposite of the current status. Therefore, if the vehicle is cold, the windows should be closed and

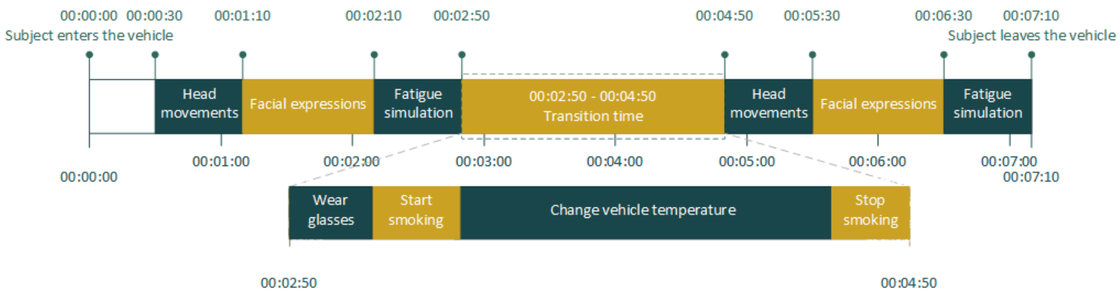


Figure 3.2: Timeline of events of a capturing session.

the air conditioning (A/C) set to maximum. Otherwise, the A/C should be set to minimum and the windows rolled down if the temperature in the exterior is low.

During the recording, the person responsible for the capture process should be in the backseat of the car, appearing in the footage and taking note of information regarding some of the frames.

Assuming a route is already defined, the steps for recording should be as follows:

1. Collect information from the subject, to be stored in file "demographics.csv":

    - ID
    - Age
    - Gender
    - Height
    - Waist-to-Head Height
    - Beard
    - Mustache
    - Race
    - Hair size

2. Remove glasses from subject.

3. Start capturing and start the clock, ensuring all devices are synchronized (start at the same time).

4. Subject enters the vehicle and sits on the front right seat.

5. Start driving.

6. If windows are open, willing subjects light up a cigarette and start smoking. The cigarette should be lighted directly from the mouth using a lighter. From time to time place the cigarette outside the vehicle (arm outside the window). Subjects should keep smoking during the following steps.

7. Wait until 00:30.

8. Perform head movements, with the following timings:

   - 1 second to adopt the final position;
   - 1 second holding in final position;
   - 1 second going back to initial position;
   - 1 second waiting time before going for next position.

   The movements should be performed in the following order:

   (a) Roll to the left (roll)

   (b) Roll to the right (roll)

   (c) Rotate to the left (yaw)

   (d) Rotate to the right (yaw)

   (e) Rotate down (pitch)

   (f) Rotate up (pitch)

9. Wait until 01:10.

10. Perform the following facial expressions, ensuring at least 2 seconds per expression and a minimum waiting time between expressions of at least 5 seconds:

    (a) Happy

    (b) Sad

    (c) Surprise

    (d) Fear

    (e) Anger

    (f) Disgust

11. Wait until 02:10.

12. Perform fatigue movements, with a minimum waiting time between movements of at least 2 seconds.

    (a) Blink in each second for 5 seconds

    (b) Yawn for 3 seconds

    (c) Close eyes for 5 seconds

    (d) Squint for 5 seconds

13. Wait until 2:50.

14. Wear glasses.

15. Change vehicle temperature.

16. If windows are open, willing subjects light up a cigarette and start smoking. The cigarette should be lighted directly from the mouth using a lighter. From time to time place the cigarette outside the vehicle (arm outside the window).

17. Wait until 4:50.

18. Perform head movements, with the following timings:

    (a) 1 second to adopt the final position;

    (b) 1 second holding in final position;

    (c) 1 second going back to initial position;

    (d) 1 second waiting time before going for next position.

    The movements should be performed in the following order:

    (a) Roll to the left (roll)

    (b) Roll to the right (roll)

    (c) Rotate to the left (yaw)

    (d) Rotate to the right (yaw)

    (e) Rotate down (pitch)

    (f) Rotate up (pitch)

19. Wait until 05:30.

20. Perform the following facial expressions, ensuring at least 2 seconds per expression and a minimum waiting time between expressions of at least 5 seconds:

    (a) Happy

    (b) Sad

    (c) Surprise

    (d) Fear

    (e) Anger

    (f) Disgust

21. Wait until 06:30.

22. Perform fatigue movements, with a minimum waiting time between movements of at least 2 seconds. If the subject is wearing sunglasses, only the yawn movement should be performed.

(a) Blink in each second for 5 seconds

(b) Yawn for 3 seconds

(c) Close eyes for 5 seconds

(d) Squint for 5 seconds

23. Wait until 07:10.

24. Continue driving until destination (if not already reached).

25. Subject leaves the vehicle.

26. Stop capturing and reset clock.

27. Finish process or repeat from step 1 with new subjects.

3.5 Data storage format

The script used for capturing automatically stores each frame of each modality in an image file. In particular, the images captured in the visible region are stored in the JPEG lossy format, due to their considerable dimensions (1440x1080). The images captured with the thermal sensor are stored in the PNG 16bit compressed lossless format, to guarantee no data is lost from the 160x120 14bit image provided by the camera, while ensuring a reduced file size.

Additional annotation and subject meta-data is also stored in the database according to the structure defined in this section. Each capturing session folder (corresponding to a subject) contains images and annotations, totaling around 1GB of data per subject.

3.5.1 Folder structure

The EU General Data Protection Regulation (GDPR) [Eur16], in force since the 25th May 2018, is an EU regulation aiming to protect and increase the data privacy of its citizens and changing the way organizations deal with that data. Some points of this regulation require the user to be able to, at any time, download all data available of him or delete it upon request. To comply with the GDPR, it is important that the database maintains a folder organization that allows for quick access or removal of data related to a particular subject.

The root folder of the dataset should contain a file “demographics.csv” with information on each subject and a folder for each of them, named by their ID. These folders contain multiple information:

- The visible component of the captured videos as a sequence of JPEG images, named by the ID of the frame.

- The thermal component of the captured videos as a sequence of 16 bits per pixel (bpp) PNG images, named by the ID of the frame.

- A file “info.csv” with information on some of the extracted frames.
- A file “annotations.xml” with the annotations of the same extracted frames in Dlib’s format.

3.5.2 File "info.csv"

The info file of a subject stores information on some (not necessarily all) of the recorded frames. Each frame can be specified either by the corresponding time or by the frame ID. The remaining columns of the table are: glasses (yes/no), smoking (yes/no), hot (yes/no, according to the temperature of the habitacle), expression (name of the facial expression the subject is performing), yawning (yes/no), eye status (open, closed or squinted) and heart rate. If, in any field, the label is not easy to determine, it should be left empty.

3.5.3 File "demographics.csv"

The demographics file stores information on some characteristics of each subject. If, in any field, the information is unknown, it should be left empty. The fields are:

- Person - identification of the subject
- Age
- Gender - male, female or other
- Height
- Waist-to-head height
- Presence of beard (yes/no)
- Presence of moustache (yes/no)
- Ethnicity - one of:
  - white
  - asian
  - black
  - hispanic
  - other
- Hair size - one of:
  - short
  - medium
  - long

3.6 Annotation process

Information on the facial bounding boxes and facial landmarks of each subject was automatically generated using existing algorithms for RGB face and landmark detection. Although this method presents disadvantages if the visible and thermal images are not aligned, it is a very fast and easy way of obtaining a considerable amount of labeled data with minimum effort. The algorithm chosen for landmark detection in RGB [BTK17] is able to output 3D coordinates of each facial point, but we are only interested in the 2D view from the camera in the context of this work.

Although the labeling process is automatic, it requires manual validation to ensure the landmarks have been correctly guessed. In this sense, an human analyses the output of the algorithm being run every 10 frames and decides if the bounding boxes and points should be saved or not, according to their accuracy. Accepting the automatically-generated labels of a frame requires that all visible faces have been successfully detected and if there are any partially visible faces the frame should be discarded to ensure consistency in the labeling method.

Overall, the labeling process of the images of a subject has the following steps:

1. Add a new line to the file "demographics.csv" with information on the new subject.

2. Run the auto-labeler every 10 images and store the results.

3. Filter the automatically-generated labels keeping only the properly labeled facial bounding boxes and landmarks.

4. Store the labels in the file "annotations.xml".

5. Add the frame IDs of the automatically labeled images to the file "info.csv".

6. Fill the remaining columns of the "info.csv" file. Not all fields need to have data, but it is particularly important to ensure that the actions that subjects are requested to perform are labeled.

7. If necessary, manually add more frames to the "info.csv" file and label them, appending the bounding boxes and landmarks to the file "annotations.xml".

For standardization, the facial bounding boxes should have well-defined limits. Its top should be delimited by the boundary between the forehead and the hair. The bottom corresponds to the limit of the chin. Left and right boundaries are defined to end exactly in front of the ears, therefore excluding them (see Fig. 3.3).

3.7 Size and characteristics of the database

In total, our database contains recordings of 38 unique subjects, captured according to the capture guide. For privacy and legal reasons, all images were collected with the vehicle standing still instead of moving. A total of five different vehicles are used, although it should be noted that

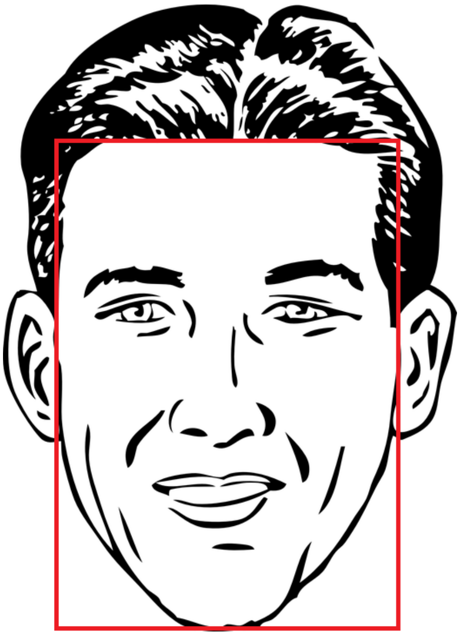


Figure 3.3: Example of face annotation.

having low variety in the vehicle used, or even using only a single vehicle, is not very problematic in the context of occupant monitoring, because it is assumed that the system would be tweaked for a specific car model and should take advantage of that property.

In terms of gender distribution, data was captured from 33 males and 5 female subjects of white ethnicity. The average age is 28.8 with standard deviation of 10.1. There are 33 subjects between 21 and 31 years old, 3 between 32 and 43, 2 between 43 and 54 and 3 between 54 and 65 years old. The height distribution is reported as a histogram in Fig. 3.4.

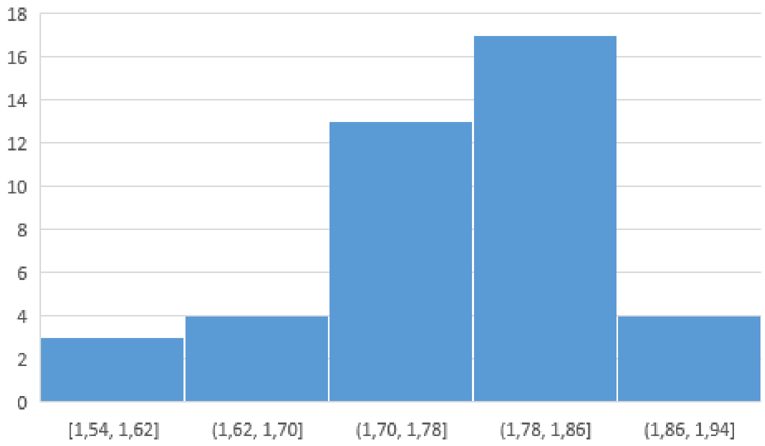


Figure 3.4: Height distribution in the dataset.

Hair size and facial hair can influence the accuracy of multiple algorithms both in RGB and thermal images. The hair size of each subject was evaluated subjectively by two human labelers using 4 categories: bald, short, medium and long. The database contains images from 3 bald subjects, 27 with short hair, 3 medium and 5 long. In terms of facial hair, we classified the subjects in terms of beard and mustache. In our database of 38 subjects, 63% have beard and 66% have a mustache.

In total, the database contains 87286 frames, where 5361 have been auto-labeled with facial bounding boxes and facial landmarks and manually filtered to remove incorrect labels generated by the automatic labeling process. From those labeled frames, 805 also contain manually labeled information on the facial expression performed by the subject and 3961 contain information on the usage of glasses, being either "no glasses", "normal glasses" or "sunglasses".

Chapter 4

Proposed Solution

A list of possible use cases of thermal imaging for vehicle occupant monitoring was compiled in Appendix A. A subset of those features were selected to be developed and, in this chapter, their implementation details are explained and the results discussed. The requirements and use cases chosen to be tackled in this dissertation were decided together with the representative of Bosch Car Multimedia considering time constraints, economical reasons (compromise between the number of cameras and what we want to detect) and thermal camera limitations (such as the low resolution and lack of high-level information).

The use cases addressed in the context of this dissertation are:

- Face detection
- Smoking detection
- Respiratory rate estimation
- Emotion recognition

Additionally, despite not being an actual use case for a final product, this work also addresses the problem of facial landmark detection, because it is relevant for some of the use cases.

An high-level architecture analysis of the solution is presented in Fig. 4.1.

A real-time video stream captured with a synchronized RGB and thermal camera is the input of the system, which is fed to the face detector, the first stage of the processing pipeline. The resulting bounding box is then passed to a facial landmark detector and its output is used for the emotion recognizer, for respiratory rate estimation and for the smoking activity detector. The latter requires information from the face detector to help avoiding false positives. In parallel with the initial face detection, the system also detects hot spots in the thermal stream to feed the smoking activity detection algorithm.

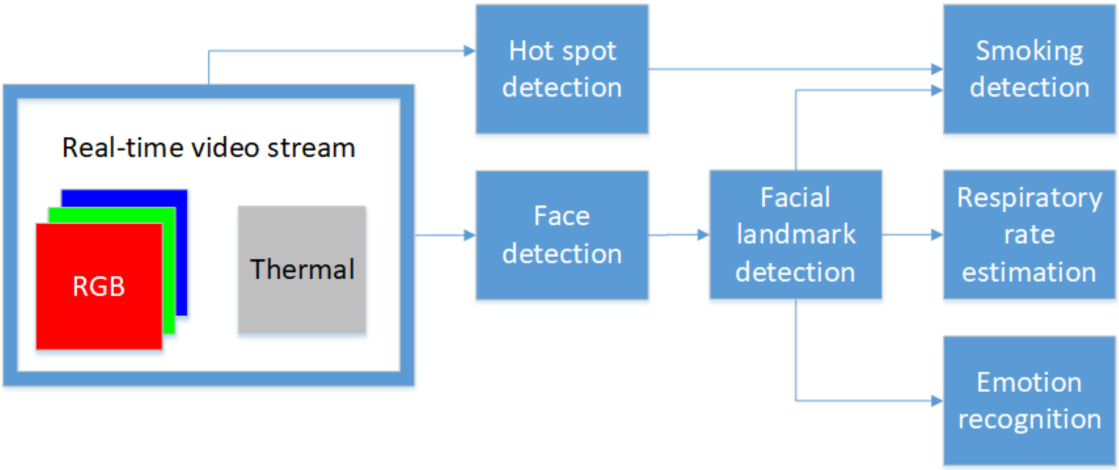


Figure 4.1: High-level architecture of the vehicle interior monitoring system.

4.1 Camera setup

This chapter discusses the advantages and disadvantages of different camera setups, by varying the number of cameras, as well as their type and position. The type of vehicle considered was a four-wheel car with a total passenger capacity of 4 or 5, driver included.

A single-camera solution is enough to perform car-seat occupancy detection with an accuracy above 90% for most seats, as proven in [MLS$^+$09] with a 360º NIR camera, but this type of setup is very limiting: it requires images to be captured by a camera having a HFOV above 180º and the variation in head rotation relative to the camera increases difficulty for further monitoring steps that require a frontal view of the face of the occupants.

We considered the possibility of having a two-camera setup. One of the cameras would be placed in the front part of the cabin, under the window. The role of this camera in the system is to capture information on the driver and front passenger. The other camera is placed under the rooftop, being responsible for capturing images of the back seats. An advantage of this setup is that all passengers can be captured without occlusions as long as the cameras' HFOV is above 50º–65º (depending on the vehicle model and exact camera position). However, the tasks of emotion recognition, heart rate estimation and breath estimation are harder, unless the thermal cameras have a considerably high resolution. The fact that the passengers do not face the camera directly also increases the difficulty.

The final camera setup chosen for this work was to place a thermal camera in front of each seat, directly facing the occupant. This setup is used in this dissertation and further research can be done to study how the same algorithms behave in different camera setups. Some advantages of this positioning are:

- More accurate face detection, as long as the occupant does not rotate the face
- Better resolution for facial landmark detection

- Easier emotion recognition

However, it also poses some disadvantages:

- Lower robustness in respiratory pattern estimation against a solution that positions the camera slightly below the subject

- One camera required per person

4.1.1 Registration

To work with the two modalities produced by the optical sensors of the camera, it is required that they are synchronized in time and aligned in space. The hardware of the FLIR ONE Pro camera is responsible for synchronizing thermal and visible images in time. However, since there is a small distance between each sensor's axis, it is important to ensure a proper alignment of the images when working with both modalities at the same time. This problem is illustrated in Fig. 4.2

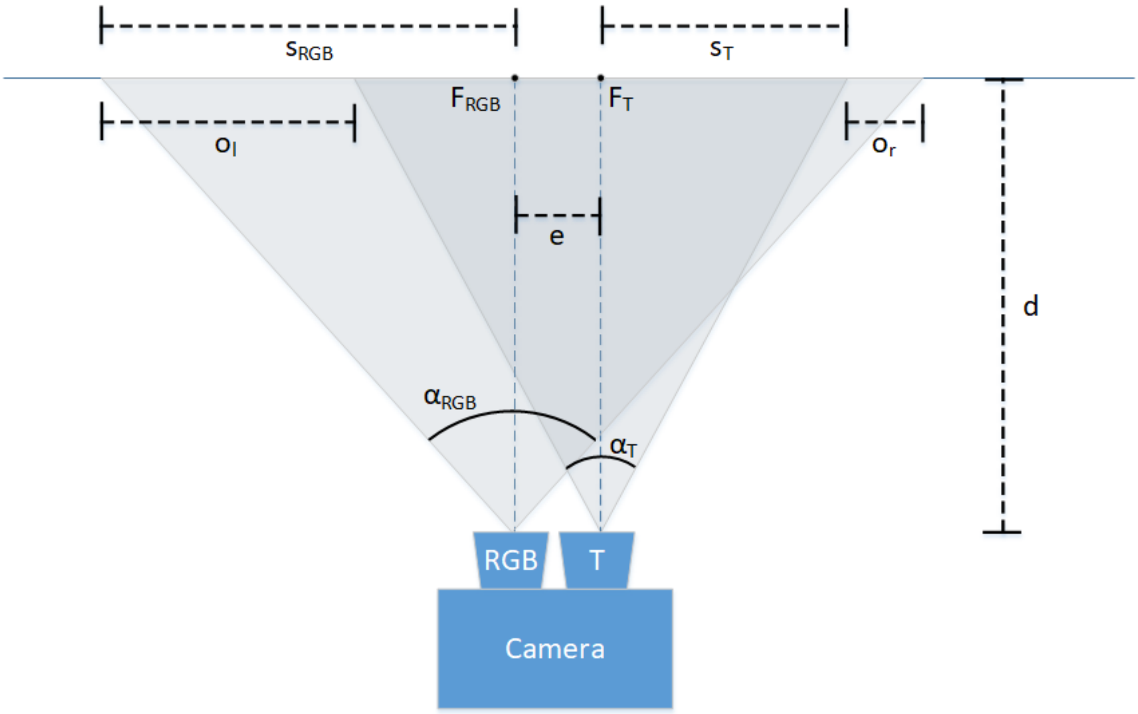


Figure 4.2: Registration of the thermal with the visible image.

In the context of this dissertation and assuming a constant camera positioning as detailed in 4.1, we estimate that the average focal distance, d, to the face of the vehicle occupant, F_{RGB} and F_T, is of 55cm. Also, the HFOV of the thermal sensor, α_T is of 55^o while the RGB sensor has a HFOV, α_{RGB} of 62^o [1] This difference in field of view allows for a compensation of the small distance between each sensor, e, of 1cm. This compensation is accomplished via software responsible for

[1] These values of FOV have been measured empirically and are subject to an error up to +- 1^o, so they should be calibrated for each camera.

cropping the RGB image in order to match the view of the thermal sensor. We can calculate the values of s_{RGB} and s_T as

$$\begin{aligned} s_T &= d * \tan(\alpha_T) \\ s_{RGB} &= d * \tan(\alpha_{RGB}) \end{aligned} \tag{4.1}$$

These two distances are, together with the distance between each lens, the required information to calculate what should be cropped on each side of the RGB image for a good alignment at the considered focal distance. Therefore, o_r and o_l are obtained as:

$$\begin{aligned} o_r &= s_{RGB} - s_T - e \\ o_l &= s_{RGB} - s_T + e \end{aligned} \tag{4.2}$$

Finally, we need to convert these two values from distance to number of pixels for cropping, p_l and p_l. Taking into account the number of horizontal pixels of the RGB image $p_{RGB} = 1440$ and the number of horizontal pixels of the thermal image $p_T = 160$, we calculate the amount to crop as

$$\begin{aligned} p_l &= \frac{o_l}{2 * s_{RGB}} * p_{RGB} \\ p_r &= \frac{o_r}{2 * s_{RGB}} * p_{RGB} \end{aligned} \tag{4.3}$$

A similar process should be applied to calculate the amount of cropping required vertically. Some variables need to be changed. First, $e = 0cm$, because both sensors are at the same height considering the image is orientated horizontally. Additionally, $\alpha_{RGB} = 47^\circ$ and $\alpha_T = 43^\circ$. Note that having $e = 0$ means that $o_r = o_l$.

Using these equations it is possible to estimate the optimal offsets for calibration considering the default focal distance of 55cm. After performing face detection (the initial stage in the system pipeline), a better alignment can be obtained by using the size of each detected facial bounding box to estimate multiple focal distances, although this method should be viewed as an approximation that is not totally robust to different face sizes. Better alignment can be achieved with more complex non-rigid transformations [MZMT15].

4.2 Validation

All the algorithms are evaluated against a custom-developed dataset specific for the context of this dissertation, even though their performance will also be tested against some datasets open to the research community, namely the NVIE dataset for face detection and emotion recognition.

4.2.1 Evaluation metrics

Some of the problems described in the next sections of this chapter require specific evaluation metrics.

In order to compare the accuracy of different object detection algorithms there are some commonly used metrics, either for computing the accuracy of a single bounding box or for evaluating the whole algorithm.

IoU (Intersection over Union). Compares two bounding boxes, usually applied in object detection to compare the predicted bounding box with the ground truth. It is calculated as the division of the intersection area by the area occupied by the union of the two boxes (4.4).

$$IoU(B_1,B_2) = \frac{B_1 \cap B_2}{B_1 \cup B_2} \tag{4.4}$$

Average Precision. Calculated for each predicted class. Usually a threshold of 0.5 is considered on the IoU and it is calculated the number of true positives (correct predictions for the class) and the number of false positives (a prediction was made for the class but is incorrect) (4.5).

$$\text{Average Precision} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}} \tag{4.5}$$

The COCO dataset also defines other more specific metrics for average precision. Their primary challenge metric is AP.

Average Precision (AP/mAP). General measure of the accuracy of object detection algorithms. It is calculated as the mean for all classes of their number of true positives divided by the number of predictions. Traditionally, in COCO dataset, this metric was referred to as "mean Average Precision", but is now simply named "Average Precision". The decision threshold for the IoU varies, however, if not specified, it is assumed that the general AP metric is averaged over multiple IoU values. Additionally, it also defines metrics for small (AP^{small}), medium (AP^{medium}) and large (AP^{large}) objects, using thresholds for the bounding box area of 32^2 and 96^2 pixels.

$$AP^{IoU=k} = \frac{\sum_{c \in classes} \frac{\text{Number of true positives}^{IoU=k}}{\text{Number of true positives}^{IoU=k} + \text{Number of false positives}^{IoU=k}}}{|classes|} \tag{4.6}$$

$$AP = AP^{IoU=[0.5:0.95:0.05]} = \frac{\sum_{i=0}^{9} AP^{IoU=0.5+0.05*i}}{10} \tag{4.7}$$

Area Under Curve (AUC). Usually the area under the ROC (Receiver operating characteristic) curves, where the true positive rate is plotted as a function of the false positive rate for a certain class or for all the classes after averaging. As the mAP, this is used in object detection algorithms as an accuracy measurement, but AUC is less suitable for imbalanced data than mAP.

The metrics above are usually related to object detection and classification. For evaluating the estimation of time series data when compared to a ground truth (e. g. respiratory and heart rate estimation), the mean squared error is a common metric.

Mean Squared Error (MSE). Measures the mean of the squared difference between each pair of predicted (Y) and expected ($\hat{Y}$) values in the series.

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \tag{4.8}$$

When considering the task of classifying a sample that belongs to a single class of multiple possible, it is common to measure performance with the categorical cross-entropy function, the F1-score or the kappa metric.

Categorical cross-entropy (H). Typically used for N-way classification. It is calculated by summing, for all classes, the symmetric of the ground truth multiplied by the logarithm of the predicted value.

$$H'_y(y) = -\sum_i y'_i * log(y_i) \tag{4.9}$$

F1-score. Measures the accuracy by calculating the harmonic average between precision and recall.

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{4.10}$$

Cohen's kappa coefficient (kappa). Used to test interrater reliability [Coh68]. The advantage of this metric when compared to a simple accuracy measure is that it takes into account the probability of the agreement occurring by chance, so it is useful to understand how an algorithm's accuracy (p_o) compares to the one of a random predictor (p_e).

$$\text{kappa} = 1 - \frac{1 - p_o}{1 - p_e} \tag{4.11}$$

The evaluation metrics to be relied on for tuning and comparison of the implemented algorithms vary depending on the task. Face detection is evaluated according to some of the official metrics from COCO dataset, more specifically the AP_{50} and the AP. To better interpret the results in the section related to face detection (section 4.3), we consider two main use cases of our detection algorithm: recognizing the existence of a vehicle occupant and accurately determining its facial location for further processing in the pipeline. For the first use case, we look at the AP_{50} metric, while for the second one we are interested in maximizing the AP, that is, the AP at $IoU = .50 : .05 : .95$. We could also consider AP_{25} for the first use case, but having bounding boxes with an IoU below 50% means that it is possible that two bounding boxes are generated for the same subject, while still passing the test and contributing positively for this metric. Finally, facial expression recognition is optimized with the categorical cross-entropy function, since it consists on a problem of classification, whereas for valence estimation we use binary cross-entropy. Additionally, for expression recognition and valence estimation, we consider the accuracy, kappa and F1-score of the solutions.

4.2.2 Cross-validation

Cross-validation in machine learning is a method of evaluating a predictive model that consists on repeated runs with different datasets, with the aim of reducing variability in the results.

In the context of this dissertation, some of the implemented algorithms are assessed using the Monte Carlo method for splitting the training and validation sets for cross-validation. This method relies on randomness to select a certain number of samples for training from the full dataset and use the remaining for validation or test purposes. Although this means that it is likely that there will be repeated data in the validation set during different runs, a big advantage of the Monte Carlo method is that there is a large number of different ways of splitting the dataset in two. In fact, being N the number of samples in a dataset and N_{train} the number of training samples to be selected, there are ${}^{N}C_{N_{train}}$ different combinations of training sets that can be generated. Another advantage of the method is that, if the dataset is not shuffled before splitting, it evades biases that could belong to certain parts of the dataset. For example, if the method for data collection changes between different recordings and the database is sorted by time, some parts of the data may have different properties and cause different results to be obtained. A cross-validation method that randomly chooses the data avoids this problem.

When splitting the dataset in training and validation, it is also important to ensure we choose a splitting method that helps assessing how well the algorithms are generalizing to unseen data. In our case, this means we need to avoid validating the algorithms in images of subjects that have been previously seen during training. For this purpose, we implemented a function responsible for, given one or more datasets, splitting them in two, ensuring that no subjects are included in both. This method is also able to handle multiple input datasets, so that it is possible to split different datasets that contain the same subjects. For example, we are able to use the NVIE spontaneous and posed database for training and just the spontaneous dataset for validation, while ensuring that the subjects chosen from the spontaneous set for training are the same as the ones chosen from the posed set.

4.3 Face detection

An advantage of thermal imaging for face detection is the temperature difference between the human skin and the surrounding environment, which means that an algorithm based on temperature segmentation should be able to detect faces up to some accuracy. However, sometimes the background is at the same temperature as the skin and more robust solutions are required. Therefore, our approach for face detection uses the YOLO object detection framework, which consists on a deep learning neural network for general object detection that we retrained, using transfer learning, for the task of face detection in thermal images.

The main reason behind our choice of the YOLO framework was its speed/accuracy trade-off. Its authors claim an inference time of 29ms and an AP_{50} of 55.3% in COCO dataset for an input size of 416x416. Considering our camera captures frames at 8.7Hz, that is, every 115ms, YOLO requires only 27% of the time between frames, leaving the remaining 84ms for handling other cameras or for execution of other tasks. Furthermore, we are also interested in reasonably good predictions of the bounding box limits, to ease the job of the remaining tasks in the processing pipeline. Although YOLO's mAP of 31% in COCO is below competitors such as RetinaNet (37.8%), we believe the compromise between speed and accuracy favors its adoption for our use case. Our dataset was split in two, for train and test, the latter containing 8 subjects, from #31 to #38, whose images are used to evaluate the algorithms. All times have been measured with a NVIDIA GeForce GTX 1080 Ti.

4.3.1 Alternatives to Deep Learning

Despite our choice of using YOLO for face detection, we also considered alternative solutions.

Some of the previous works mentioned in chapter 2 take advantage of the natural temperature of the human body. We experimented this approach and implemented a face detector based on the horizontal and vertical projections of the thermal image where pixels below a certain threshold were previously zeroed. The lower bound is located by finding a local minimum in the horizontal projection. The gradient of the projections is smoothed using a moving average and analyzed to find the maximum increase and maximum decrease in value, theoretically corresponding to the beginning and end of the remaining borders of the facial region. After this initial step for locating the face, we employed an algorithm for refinement of the upper and lower limits of the bounding box. The upper limit is located from the top 40% assuming a considerable temperature change between the hair and the forehead, that can be detected using the gradient of the horizontal projection. The lower limit is estimated from the bottom 30% of the image by calculating the mean vertical coordinate of the points generated by the Harris corner detector [HS88]. Although this method has been proven to work well for some datasets where the background is colder than the subject, we run a test on images from a subject inside a vehicle and obtained 20% of AP_{50} and 5% of AP, due to the fact that the algorithm performs very poorly when there are hot objects in the background.

We also experimented using Viola-Jones, feeding it with around 200 images followed by augmentation and using 3019 grayscale (non-thermal) images as negatives. The initial results with this approach were very poor, either using LBP, HOG or HAAR features, so we decided not to continue investing on this method.

Eventually, we decided on using YOLO to perform face detection. We initially attempted the version 2 of the algorithm, with a relatively small number of images when compared to our final solution that is trained in multiple frames from our dataset and other additional images. For this reason, the results were not satisfactory, scoring around 60% of AP_{50} in inner-vehicle images. To improve this result, an ensemble was created that combined the YOLOv2 model with a SVM (based on Local Binary Patterns and Histogram of Gradients) for classification. Then, the decision threshold of YOLO that decides if a bounding box contains a face was removed and the confidence for each bounding box was merged with the output of the SVM with a weight, respectively of 0.8 and 0.2. This ensemble resulted in an improvement in AP_{50} of 2%, a value that is not satisfactory to fulfill the required use cases of this dissertation. In order to obtain better accuracies, we increased the size of the dataset, upgraded to version 3 of YOLO and tweaked some parameters, as further explained in section 4.3.2.

4.3.2 Transfer learning with YOLOv3

Both YOLOv2 and YOLOv3 are trained on a large amount of visible images. In order to perform face detection in thermal images, we can take advantage of those pretrained weights and adapt them to our scenario where the input should be a single-channel temperature matrix and the output the bounding boxes of all faces. We studied and compared different ways of adapting the network to our input, which will be later discussed.

In order to retrain the network, we need not only facial images but also images of high-temperature objects that could be confused with our target class. Most of the datasets of infrared images have a clean background (cold) and are not suitable for good learning of negatives. Therefore, we collected and hand-labeled a total of 2075 additional images of faces in multiple random scenarios, but ensuring that no appearing subject is included in the data reserved for testing. The validation of all models reported in this section was performed using images from subjects #12 to #18 (7 subjects).

For the pretrained output to match our objective of face detection, the network should be adjusted so that each bounding box only predicts one class. Additionally, since facial bounding boxes have a certain aspect ratio, we run the K-means algorithm to cluster the sizes of all faces in the dataset and generate anchor boxes that are better tuned for the use case, unlike the pretrained version of YOLOv3 which is prepared to receive multiple classes of objects of varying size.

Furthermore, since the original network expects 3-channel RGB images, we need to either convert our thermal images to RGB, or adapt the network to accept the new input format. We have made experiments with a color palette to convert the temperature information into RGB images. Feeding the 2075 images to YOLOv2 in this new format results in an AP_{50} of 87.95% and an AP of 39.45%.

One of the main improvements in version 3 of YOLO is the accuracy of the limits of the predicted bounding boxes, therefore vastly increasing the score on the AP metric. We trained YOLOv3 with the same images, but this time combined with the first 11 subjects from our dataset, achieving an AP_{50} of 96.10% and an AP of 59.24%.

In an attempt to obtain more images for training and improve the prediction accuracy, we experimented adding the NVIE spontaneous infrared dataset to YOLOv3. Since this dataset has only been labeled with the facial points (13 in total), we employ a simple algorithm to estimate the position of the facial bounding box. First, we define the smallest box that encloses all facial points. Then, its left and right borders are increased by 15% on each side and its top border is increased by 40%. Since the bottom limit is already defined by the position of the chin, we do not need to adjust it. After adding NVIE to our list of training images we obtained much worse results, with an AP_{50} of 69.21% and an AP of 39.30%. Manual analysis of the output of the object detector in the training images shows us that YOLOv3 was outperforming the bounding boxes that automatically generated from the facial points. Therefore, due to the lack of accurately labeled facial bounding boxes and because NVIE's licensing disallows its inclusion in a commercial product, we decided not to include it in subsequent experiments.

Our next experiment was to feed the network directly with the thermal image, without preprocessing it with a color palette. We experimented tripling the single-channel input in order to match the number of input channels of the pretrained network, without applying any palette or making any other change to it. We take advantage of the fact that all the images in our dataset contain temperature information and we do not perform any kind of equalization to avoid losing that important data, considering that the facial temperatures have a limited expected range [KK09]. This experiment slightly decreased the accuracy of our model in AP_{50} but made a great improvement in the general AP metric when compared to the initial attempt with YOLOv3 without NVIE, resulting in an AP_{50} of 95.25% and an AP of 64.41%.

To avoid repeating the input, we experimented passing a single channel to the network, corresponding to the temperature matrix captured by the thermal camera. In order to prepare the network for the new input, we should understand how the first convolutional layer of YOLOv3 works and how it should be adapted to accept the new input. The output of a convolution layer is visualized in Fig. 4.3. In YOLOv3, the first layer contains 32 filters, also known as kernels, of size 3x3, which means that each value in the output convolved feature is a linear combination of the pixel values in a 3x3 square around it. A kernel is therefore defined by 9 trainable weights and, considering 3 color channels, there are $32 * 3 = 96$ kernels.

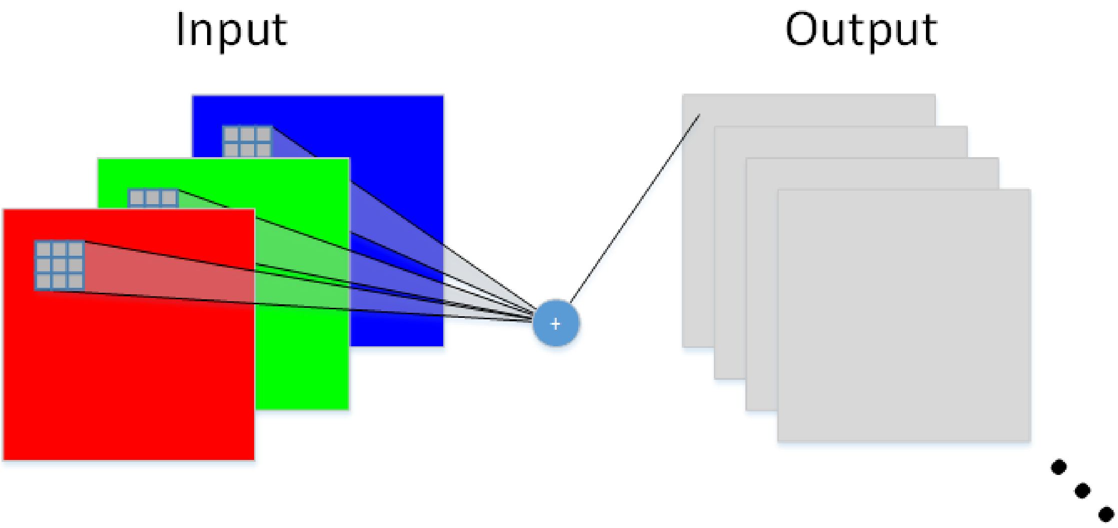


Figure 4.3: Convolutional layer connected to RGB input.

Since the input shape suffered a reduction in the number of channels from 3 to 1, we discarded the weights corresponding to the kernels of the first convolutional layer and initialized them with random values. Comparing the results of this model to the triple thermal input version, we notice an improvement in AP_{50} to 95.51%, and an increase in AP to 65.65%.

To improve those values, we tried to take advantage of the old weights of the convolutional kernels of the first layer. Since there are 3 kernels per filter (one for each color channel), it is necessary to properly combine the weights of those kernels into one. Convolutional layers calculate the output values for each filter according to the formula

$$h_j^n = max(0, \sum_{k=1}^{K} h_k^{n-1} * w_{kj}^n), \tag{4.12}$$

where h is a feature map, n is the index of the convolutional layer in the model, j is the index of a filter, K is the total size of the kernel and w is a weight matrix. Note that the output of a convolutional layer for a multi-channel input is related to the sum of the convolution operation on each channel, and not to its mean. For this reason, we add the weights of each kernel for each filter to adapt from a multi-channel to a single-channel input, in an attempt to feed similar data to the rest of the network, and therefore taking as much advantage as possible from the previously learned weights. This resulted in our best model both in terms of AP_{50}, with a score of 96.98%, and AP, scoring 65.85%. An example of detection is demonstrated in Fig. 4.4

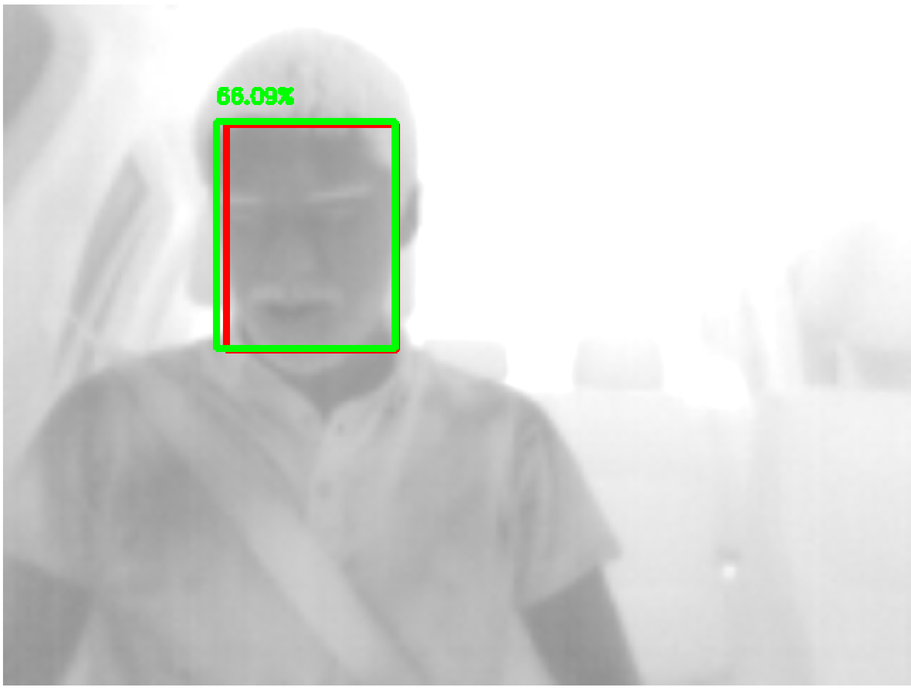


Figure 4.4: Example of a successful face detection with a high ambient temperature. The red bounding box represents the ground truth and the prediction bounding box is drawn in green together with the confidence value (66.09%).

4.3.3 Optimizing for speed

In the context of this dissertation, we are not very interested in detecting small objects, as we assume a minimum size of the faces of the occupants and distance to the camera. Therefore, it is possible that parts of the network are not contributing to the overall accuracy, because we do not have small objects in our dataset. To test this hypothesis, we grabbed the weights of our best predictor (single-channel input, reusing weights from the first layer) and removed the last output from the model, including all the convolutional layers that precede it up to the second output. The first output of YOLOv3 is given by the 82nd layer and the second output (medium-sized objects) is given by the 94th layer, so the rest of the network can be eliminated for our purposes. Inferencing with this pruned model results in exactly the same AP_{50} and AP score as the original one, which means that, indeed, the last few layers of the network are not being used at all to improve the prediction accuracy. The big advantage of making this conclusion is that we can predict in the pruned model, which means a considerable improvement in terms of speed. In our implementation, the full model takes 69ms to predict one frame of resolution 416x416, while the pruned one only takes 48.3ms, which means we get a reduction of 30% in inference time. There is still potential for improvement in terms of performance, because we are using a non-optimized implementation of YOLOv3 for Keras. Additionally, we are predicting frame-by-frame instead of in batches and we are using a different GPU from the one used by the authors of YOLOv3. Considering the 29ms of inference time claimed in the original YOLOv3 paper, we speculate that, after optimization, our pruned version should be even faster than that. We also experimented ignoring the output of the other layers but, as expected, the results deteriorated. Nevertheless, it should be noted that there is no performance gain in ignoring output from previous layers because they will always need to be part of the model, since the output layers that follow require their upsampled feature map.

4.3.4 Training without last output

Since we have concluded that the last layer of YOLOv3 is not helpful in our use case, we were able to increase its speed at inference time, but it is also possible to totally prune it during training so that it no longer contributes to the total loss (and decreasing training times). If we look at the loss function of the system (2.2), there is an enclosing sum that is responsible for adding the individual losses of each output at each scale. Therefore, if we decrease the number of output layers, L, from 3 to 2, we are effectively excluding the loss of the last output in the overall loss function. Our experiments show that there is no change in accuracy but the training times are reduced from around 4 to 3 hours.

Table 4.1 presents a comparison of results between the different approaches mentioned in this section. Using a single channel as input and reusing the weights of the initial layer provided the best AP score and a low inference time, so we trained this model again but using the full training dataset and repeated the process for a total of 5 times. Then, we used the best model in terms of validation loss (2.1297) for testing, reaching an AP_{50} of 99.60% and an AP of 79.47%.

Model	**AP25**	**AP50**	**AP75**	**AP**	**Inference time**
v2 palette	93.40%	87.95%	28.11%	39.54%	37.3ms
v3 palette NVIE	69.64%	69.21%	43.39%	39.30%	58.9ms
v3 palette	96.10%	96.10%	70.49%	59.24%	59.4ms
v3 gray 3 chan	96.84%	96.36%	82.86%	64.47%	57.6ms
v3 1 chan random weights	95.53%	95.31%	77.72%	63.15%	48.8ms
v3 1 chan reused weights	96.06%	95.98%	82.35%	65.09%	48.3ms
v3 1 chan reused weights trained on subjects #1-#30 and tested on #31-#38	99.60%	99.60%	98.06%	79.47%	48.5ms

Table 4.1: Comparison of face detection results obtained with different versions of YOLO and transfer learning techniques.

We believe that the results are very satisfactory to comply with the requirements of the system. After manual observation of the output of the face detector, one can notice that the reason the score on the *AP* metric does not increase more is mainly due to the fact that the labeling process was automatic and the face detector used for RGB images sometimes produces bounding boxes of slightly smaller size than what has been defined as the rules for labeling facial bounding boxes in the context of this project (section 3.6). For this reason, although the face detector reaches 98.06% in accuracy with an IoU above or equal to 75%, it is harder for the algorithm to exactly match the automatically generated ground truth and reach such values when the IoU threshold is higher.

4.4 Facial landmark detection

We use the ensemble of trees proposed in [KS14], a fast landmark detection algorithm that achieves a good trade-off between speed and accuracy [WJ18], to predict the position of facial landmarks in our images. To artificially increase the size of our dataset we use multiple image augmentation techniques on all labeled frames, such as brightness variation, rotation, scaling, translation, blur, sharpen, shear and horizontal flipping. In particular, for shearing, we use a small value (5^o) to ensure that the facial properties are not significantly distorted in an unrealistic way. Some of those transformations also require the bounding box and facial landmarks to be repositioned according to the new image. Additionally, for horizontal flipping, the points need to be swapped, taking advantage of the human facial symmetry. Considering relative coordinates between 0 and 1, let A be a facial point and B be its mirror in the other side of the face, the new position of the point in the image, A', is calculated as

$$\begin{cases} A'_x = 1 - B_x \\ A'_y = B_y \end{cases} \tag{4.13}$$

An illustrative example of this horizontal flip is given in Fig. 4.5.

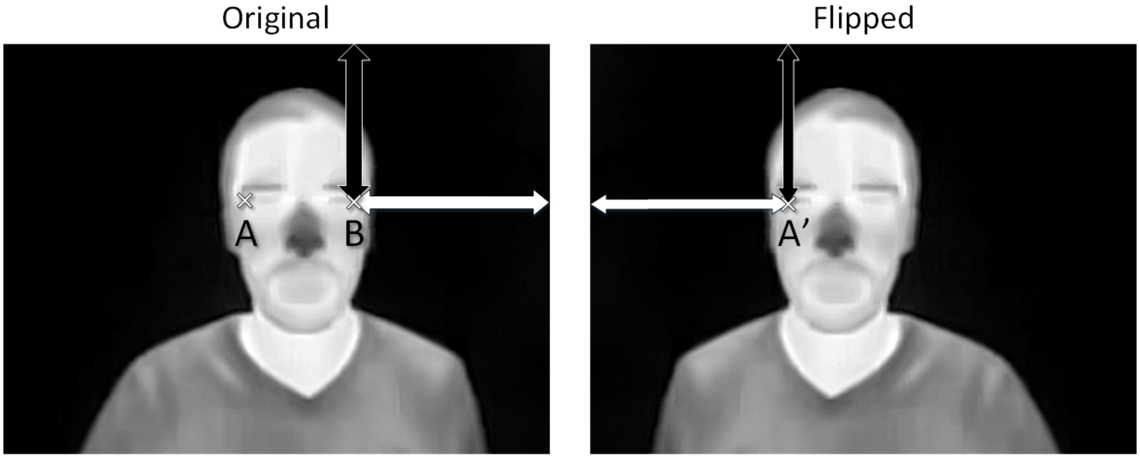


Figure 4.5: Horizontal flip of the facial landmark corresponding to the left corner of the left eye.

4.4.1 Eyeglasses classifier

Eyeglasses are opaque to infrared light and block the view to the eyes (and possibly eyebrows) of a subject. Therefore, the landmarks predicted by the ensemble of trees near the eyes in subjects wearing glasses are simply based on a generalization from the images where subjects are not wearing glasses and the eyes are visible, and the position estimation is done according to the average location of those points in the dataset.

To distinguish between images of subjects with an without glasses, we created an eyeglasses classifier based on a simple neural network, composed by the traditional architecture of convolutional and pooling layers for feature extraction a dense layers to make the final binary decision, as detailed in Table 4.2. We also reduce the input of this model by extracting only the eye region and feeding it to the network. However, since the eyeglasses classifier comes before the landmark detector in the processing pipeline, we do not have an accurate position of the eyes available to cut the RoI from, so we cut the facial image using the limits of the facial bounding box, from 10% to 50% of the pixels, counting vertically from the top. Then, the cropped image is resized to 60x30 and equalized.

To test the accuracy of the eyeglasses classifier, we split the dataset in two, leaving 30 subjects for training and 8 for test. The model is trained with augmented images generated from the labeled frames of the recordings of the training subjects. Testing the glasses classifier applied on the ground truth bounding boxes results in an accuracy of 97.96%. Using the bounding boxes generated by our face detector (section 4.3), the accuracy drops to 95.92%.

By having a glasses detector and placing it right after face detection in the processing pipeline, it is possible to tune the task of facial landmark detection for each separate case (with or without glasses), with the objective of improving the accuracy of the predictions both with and without glasses. We decided to divide the problem in two and train a model for each of the two cases. Moreover, we developed a classifier that detects the presence of glasses in a subject, so that its

Layer	Filters/Units	Parameters
Conv2D	32	kernel = 3
BatchNormalization		
MaxPooling		size = 2
Conv2D	32	kernel = 3
BatchNormalization		
MaxPooling		size = 2
Dense	32	
Dropout		amount = 0.5
Dense	1	

Table 4.2: Architecture of the eyeglasses classifier

output can be used as a decision on which landmark predictor should be further employed in the pipeline. For the landmark predictor that is specialized on glasses, we removed all points in the eyes and in the eyebrows. Nevertheless, we also compared the results using this approach with the results using a single model for subjects with or without glasses.

4.4.2 Training and results

Hyper-parameter optimization was performed both automatically using the grid-search technique and manually (due to the high search-space) and the results of that process are reported in Appendix B. The parameters we optimized with the objective of decreasing the MSE were the *nu*, three depth and multiple configurations for image augmentation. Table 4.3 contains the MSE of our facial landmark detector in our test dataset, compared to a RGB solution [KS14] and Fig. 4.6 demonstrates a prediction of the landmark detector. Results show that separating the predictor in two models (with and without glasses) is not beneficial and increases the MSE considerably, whereas using a single model the algorithm is able to outperform the RGB version of the algorithm. Additionally, it is observable that the RGB predictor has much lower accuracy when the subjects are wearing glasses than in the other cases, whereas the Thermal version of the algorithm actually performs slightly better in that situation. A manual inspection of the output of the RGB predictor in images with sunglasses shows that the occlusion of the eyes also affects the prediction of the remaining points, a problem which could possibly be mitigated by retraining the model with more images of people wearing glasses. Inference time of the algorithm in thermal images is around 15.9ms per face.

4.5 Smoking

The output of a thermal image sensor is an image where each pixel has a value that represents the temperature of the corresponding object, enabling the distinction between objects that have different temperatures. Although smoking is not limited to highly heat-emitting objects, a considerable amount of smokers use cigarettes or similar products, which produces high amounts of heat, therefore a high potential candidate to be detected by thermal cameras. Nonetheless, relying

Facial bounding box	**Glasses**	**Thermal two models MSE** ($*10^{-4}$)	**Thermal single model MSE** ($*10^{-4}$)	**Visible MSE** ($*10^{-4}$)
Ground truth bounding box	Yes	8.480	2.637	5.153
	No	4.519	2.709	3.266
Bounding box generated with our algorithm (section 4.3)	Yes	7.190	2.723	5.716
	No	4.607	2.853	3.454

Table 4.3: Comparison of the mean-squared error of our thermal landmark detector against a RGB detector [KS14] in our test dataset (using relative coordinates between 0 and 1).

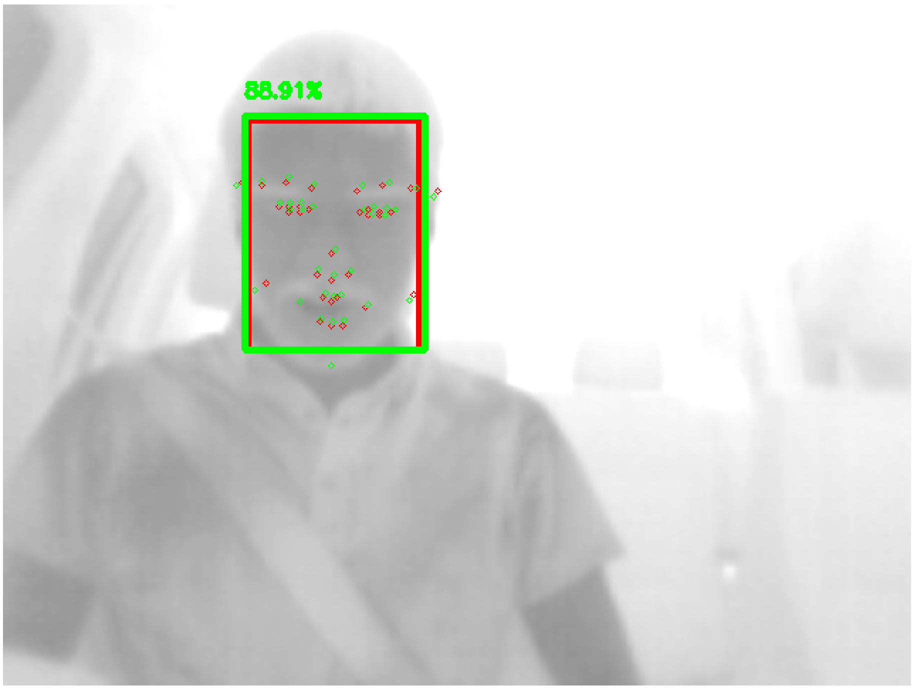


Figure 4.6: Example of a landmark detection with a high ambient temperature. Red represents ground truth and green corresponds to the predictions of the face detector and the facial landmark detector.

solely on temperature information to detect cigarettes may cause false positives and even though any object considerably hot inside a vehicle should be a red flag, it does not necessarily have to be marked as a smoking activity. Moreover, the interest behind detecting the activity of smoking also includes understanding which of the vehicle occupants was performing the activity.

To properly tackle false positives and identify the smoking subjects, the proposed solution takes advantage of the common movement pattern of smokers and the relative position between the cigarette, face and mouth. An usual smoking pattern consists on moving the cigarette to the mouth for inhaling and then putting it away for exhaling the smoke. Our algorithm tries to guess the temporal moments when the occupant is taking a puff to determine a probability of the person being smoking.

Our solution is composed of two stages. First, hot spots and faces are detected in thermal sequences. Then, a classifier uses that information to estimate the smoking probability.

4.5.1 Hot spot detector

For hot spot detection, we rely on the fact that our dataset contains temperature information for each pixel. First, a mask is created by applying a threshold on the temperature matrix with a default minimum value of 100°C. Then, we perform a 5 by 5 dilation on this mask. After that, we equalize the temperature matrix considering 90°C as the minimum temperature and 130°C as maximum. Finally, we apply the simple bob detector of OpenCV, filtering for areas with between 2^2 and 100^2 pixels and filtering by convexity with a minimum value of 0.5. This results in a list of hot spots that may or not be related to smoking activities.

An experiment was made using a smaller threshold of 60°C for hot spot detection and registering the highest temperature captured. The results of this experiment are documented in the second column of Table 4.4 (subjects 19 and 20 should not be considered for this purpose since they were smoking). Due to the high temperatures sometimes occurring in the background and to avoid a considerable amount of false positives, the value chosen for minimum temperature threshold was 100°C.

Subject	Maximum temperature of detected hot spots	Smoking device
1	69.5°C	none
2	73.7°C	none
3	72.9°C	none
4	70.6°C	none
19	118.6°C	heated tobacco
20	191.1°C	cigarette
22	60.8°C	none
31	60.9°C	none

Table 4.4: Maximum temperatures of the hot spots detected with minimum temperature threshold of 60°C. High temperatures are recorded in the vehicle for some non-smoking subjects due to the heat absorbed by some parts of the vehicle in exposure to direct sunlight.

One of the main limitations of using thermal cameras for smoker detection is different types of cigarettes such as electronic or heated tobacco. We experimented running the algorithm with a smoker using heated tobacco (Fig. 4.7). After observation of the temperature of the device during time, we noticed that it took 3 minutes and 47 seconds until it surpassed 60°C and stabilized below 70°C. Considering that about 19% of smokers usually smoke a cigarette in less than 6 minutes [Gra67], we believe that a hot spot detector using thermal imaging with a temperature threshold of 60°C should be able to detect people smoking heated tobacco with a good success rate, but using such threshold allows for multiple false positives in hot environments. For this reason, we defined a minimum temperature threshold of 100°C for the vehicle occupant monitoring system. However, the hot spot detector with this higher threshold is able to detected heated tobacco if the cigarette is removed from the device, an event which caused a temperature of 118.6°C to be registered by the thermal camera (Table 4.4).

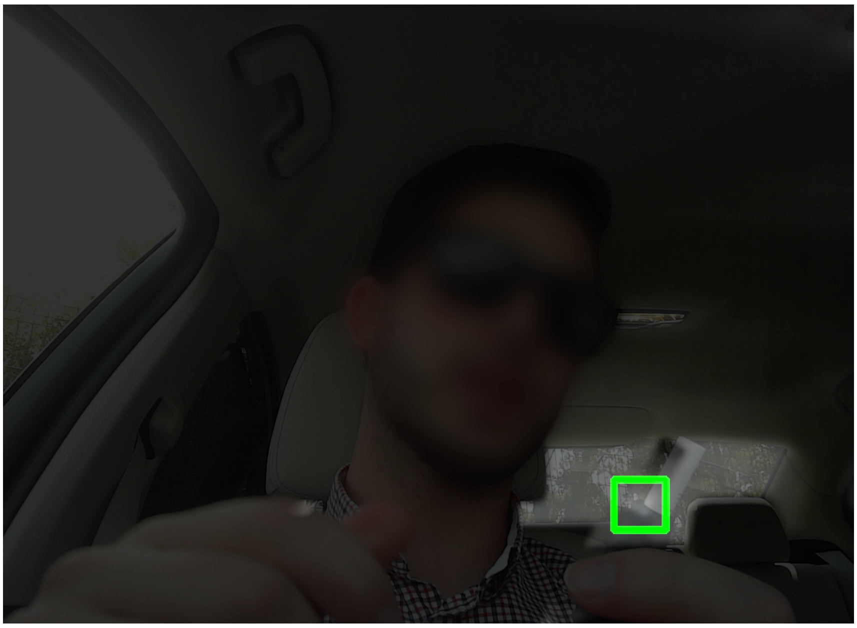


Figure 4.7: Fused RGB and thermal image demonstrating a successful detection of heated tobacco using a minimum temperature threshold of 60°C. The rear window of the vehicle is at a temperature of 52°C.

In order to compare different approaches for smoking detection in a vehicle environment, experiments have been conducted to examine the detections of an hot spot detector based on thermal imaging and the readings obtained from particle detectors installed inside a vehicle in a test scenario where an occupant in the front right seat is smoking. The air particle sensors used in the experiments detect particles of 10 parts-per-million (PPM), 2PPM and 1PPM. The testing subject was asked to smoke freely while inside the vehicle. The front windows were left open to test the case where, theoretically, the performance of the air particle sensor would be worst. A chart with a comparison of the output of each sensor is reported in Fig. 4.8.

As observable, every time the subject took a puff the thermal camera was able to detect a hot spot. At $t = 28s$, the subject lighted a cigarette and it was immediately detected by the thermal sensor. The air particle sensors were also able to capture the presence of smoke and the first peak of particle density was recorded less than 15 seconds after the subject took his first puff. We conclude that both sensors are helpful to detect the activity of smoking and can additionally be combined using an algorithm for sensor fusion. Additionally, we might observe from the output of the hot spot detector that, during this capturing session where the subject was asked to smoke without any restrictions, in 5% of the frames the cigarette was not visible by the camera due to occlusion or limited field-of-view.

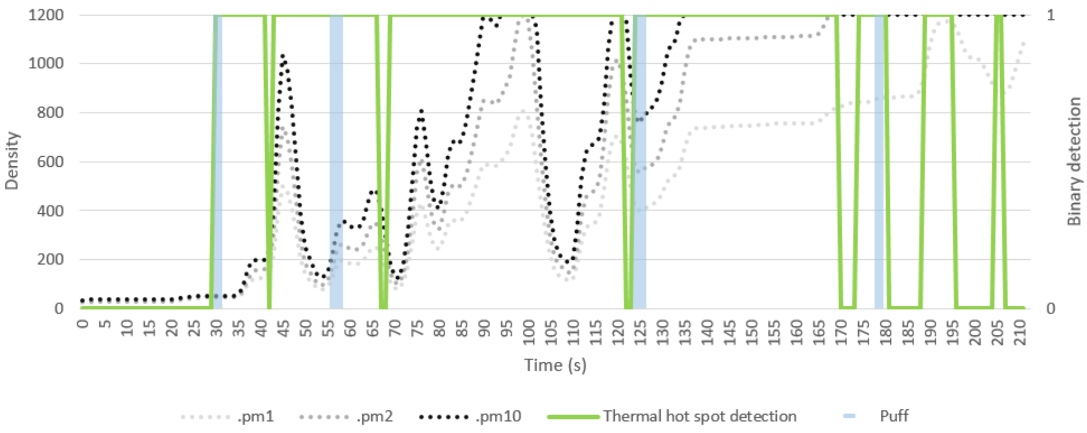


Figure 4.8: Comparison between the output of multiple air particle sensors and an hot spot detector based on thermal imaging during a smoking experiment inside a vehicle with the front windows.

4.5.2 Smoking classifier

After detecting faces and hot spots in the thermal image, a classifier estimates the probability of the subject to be performing the activity of smoking.

For classification, our base assumption is that smokers commonly perform the gesture of moving the hand holding the cigarette close to the mouth, taking a puff, and moving the hand away again [WHC+10]. Using machine learning to understand common smoking patterns would require a considerable amount of data, so we approached the problem by defining a set of rules and mathematic formulas to determine a probability of smoking.

The algorithm requires, as input, the facial bounding box of the person, the location of the center of the mouth and the position of any hot spot that is being captured by the thermal camera. In order to obtain the bounding box of the face, we employ the face detection algorithm previously described in section 4.3. Following that, we use our facial landmark detector (section 4.4) to estimate the position of the left and right corners of the mouth, which are then averaged to obtain its center. We then store, for each visible subject and frame, the distance from the mouth (obtained from our landmark detector) to the closest hot spot, or a flag indicating that there is no hot spot present in the view of the camera. However, in order to adapt to different head sizes and subject positioning inside the vehicle, this distance is divided by the diagonal of the facial bounding box, as estimated by our face detector previously in the pipeline. This results in an array per subject of distances over time, that is then processed to estimate the probability of the movement pattern being related to the activity of smoking.

After limiting the array of distances by discarding the oldest data so that its size does not exceed the maximum size of the temporal window considered (MAX_TIME_WINDOW), a probability is estimated according to algorithm 1.

We define an hysteresis band with a start ($PUFF\_START\_DISTANCE\_THRESH$) and end threshold ($PUFF\_END\_DISTANCE\_THRESH$) to decide if the cigarette is close enough to the subject's mouth in order to consider that a puff is being taken. First, if the array of distances of the

Algorithm 1 Estimate smoking probability

```
Require: D = array of normalized distances between a subject's mouth and the closest hot spot
  or a flag when no hot spot is visible
  currentTime = current time
  times = array of the same size as D that stores the time when the frame of index equal to the
  position in the array was captured
  if $\forall_d \in D, d = NaN$ then
    return 0
  end if
  $puffs \leftarrow 1$
  $insidePuff \leftarrow false$
  for $i < D.size$ do
    if $insidePuff$ then
      if $D_i >= PUFF\_END\_DISTANCE\_THRESH$ or $currentTime - hotSpotLastTimeSeen > MAX\_PUFF\_TIME\_IF\_NO\_HOT\_SPOT\_VISIBLE$ then
        $insidePuff \leftarrow false$
      else if $D_i <= PUFF\_START\_DISTANCE\_THRESH$ and $(puffs.size == 0$ or $currentTime - times[puffs.size - 1] >= MIN\_TIME\_BETWEEN\_PUFFS)$ then
        append $i$ to $puffs$
        $insidePuff \leftarrow true$
      end if
    end if
  end for
  $x \leftarrow puffs.size - MIN\_NUM\_PUFFS - TIME\_DECAY * hotSpotLastTimeSeen$
  return $\frac{1}{1+e^{-x}}$
```

analyzed time window contains no data because no hot spot was detected, we return 0 as the probability of smoking. Then, using the hysteresis thresholds, we count the number of puffs the subject has taken during the time window. In order to handle possible occlusion of the hot spot by other objects, we establish a threshold on the maximum amount of time that we consider the subject is taking a puff if the hot spot is not visible ($MAX\_PUFF\_TIME\_IF\_NO\_HOT\_SPOT\_VISIBLE$). The final calculation of the probability of smoking, $P(smoking)$, is given as the number of puffs taken minus the minimum number of inhalations required to reach a probability of 0.5 or above (MIN_NUM_PUFFS) minus a time decay ($TIME\_DECAY$) according to the time since a hot spot was last detected by the camera (4.14). To ensure the resulting value is between 0 and 1, we finally apply the sigmoid function to it.

$$\begin{aligned} P(\text{smoking}) = \\ \text{number of puffs taken} - MIN\_NUM\_PUFFS \\ -TIME\_DECAY * \text{time since a hot spot was last seen} \end{aligned} \tag{4.14}$$

Note that, despite the way the algorithm is described in this document, it can be optimized for running in real time by storing the value of the variable p calculated in the previous frame, therefore avoiding the necessity of iterating through the array of distances every frame. The algorithm would also need to receive as input the last distance measured between the mouth and closest hot spot, the last frame where an hot spot had been seen and the last time the subject took a puff.

Table 4.5 details the configurable parameters of the smoking classifier. The default values were chosen by manual observation of smokers in our dataset, so the values can be refined for more accurate predictions.

We tested the classifier in the scenario explained in 4.5.1 of a single smoker, a RGB+Thermal camera and multiple particle sensors. The output of the classifier can be visualized in Fig. 4.9. At $time = 87s$ and $time = 160s$, the subject placed the cigarette very close to the mouth, so the algorithm considered the movement has a puff. To guarantee classification robustness when puffs are erroneously registered, it is mathematically guaranteed that the smoking classifier will never predict a probability above 50% until more than MIN_NUM_PUFFS puffs are recorded.

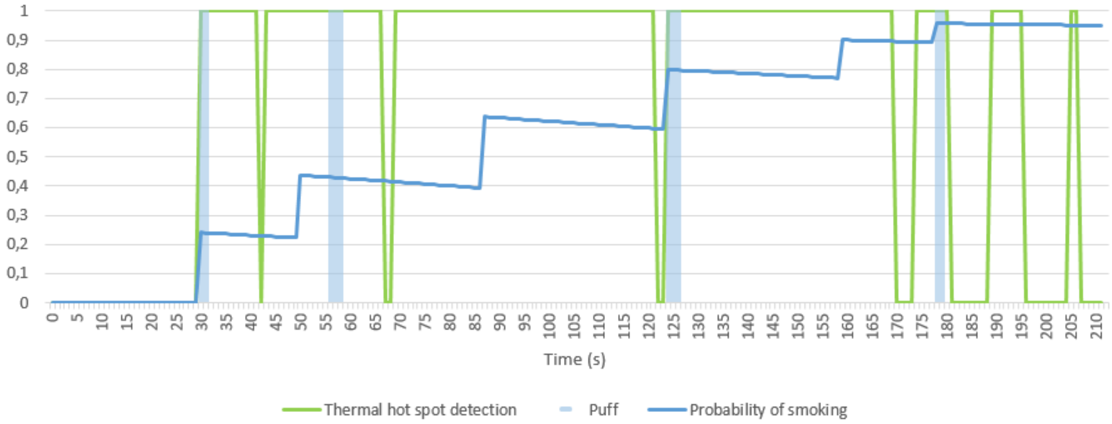


Figure 4.9: Output of our smoking classifier during a smoking experiment inside a vehicle.

Finally, we should note that two big limitations of our algorithm are electronic cigarettes and heated tobacco, which may not produce enough heat for detection, and smokers that keep the cigarette in their mouth while smoking. Therefore, in a smoking detection system, and depending on the use case, it may be wise to trigger some sort of alert when a hot spot is detected, even if the classifier outputs a small value of confidence.

4.6 Respiratory rate estimation

As mentioned in section 2.7.2, it has been demonstrated that thermal imaging can be used to measure the respiration, taking advantage of the variations in temperature that occur during breath in the nostrils and philtrum region. In our monitoring system, the camera is positioned only slightly below the occupant's face, so, although using the temperature of the nostrils allows for a more robust estimation of the respiratory pattern [CJMBB17], our system uses the philtrum as the RoI for this purpose. In order to locate the RoI, we first apply the face detection algorithm described in section 4.3 and the facial landmark detection algorithm described in section 4.4.

Parameter	**Default**	**Description**
Time decay	0.005	Controls how much the smoking probability should be reduced since the last time an hot spot was seen near a person.
Minimum number of puffs	2	Minimum number of times an hot spot needs to be moved closed to a person's mouth in order to allow for a smoking probability above 50%. Studies show that, to smoke a full cigarette, at least 85% of smokers take more than 5 puffs [Gra67].
Puff start distance threshold	0.25	During a puff, if the distance between the mouth and the closest hot spot divided by the face diagonal is equal to or above this threshold, the puff will be considered as terminated.
Puff end distance threshold	0.3	When not taking a puff, if the distance between the mouth and the closest hot spot divided by the face diagonal is equal to or below this threshold, it will be considered that the subject is starting a puff.
Minimum time between puffs	3	Minimum time required between the start of two different puffs, otherwise a single puff will be considered.
Maximum puff time if no hot spot is visible	3	During an inhalation, if the hot spot is no longer visible for this amount of time, the inhalation will be considered as ended.
Maximum size of the time window to be analyzed	300	Defines the amount of time in the past, and therefore number of frames, that should be processed by the algorithm.

Table 4.5: Configurable parameters of the smoking classifier.

The averaged temperatures extracted from the ROI are considerably noisy. To allow for extraction of the respiratory frequency, the signal is passed through an elliptic third-order passband filter which extracts frequencies between 0.1Hz and 0.85Hz that we consider, based on previous work, to be the expected breathing rate [CBBJM18]. We also define a minimum time window of 1 minute, based on the same work, in order to output an estimation, otherwise we consider there is not enough data to reliably predict the respiratory rate. Fig. 4.10(c) shows an example of the mean temperature of the RoI extracted during one minute from the video of one of the subjects in our database, performing very small movements. The chart also compares the filtered signal with a moving average, where the latter is visibly affected by high-frequency noise and a low-frequency temperature variation with a period of ∼40s and positive peaks at ∼15s and ∼55s.

To track the region of interest three different approaches have been attempted. First, a polygon is defined as RoI, using the following points obtained from the landmark detector:

1. Top left of the mouth
2. Bottom left corner of the nose
3. Bottom right corner of the nose
4. Top right of the mouth

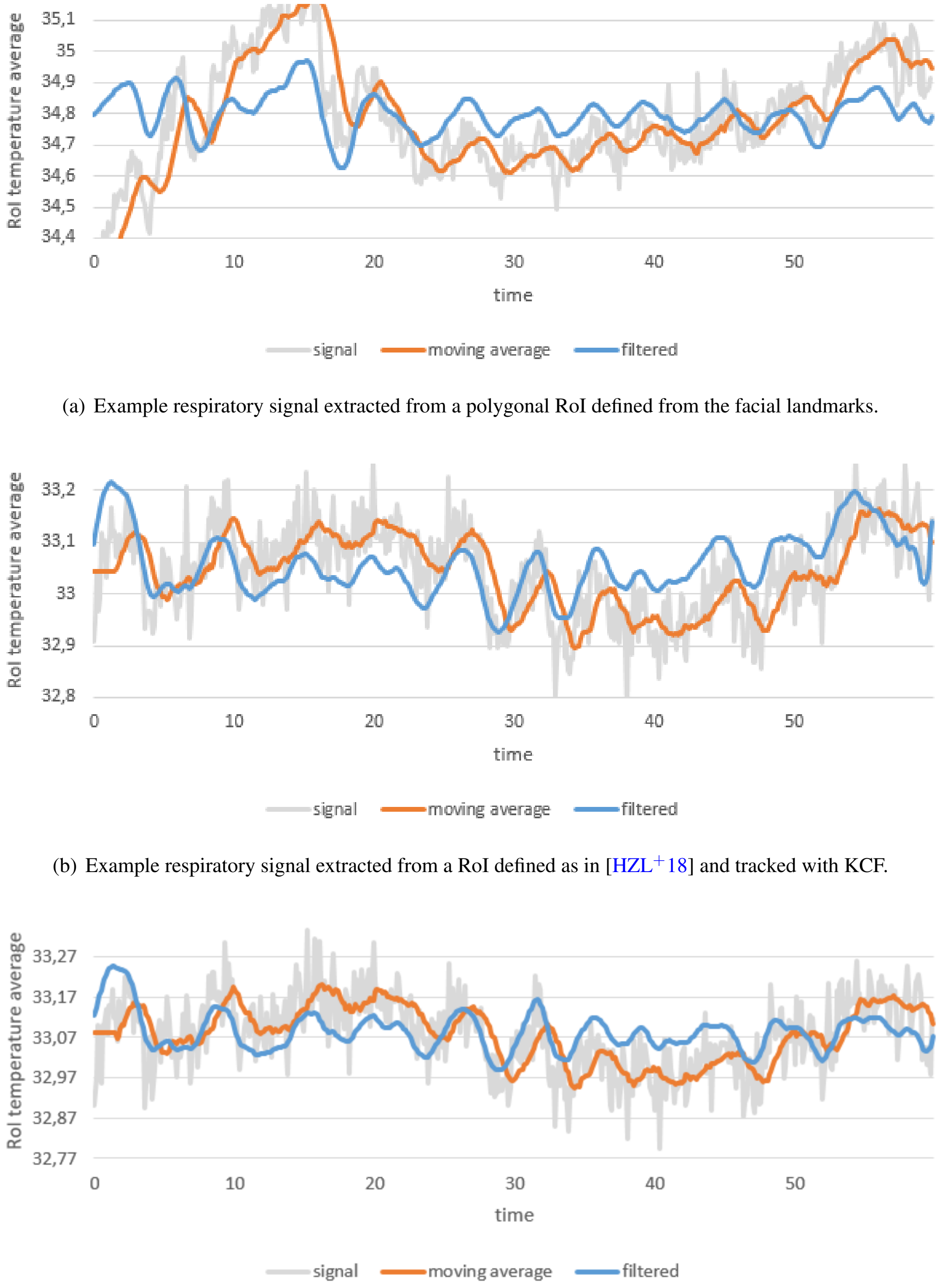
To account for small deviations in the prediction of the landmarks, this polygon is dilated with an offset corresponding to 2% of the diagonal of the facial bounding box. Unlike the other two approaches, here the RoI location is reset according to the predictions of the facial landmark detector, without any tracking between frames.

On a different approach, a rectangular RoI was defined using the points suggested in [HZL+18] (left corner of the mouth, center of the nose and right corner of the mouth), and tracked using Kernelized Correlation Filters (KCF) [HCMB15]. However, one problem with the choice of this RoI is that it can include temperatures from an open mouth. This is not negative in the context of the work of Hu et al. where the subjects are sleeping, but is not suitable for our use case, where the mouth can be open, for instance, if the occupant is talking to someone else.

The third approach uses the same tracker, but the selected RoI is a smaller rectangle. The top limit is defined by the bottom center point of the nose, extended by 10% of the facial height so that the nostrils are included when visible. The bottom is delimited by the top of the mouth with a negative offset equal to 3% of the facial height, to avoid capturing the mouth. Left and right boundaries are defined by subtracting and adding 12% of the facial width to the horizontal coordinate of the center of the nose.

Despite the absence of ground truth information, a visual analysis of Fig. 4.10 allows one to conclude that using a RoI defined by facial landmarks without tracking causes higher levels of noise than a solution that uses a tracker between the consecutive frames. Deciding between the RoI defined in [HZL+18] and the third approach is difficult due to the lack of ground truth. However, considering that the larger RoI may include the mouth region, we decided to use the smaller one in our system, which is likely to be more robust to subjects opening and closing their mouth for actions other than breathing.

In many situations, it is not possible to extract the respiratory rate due to occlusion or movement of the subject's face. To account for this, we propose an algorithm (2) to automatically handle failures of the tracker and avoid estimating the respiratory rate if the philtrum RoI is not visible. It takes into account the possibility of no face being detected due to a failure of the face detector, but if the tracker fails to update and the facial bounding box is not available, then all RoI temperature history is discarded and the process is restarted as soon as a face is visible again. A restart also happens if the subject is not facing the camera. In our experiments, we verify this by measuring the distance between the center of the nose to the horizontal limits of the bounding box, and ensuring it is contained in the middle 25% of such pixels. However, this could be improved by using more complex pose estimation methods [MCT09]. When restarting the tracker, its change in position can cause a big change in the temperature of the RoI in comparison to the previous



(a) Example respiratory signal extracted from a polygonal RoI defined from the facial landmarks.

(b) Example respiratory signal extracted from a RoI defined as in [HZL+18] and tracked with KCF.

(c) Example respiratory signal extracted from a RoI defined from facial landmarks corresponding to the top of the mouth and bottom corners of the nose. The RoI is tracked with KCF.

Figure 4.10: Example respiratory signals using different RoIs, before and after performing a moving average and the application of an elliptic filter.

frame. To account for this, we store an offset value that records the difference in temperature between the frame before and after the reset. This offset is then added to all subsequent temperature calculations, to avoid big variations in the signal caused by tracker resets.

Algorithm 2 Update and validate philtrum tracker

```
Require: frame = input image
  face = facial bounding box (if detected) and landmarks
  offset = value used to avoid big temperature changes due to tracker resets, initially 0
  try to update tracker position
  if tracker reported failure then
    $trackerSuccess \leftarrow false$
  else
    $trackerSuccess \leftarrow true$
    $roi \leftarrow$ new tracker position
  end if
  if face was detected then
    if subject is not facing forward then
      return $false$
    end if
  end if
  if $trackerSuccess == true$ then
    $p \leftarrow$ bottom center point of the nose
    $roiCenter \leftarrow (roi_x + \frac{roi_{width}}{2}, roi_y + \frac{roi_{height}}{2})$
    $trackerToNoseDistance \leftarrow \sqrt{(roiCenter_x - p_x)^2 + (roiCenter_y - p_y)^2}$
    if $trackerToNoseDistance > MAX\_TRACKER\_NOSE\_DISTANCE$ then
      $trackerSuccess \leftarrow false$
    end if
  end if
  if $trackerSuccess == false$ then
    if face was not detected then
      return $false$
    else
      $previousMean \leftarrow$ mean of temperatures in $roi$
      restart tracker using facial landmarks
      $roi \leftarrow$ new tracker position
      $newMean \leftarrow$ mean of temperatures in$roi$
      $offset \leftarrow offset - (newMean - previousMean)$
    end if
  end if
  return $roi$
```

Respiratory rate is only estimated if the algorithm 2 succeeds for at least MIN_TIME_WINDOW seconds and if the face of the subject is being detected. The filtered signal is then decomposed into frequencies by extracting the Power Spectral Density (PSD), similarly to the work in [CBBJM18]. Fig. 4.11 shows the extracted PSD and FFT (for comparison) from the example signal in Fig. 4.10(c).

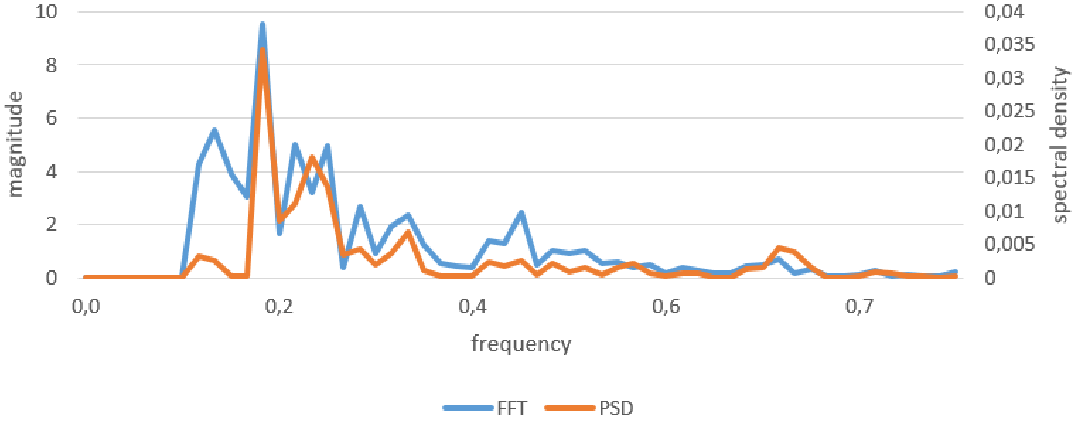


Figure 4.11: Example of extracted frequencies from a signal of 1 minute (Fig. 4.10(c)) of the philtrum RoI.

Table 4.6 details the configurable parameters of the respiratory rate estimator. With the exception of the minimum time window, the values have been defined manually by observation of results and can be adjusted with further experiments.

Parameter	**Default**	**Description**
Minimum time window	60	Minimum amount of time in seconds tracking the RoI required in order to estimate the respiratory rate.
Maximum time window	180	Maximum amount of time in seconds tracking the RoI that should be considered for the calculation of the respiratory rate. Increasing this value causes variations in respiratory rate during time not to be captured.
Maximum tracker to nose distance	0.05	Maximum distance allowed between the center of the tracker and the bottom center point of the nose before the tracker is reset. This value is defined in relative coordinates to the size of the image.
Maximum head rotation	0.2	Percentage of yaw allowed in relation to the camera while measuring respiration. This value is relative to the width of the facial bounding box and defines how much rotation is allowed in each direction.

Table 4.6: Configurable parameters of the respiratory rate predictor.

Experiments were run using this algorithm in our test dataset. None of the subjects in our database was asked to remain still or told that we would try to extract the respiratory rate from their recordings. Therefore, a considerable amount of images from our dataset contain head movement or even occlusion of the nasal area, and there is the need of an algorithm that understands when the respiratory rate can be measured. By analyzing the behavior of the tracker in our images, it is possible to visually conclude that it is able to handle large yaw movements and restart the process. However, when there is a considerable movement of the face, the extracted signal becomes too noisy, as can be observed from the charts in Appendix C. The respiratory rates obtained from the first successful reading for each subject are reported in Table 4.7.

Subject	**End time of measurement (s)**	**FFT argmax (Hz)**	**PSD argmax (Hz)**	**FFT max**	**PSD max**
31	145	0.183	0.283	13.1	0.089
32	84	0.200	0.200	83.6	3.074
33	380	0.183	0.283	9.8	0.044
34	144	0.150	0.150	75.3	1.314
35	-	-	-	-	-
36	92	0.183	0.333	13.0	0.091
37	66	0.217	0.233	49.3	1.899
38	-	-	-	-	-

Table 4.7: Results of the respiratory rate estimator in our test dataset. Only the first successful measurement is reported for each subject and a maximum window size of one minute is defined. The algorithm was unable to predict the respiratory rate for subjects 35 and 38. The third and fourth columns report the respiratory rate calculated using, respectively, FFT and PSD. The last two columns report the magnitude and spectral density for these frequencies.

It is difficult to interpret the results without ground truth information for comparison. However, it is known that the average respiratory rate of a healthy adult at rest lies between 0.2Hz and 0.3Hz [Gan12], so we expect to obtain values in that range. The results show that this verifies for 2 subjects with FFT and 4 subjects with PSD, from a total of 6 subjects for which the algorithm considered that the prediction was valid. The runtime of the algorithm is around 8ms per face.

4.7 Emotion recognition

As previously mentioned in chapter 2, different facial expressions cause different variations in the temperature of different areas of the facial region that may be captured by thermal cameras. In order to properly register those variations, the dataset being used must contain spontaneous instead of posed images and with a considerable waiting time between different expressions [NKCL14]. We cannot rely on this valuable temperature information in the test case of this dissertation, because, due to time constraints, our custom dataset contains only posed facial expressions. However, we can experiment developing a model and testing it using data from NVIE and compare its performance when applied to our scenario where most of the facial images are of lower resolution due to the distance of the subjects to the camera.

One of the limitations of the NVIE dataset is the fact that, although the images are accompanied by a file with the corresponding temperature matrix, this file is corrupt for a significant amount of frames and cannot be relied upon. To avoid discarding a large amount of frames, we try to recover the temperature matrix by converting each corrupt frame to grayscale and considering that the pixel with the lowest value has a temperature of 10ºC and the pixel with the highest value is at a temperature of 37.5ºC (the maximum temperature should be the corner of the eyes and should be very close to the one of the human body). Previously, we had tried reverting the colored image by, for each pixel, finding the index of the palette that minimizes the difference between colors, but this provided worse results than converting directly to grayscale, probably because we did not have access to the right palette. Another limitation of the database is that the images are taken by two totally different cameras and are not well aligned.

The solution proposed in [NRM09] for valence prediction using Deep Boltzmann Machines does not take advantage of the full size of the NVIE dataset in its current version. Instead it only uses images from 32 subjects for training and 6 for testing, from a total of 103 subjects that are contained in the current version of the spontaneous NVIE database. Therefore, there is the possibility of trying to estimate valence using more data and possibly larger models to adapt to it. Furthermore, Salakhutdinov et al. use frames around apex from the spontaneous database but do not take advantage of posed images to help during training.

Considering the possibility of training a model with more data, including our custom-made dataset (chapter 3), we built a small deep learning model based on convolutional layers that is applied to the whole face of the subject after a preprocessing step of alignment and equalization. The model, summarized in Table 4.8, was obtained after tuning of the number of layers, type of activations, size of filters, number of hidden units, optimizer settings and regularization parameters, with the objective of decreasing the validation loss.

Layer	**Filters/Units**	**Parameters**
Conv2D	32	kernel = 5
MaxPooling		size = 2
Conv2D	32	kernel = 3
MaxPooling		size = 2
Dense	32	
Dropout		amount = 0.3
Dense	16	
Dropout		amount = 0.1
Dense	1	

Table 4.8: Architecture of the thermal imaging valence predictor.

The model was trained with images gathered from the NVIE spontaneous database. We extract not only the apex frame, but also up to 12 frames for each facial expression performed by each subject as done in previous work [WHG+14]. Additionally, we also feed the neural network with images from the posed database, to increase the amount of data, although we only use spontaneous data for validating and testing the accuracy of the predictions. We split the dataset in two, where one fifth is allocated for testing purposes, leaving us with data from 95 subjects for training (spontaneous and posed) and 20 for testing (spontaneous). Cross-validation is performed using the Monte Carlo method and setting aside 19 of the 95 subjects for validation, which means the training process occurs with data from 76 subjects. After training the model multiple times, we choose the one that minimizes the validation loss and evaluate it in images from the 20 test subjects, obtaining an accuracy of 80.1%, kappa coefficient of 0.469 and F1-score of 0.596, with an average inference time of 5.5ms (full results in Appendix D). Although we report the test score of the model with lower validation loss, we also experimented averaging the test results of all models trained during cross-validation and obtained an average accuracy of 79.7%, kappa of 0.454 and F1-score of 0.578 with a standard deviation of 0.082, 0.095 and 0.054, respectively for each metric.

To estimate a decision threshold that maximizes the F1-score for predicting positive/negative valence, we employ the basin hopping global optimization method. Note that the F1 function is not guaranteed to be convex and therefore should not be maximized by local optimization.

The relation between the model predictions and the ground truth labels is presented as a confusion matrix in Table 4.9. Additionally, Table 4.10 compares our solution to previous work on the same database, demonstrating that our model outperforms the others in all listed metrics.

		prediction	
		low	**high**
truth	**low**	730	158
	high	64	164

Table 4.9: Confusion matrix of the valence predictor in NVIE spontaneous database.

Method	**Accuracy (%)**	**Kappa**	**Average F1-score**
CNN with NVIE	80.1	0.469	0.596
DBM with mixDataset [WHG+14]	68.2	0.277	0.503
StatisticFea+PCA+SVM [WHG+14]	61,0	0,190	0,466
2D-DCT+PCA+SVM [WHG+14]	62,9	0,174	0,438
GLCM+PCA+SVM [WHG+14]	55,3	0,049	0,366
Laplacianfaces [HYH+05]	59,8	0,062	0,347
Gabor+PCA+SVM [WHG+14]	55,3	0,007	0,324
Sparse representation [WYG+09]	69,5	0,106	0,282
Pixels+PCA+SVM [WHG+14]	51,3	-0,090	0,251

Table 4.10: Comparison between our valence predictor and previous work.

We also tested the performance of the model in our custom dataset with images captured inside vehicles. Face detection on the input temperature matrix is performed using our YOLOv3 thermal face detector (section 4.3). Then, we use the output of our facial landmark detector (section 4.4) to apply a transformation on the image [Ume91] so that the face becomes aligned with the camera. The prediction for each face takes an average time of 9.57ms, being 15% of that time dedicated to the alignment process. Our dataset contains facial images of significantly lower resolution compared to NVIE's and the facial expressions that the subjects expressed are all posed. The model is unable to predict correctly and has a negative kappa score of -5.15%, an accuracy of 46.51%, and a F1-score of 46.23%, which means the model is not adapting well to the images from our dataset. Even though we apply the same preprocessing algorithm, NVIE's images do not contain temperature information so we had to estimate the temperatures, which together with the low resolution of our images may justify the bad performance of the predictor. Therefore, to adapt the algorithm to the new data, we fine-tuned the network using the pre-trained weights on NVIE posed+spontaneous and training it with images from 24 subjects from our dataset. We performed cross-validation using data from 6 subjects and the best model was selected and tested on the remaining 8 subjects (Appendix D). Despite the low resolution, the model performs better in our database than in NVIE's, with a kappa score of 0.345, an accuracy of 67.8% and a F1-score of 0.72 in valence prediction. We should observe, however, that these results are subject to some variance and more data is required to obtain results that are more statistically meaningful, since our test dataset contains only 87 images of positive or negative valence. The resulting confusion matrix is reported in Table 4.11.

		prediction	
		low	**high**
truth	**low**	23	11
	high	17	36

Table 4.11: Confusion matrix of the valence predictor in our custom database.

4.7.1 Combining RGB and Thermal

We compare the performance of our model against a different model trained in RGB images from a Facial Expression Recognition Challenge organized by Kaggle in 2013 (FER2013) [Kag13], with a dataset containing 28709 images of the same 7 different types of facial expressions we considered in our custom-made dataset. We use the model proposed by Arriaga et al. [AVTP17], which is inspired in the Xception model [Cho16] and obtains an accuracy of 66% in FER2013. When predicting valence with this network in the NVIE spontaneous database, we get a weighted accuracy of 81.6%, a kappa of 63% and a F1-score of 77.8%.

As previously mentioned in this document (section 2.4), a combination of different image modalities can contribute to an increase in the accuracy of predictions when compared to a single-modality solution. We experimented mixing the output of two separate classifiers, one for RGB and one for Thermal, and applied the stacked ensemble technique by concatenating feature-rich layers from each separate model (last convolutional layer from the RGB-based model and penultimate dense layer from the Thermal-based model) and appending fully-connected layers to be responsible for the final decision (Fig. 4.12). The stacked part consists of 2 fully connected layers followed by another one with 7 units, corresponding to the six facial expressions to be predicted plus "neutral", and using the softmax activation function. The other two dense layers contain 32 and 16 units and are followed, respectively, by a dropout of 0.3 and 0.1 for regularization.

For the training of this ensemble, we start by freezing the layers of the underlying models. This helps avoiding overfitting in our case, where most of the weights are already adjusted to their

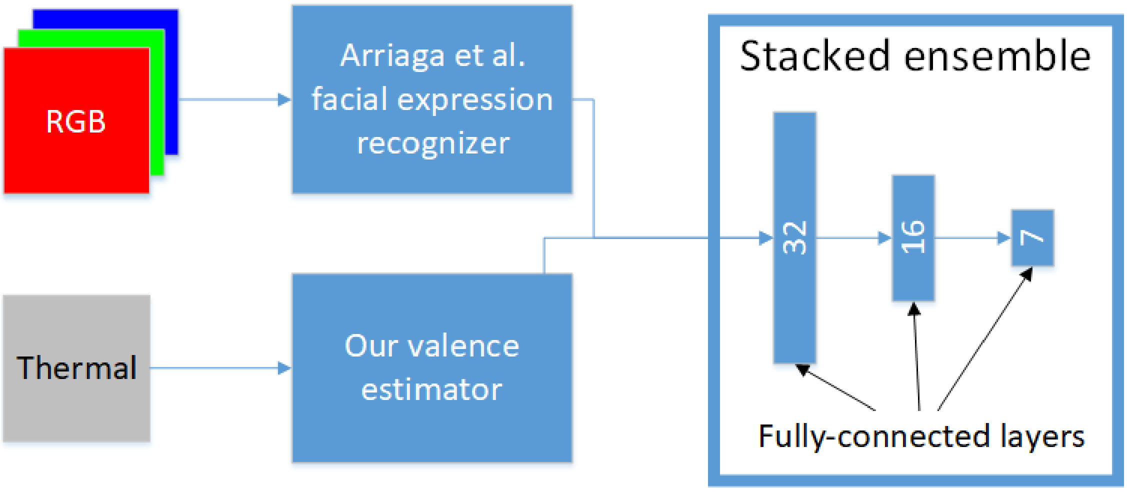


Figure 4.12: Stacked ensemble of a RGB facial expression recognizer and a thermal valence predictor combined to predict facial expressions in RGB+T images.

final purpose, there is a large amount of parameters (mainly in the RGB network) and the size of the dataset is not very large. If we had a larger dataset or a smaller number of parameters, we could perform fine-tuning instead of freezing the transfered layers, or even training the full network from scratch [YCBL14].

Table 4.12 reports the confusion matrix of the ensemble when tested on NVIE spontaneous database predicting valence and Table 4.13 reports the same data for the prediction of six facial expressions and "neutral". Additionally, Table 4.14 compares the performance of the ensemble with the RGB model proposed by Arriaga et al. in different datasets.

		prediction	
		low	**high**
truth	**low**	826	62
	high	28	200

Table 4.12: Confusion matrix of the RGB+T ensemble predicting valence in NVIE spontaneous database.

		prediction						
		neutral	**happiness**	**disgust**	**fear**	**surprise**	**anger**	**sadness**
truth	**neutral**	20	6	13	3	24	37	9
	happiness	0	194	1	10	11	12	0
	disgust	11	29	36	0	28	100	12
	fear	29	34	28	14	15	86	10
	surprise	19	32	28	4	82	48	15
	anger	30	8	5	13	49	100	23
	sadness	21	0	17	0	20	114	56

Table 4.13: Confusion matrix of the RGB+T ensemble predicting facial expressions in NVIE spontaneous database.

By analyzing the results in Table 4.14 and Appendix D, it is observable that the ensemble brings no improvement against the RGB-only solution in our dataset, even after fine-tunning the network to our data after a pre-train on NVIE. There may exist improvements in terms of valence prediction, but more test data would be required to extract more statistically significant results. It is also clear that the addition of temperature information is more beneficial for facial expression recognition in the spontaneous case than in posed images. This difference in usefulness of the thermal information between posed and spontaneous is expectable since temperature changes occur in the human face with true facial expressions, a property that makes thermal imaging good for differentiating between true and fake facial expressions [GWWJ15] [GNWJ17].

		facial expression recognition			**valence estimation**		
test dataset	**model**	**weighted acc.**	**kappa**	**f1-score**	**acc.**	**kappa**	**f1-score**
NVIE spontaneous	**T**	-	-	-	80.1%	0.469	0.596
	RGB	25.4%	0.149	0.185	81.6%	0.634	0.778
	RGB+T	34.5%	0.231	0.323	89.6%	0.735	0.8050
NVIE posed	**T**	-	-	-	76.4%	0.407	0.558
	RGB	28.0%	0.181	0.219	78.6%	0.574	0.738
	RGB+T	30.0%	0.170	0.267	88.8%	0.704	0.778
Our dataset	**T**	-	-	-	67.8%	0.345	0.720
	RGB	58.9%	0.406	0.6132	87.9%	0.683	0.918
	RGB+T	51.6%	0.365	0.574	88.9%	0.754	0.916

Table 4.14: Comparison between a thermal-only model, a RGB-only model and a RGB+T ensemble for facial expression recognition and valence estimation in multiple datasets.

Chapter 5

Conclusions and Future work

We have developed a vehicle occupant monitoring system using thermal images and focused on face detection, smoking detection, emotion recognition and respiratory rate monitoring. This chapter discusses the results for the use cases, extracting conclusions and presenting future work for each. It also presents new ideas, that surged during the development of this work, and that could improve the results in a future work. Those improvements are either in terms of accuracy/speed or to extend the functionalities of the system.

First, in order to improve the results of the trainable algorithms such as face detection, landmark prediction and facial expression recognition, more data can be gathered from different subjects. One of the objectives of our work was to understand the possibilities of thermal imaging for vehicle occupant monitoring and this includes having a database of images inside the vehicle to test the algorithms in a real scenario. However, more data could be obtained to better train the algorithms and to no longer depend on datasets that have limited licensing (non-commercial) and cannot be integrated into a final product.

For face detection, the results are satisfactory for the use cases, and we believe that one major contributor to the high accuracy is the possibility of overfiting on a particular vehicle model. To the best of our knowledge, no previous research obtained such high score in AP_{50} performing face detection in thermal images. Furthermore, one possible justification to why the detector does not score higher in the AP metric is due to the fact that the ground truth is generated automatically and the limits are not perfectly defined. Overall, we believe that our work presents new ways of transferring existing algorithms from RGB to thermal and demonstrates good results in a vehicle scenario. Yet there are some ideas that could be experimented but could not be attempted due to time constraints. One of the models presented in section 4.3 uses a color palette to map the temperatures to different colors and, although the results were not promising compared to our other approaches, there is the possibility of experimenting different palettes and see how they impact the prediction accuracy. Additionally, at the moment, the face detection algorithm here presented for thermal images relies on pretrained weights fitted to RGB images from ImageNet and adapted to work with thermal images. Instead, we could retrain the whole YOLOv3 neural network with the grayscale version of those images, so that the network is prepared from scratch to accept input in a single-channel format, therefore removing the necessity of readjusting the weights from a three-color system to single-color. Furthermore, experiments can be conducted to understand how the input resolution affects the accuracy of the predictor, and what is the expected trade-off between inference speed and quality of predictions.

After performing face detection, it is possible to align the visible and thermal images by calculating the distance to the camera using an estimation of the size of each subject's face. However, better alignment can be achieved using non-rigid transformations [MZMT15] and therefore provide a better input for other algorithms in the processing pipeline that take advantage of a good face alignment. Another step that should be performed is tracking of different occupants, which is critical for the algorithms that follow in the pipeline. In the context of this dissertation, this step was not considered, and the smoking and respiratory rate estimation algorithms assume no more than one subject is visible at a time, but a tracking algorithm should be added to the pipeline for the monitoring system to properly handle multiple faces.

Regarding the smoking classifier, we have proposed an algorithm to detect smoking activities and avoid false positives. Additionally, we concluded that using thermal imaging to detect heated tobacco is possible but not trivial, because the temperature of the device is not high enough to be distinguished from the background, especially in hot environments. Some additional experiments can be conducted, namely to verify the correlation between the heat of a cigarette and the moment a puff is taken. Visually, we noticed that when the smoker takes a puff, cigarettes tend to increase slightly in temperature. This correlation could be exploited to increase the robustness of the smoking classifier. Furthermore, combining the thermal information with a RGB cigarette detector using an ensemble could also increase the robustness of the classifier, although a cigarette detector based on visible light could present problems in low or varying light conditions.

In terms of respiratory rate estimation, the results are not fully satisfactory. Due to the lack of ground truth during the recordings for our database, it is not possible to extract objective measures of the correlation between the real breathing rate and the predictions of the algorithm. Running our algorithm in the test set of our database we observe that it is able to properly recognize RoI tracking failures and recover from them when possible. However, by visually analyzing the history of temperatures in subjects from the test set, it is noticeable that there is a considerable amount of noise. This appears to be related to subject movement, and cleaner readings can be obtained when the subject's face is standing still (Fig. 4.10(c)). We evaluated the results subjectively, by considering an average adult breathing rate between 0.2Hz and 0.3Hz. This occurred for 4 out of 6 subjects when extracting the maximum frequency with the PSD. More research is required to evaluate the method with appropriate ground truth data. One possible way of improving the results is to combine multiple time segments of valid temperature readings, then averaging the multiple extracted PSD to obtain a new one with less variance.

For emotion recognition, we developed a model that measures valence using thermal images with an accuracy above previous work. However, we believe that using this valence information to perform automated responses in a vehicle would require fusing the data with other sensors, since there is an error rate in our dataset of 32.2%. We have also tested this model on the NVIE database, obtaining an accuracy of 80.1% in spontaneous expressions. Further tests need to be employed with a database of vehicle occupants performing facial expressions in a spontaneous manner instead of posed, to better assess the quality of the algorithms. Also, other solutions can be studied that take advantage of temporal information, that is, comparing onset with apex frames and using the temporal difference between facial features to estimate the facial expression of the subject. Moreover, for the RGB+T ensemble, more experiments need to be made to assess which layers of the original models should be frozen and which could be left unfrozen, as well as deciding if the input of the stacked part of the ensemble should be the penultimate layers of each separate model, or if there is an advantage in using information from previous layers from each model to the ensemble. Furthermore, to optimize the facial expression recognition models to the task of valence prediction, the output could be changed to binary and trained and tested as such. In our experiments that include visible light information, the valence prediction accuracy was calculated from the values of the confidence matrix, without tuning the network for the valence prediction task.

References

[Abi07] B Abidi. IRIS Thermal/Visible Face Database. http://vcipl-okstate.org/otcbvs/bench/, 2007.

[AHJ$^+$11] Abbas K. Abbas, Konrad Heimann, Katrin Jergus, Thorsten Orlikowsky, and Steffen Leonhardt. Neonatal non-contact respiratory monitoring based on real-time infrared thermography. *BioMedical Engineering Online*, 10(1):93, 2011.

[ANTV16] Bauyrzhan Aubakir, Birzhan Nurimbetov, Iliyas Tursynbek, and Huseyin Atakan Varol. Vital sign monitoring utilizing Eulerian video magnification and thermography. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2016-Octob:3527–3530, 2016.

[AVTP17] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time Convolutional Neural Networks for Emotion and Gender Classification. *arXiv preprint arXiv:1710.07557*, 2017.

[Bak74] R. R. Baker. Temperature distribution inside a burning cigarette. *Nature*, 247(5440):405–406, 1974.

[BES$^+$11] René Braun, Jens Eichmann, Samer Sabbah, Roland Harig, and Chris R. Howle. Remote detection of liquid surface contamination by imaging infrared spectroscopy: measurements and modelling. In *Optics and Photonics for Counterterrorism and Crime Fighting VII; Optical Materials in Defence Systems Technology VIII; and Quantum-Physics-based Information Security*, page 81890G, 2011.

[BGC17] Arwa M Basbrain, John Q Gan, and Adrian Clark. Intelligent Computing Methodologies. In *Intelligent Computing Methodologies*, volume 10363, pages 71–82. Springer, 2017.

[BGK16] Stephanie L. Bennett, Rafik Goubran, and Frank Knoefel. Adaptive eulerian video magnification methods to extract heart rate from thermal video. *2016 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2016 - Proceedings*, 2016.

[BHGK17] Stephanie Bennett, Tarek Nasser El Harake, Rafik Goubran, and Frank Knoefel. Adaptive Eulerian Video Processing of Thermal Video: An Experimental Analysis. *IEEE Transactions on Instrumentation and Measurement*, 66(10):2516–2524, 2017.

[BHP93] Steven D. Burch, Vahab Hassani, and Terry R. Penney. Use of infra-red thermography for automotive climate control analysis. Technical report, SAE Technical Paper, 1993.

[BLA^{+}13] The Boulevard, Langford Lane, Simon Atkinson, Bob Flitney, Lin Lucas, and Lin Lucas. Thermal imaging helps Ford detect air leaks in vehicle cabins. *Sealing Technology*, 2013(9):2–3, 2013.

[Blu17] Richard Blumenthal. Helping Overcome Trauma for Children Alone in Rear Seats Act. https://www.congress.gov/bill/115th-congress/senate-bill/1666, 2017.

[Bri90] R. J. Brison. Risk of automobile accidents in cigarette smokers. *Canadian Journal of Public Health*, 81(2):102–106, 1990.

[BRSD15] Anushree Basu, Aurobinda Routray, Suprosanna Shit, and Alok Kanti Deb. Human emotion recognition from facial thermal image based on fused statistical feature and multi-class SVM. *Annual IEEE India Conference (INDICON), At New Delhi, India*, pages 1–5, 2015.

[BTK17] Adrian Bulat, Georgios Tzimiropoulos, and United Kingdom. How far are we from solving the 2D & 3D Face Alignment problem? *Iccv*, 2017.

[CBBJ18] Youngjun Cho, Nadia Bianchi-Berthouze, and Simon J. Julier. DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, 2018-Janua:456–463, 2018.

[CBBJM18] Youngjun Cho, Nadia Bianchi-Berthouze, Simon J. Julier, and Nicolai Marquardt. ThermSense: Smartphone-based breathing sensing platform using non-contact low-cost thermal camera. *2017 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2017*, 2018-Janua:83–84, 2018.

[Cho16] François Chollet. Xception: Deep Learning with Separable Convolutions. *arXiv preprint arXiv:1610.02357*, pages 1–14, 2016.

[CJMBB17] Youngjun Cho, Simon J. Julier, Nicolai Marquardt, and Nadia Bianchi-Berthouze. Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. *Biomedical Optics Express*, 8(10):4480, 2017.

[Coh68] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

[CSCG16] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F. Cohn, and Sergio Escalera Guerrero. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, 2016.

[CYN14] Yuen Kiat Cheong, Vooi Voon Yap, and Humaira Nisar. A Novel Face Detection Algorithm Using Thermal Imaging. In *2014 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, pages 208–213, 2014.

[ECMFZ12] Anna Esposito, Vincenzo Capuano, Jiri Mekyska, and Marcos Faundez-Zanuy. A naturalistic database of thermal emotional facial expressions and effects of induced emotions on memory. *Lecture Notes in Computer Science (including* *subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7403 LNCS:158–173, 2012.

[EHDF15] Mohamed A. Elgharib, Mohamed Hefeeda, Frédo Durand, and William T. Freeman. Video magnification in presence of large motions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 4119–4127, 2015.

[Eur16] European Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Coucil on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regu. *VoteWatch Europe*, L119:1–88, 2016.

[Fli16] Flir. Flir thermal imaging cameras allow machines to read human emotions. http://www.flir.co.uk/cs/display/?id=67117, 2016.

[FP10] Jin Fei and Ioannis Pavlidis. Thermistor at a distance: Unobtrusive measurement of breathing. *IEEE Transactions on Biomedical Engineering*, 57(4):988–998, 2010.

[Fre15] Marie (Photonics Media) Freebody. Consumers and Cost Are Driving Infrared Imagers into New Markets. https://www.photonics.com/a57307/, 2015.

[FS95] Yoav Freund and Robert E. Schapire. A desicion-theoretic generalization of online learning and an application to boosting. *Journal of computer and system sciences*, 55(1):23–37, 1995.

[FZMFA14] Marcos Faundez-Zanuy, Jiri Mekyska, and Xavier Font-Aragonès. A New Hand Image Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums. *Cognitive Computation*, 6(2):230–240, 2014.

[Gan12] W. F. Ganong. *Review of Medical Physiology*. Appleton & Lange Norwalk, CT, 2012.

[GCLL10] Q Guo, S Chen, H Leung, and S Liu. Covariance intersection based image fusion technique with application to pansharpening in remote sensing. *Information Sciences*, 180(18):3434–3443, 2010.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[GF13] Travis R. Gault and Aly A. Farag. A Fully Automatic Method to Extract the Heart Rate from Thermal Video. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 336–341, 2013.

[Gir15] Ross Girshick. Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:1440–1448, 2015.

[GM14] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision and Applications*, 25(1):245–262, 2014.

[GNWJ17] Q Gan, S Nie, S Wang, and Q Ji. Differentiating between posed and spontaneous expressions with Latent Regression Bayesian Network. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 4039–4045, 2017.

[GPVPW14] FLIR Systems) Guy Pas (Vice-President Worldwide, Systems Sales. Demand for FLIR thermal imaging technology sees prices drop. https://www.theneweconomy.com/technology/demand-for-flir-thermal-imaging-technology-sees-prices-drop, 2014.

[Gra67] Saxon Graham. Cancer of Lung Related T O Smoking Behavior. *American Public Health Association*, 1967.

[GSMP07] Marc Garbey, Nanfei Sun, Arcangelo Merla, and Ioannis Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Biomedical Engineering*, 54(8):1418–1426, 2007.

[GWWJ15] Q Gan, C Wu, S Wang, and Q Ji. Posed and spontaneous facial expression differentiation using deep Boltzmann machines. *Affective Computing and . . .* , 2015.

[HBD+08] Roland Harig, René Braun, Chris Dyer, Chris Howle, and Benjamin Truscott. Short-range remote detection of liquid surface contamination by active imaging Fourier transform spectrometry. *Optics Express*, 16(8):5708, 2008.

[HBM16] Kian Hamedani, Zahra Bahmani, and Amin Mohammadian. Spatio-temporal filtering of thermal video sequences for heart rate estimation. *Expert Systems with Applications*, 54:88–94, 2016.

[HCE+17] Shawn Hershey, Sourish Chaudhuri, Daniel P.W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 131–135, 2017.

[HCMB15] Joao F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015.

[HS88] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Procedings of the Alvey Vision Conference 1988*, pages 1–23, 1988.

[HYH+05] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong Jiang Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[HZL+18] Menghan Hu, Guangtao Zhai, Duo Li, Yezhao Fan, Huiyu Duan, Wenhan Zhu, and Xiaokang Yang. Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep situation. *PLoS ONE*, 13(1), 2018.

[IGM14] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology*, 51(10):951–963, 2014.

[IMB+17] Stephanos Ioannou, Paul H. Morris, Marc Baker, Vasudevi Reddy, and Vittorio Gallese. Seeing a Blush on the Visible and Invisible Spectrum: A Functional Thermal Infrared Imaging Study. *Frontiers in Human Neuroscience*, 11(November):1–15, 2017.

[ISH+09] Tatsuya Izumi, Hiroaki Saito, Takeshi Hagihara, Kenichi Hatanaka, and Takanori Sawai. Development of occupant detection system using far-infrared ray (FIR) camera. *SEI Technical Review*, 1(69):72–77, 2009.

[JLL+17] Shangjie Jiang, Bin Luo, Jun Liu, Yun Zhang, and Liang Pei Zhang. UAV-based vehicle detection by multi-source images. In Jinfeng Yang, Qinghua Hu, Ming-Ming Cheng, Liang Wang, Qingshan Liu, Xiang Bai, and Deyu Meng, editors, *Communications in Computer and Information Science*, volume 773, pages 38–49, 2017.

[JLM10] V. Jain and E. Learned-Miller. FDDB: A Benchmark for Face Detection in Unconstrained Settings. Technical report, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[JP02] Bryan F. Jones and Peter Plassmann. Digital infrared thermal imaging of human skin. *IEEE Engineering in Medicine and Biology Magazine*, 21(6):41–48, 2002.

[JT94] Jianbo Shi and Tomasi. Good features to track. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*, pages 593–600, 1994.

[JWS+09] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[KA12] Georgia Koukiou and Vassilis Anastassopoulos. Drunk person identification using thermal infrared images. *International Journal of Electronic Security and Digital Forensics*, 4(4):229, 2012.

[KA13] G. Koukiou and V. Anasassopoulos. Face locations suitable drunk persons identification. *2013 International Workshop on Biometrics and Forensics, IWBF 2013*, 2013.

[KA15] Georgia Koukiou and Vassilis Anastassopoulos. Neural networks for identifying drunk persons using thermal infrared imagery. *Forensic Science International*, 252:69–76, 2015.

[Kag13] Kaggle. Challenges in Representation Learning: Facial Expression Recognition Challenge. https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge, 2013.

[KK09] M. Özgün Korukçu and Muhsin Kilic. The usage of IR thermography for the temperature measurements inside an automobile cabin. *International Communications in Heat and Mass Transfer*, 36(8):872–877, 2009.

[KK12] Mehmet Özgün Korukçu and Muhsin Kılıç. Tracking hand and facial skin temperatures in an automobile by using IR-thermography during heating period. *Gazi University Journal of Science*, 25(1):207–217, 2012.

[KMM10] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Forward-backward error: Automatic detection of tracking failures. *Proceedings - International Conference on Pattern Recognition*, pages 2756–2759, 2010.

[KMM12] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2012.

[Koc15] Ed (Eastern Region Sales Director) Kochanek. Are We in the Golden Age of Thermal Imaging? https://www.irinfo.org/10-01-2015-kochanek/, 2015.

[KS14] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[KYNT09] Y Koda, Y Yoshitomi, M Nakano, and M Tabuse. A facial expression recognition for a speaker of a phoneme of vowel using thermal image processing and a speech recognition system. *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 955–960, 2009.

[LAE$^+$16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:21–37, 2016.

[LGG$^+$17] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:2999–3007, 2017.

[LMB$^+$14] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[LWS11] Yanpeng Lv, Shangfei Wang, and Peijia Shen. A real-time attitude recognition by eye-tracking. *Proceedings of the Third International Conference on Internet Multimedia Computing and Service - ICIMCS '11*, page 170, 2011.

[MBP10] Brais Martinez, Xavier Binefa, and Maja Pantic. Facial component detection in thermal imagery. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pages 48–54, 2010.

[MCT09] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[MEDFZ10] Jiří Mekyska, Virginia Espinosa-Duró, and Marcos Faundez-Zanuy. Face segmentation: A comparison between visible and thermal images. *Proceedings - International Carnahan Conference on Security Technology*, pages 185–189, 2010.

[MLS$^+$09] Andrey Makrushin, Mirko Langnickel, Maik Schott, Claus Vielhauer, Jana Dittmann, and Katharina Seifert. Car-seat occupancy detection using a monocular 360° NIR camera and advanced template matching. *DSP 2009: 16th International Conference on Digital Signal Processing, Proceedings*, pages 0–5, 2009.

[MP06] Ramya Murthy and Ioannis Pavlidis. Noncontact measurement of breathing function. *IEEE Engineering in Medicine and Biology Magazine*, 25(3):57–67, 2006.

[MP07] G Mangiaracina and L Palumbo. [Smoking while driving and its consequences on road safety]. *Annali Di Igiene: Medicina Preventiva E Di Comunità*, 19(3):253–267, 2007.

[MZMT15] Jiayi Ma, Ji Zhao, Yong Ma, and Jinwen Tian. Non-rigid visible and infrared face registration via regularized Gaussian fields criterion. *Pattern Recognition*, 48(3):772–784, 2015.

[Nat15] National Highway Traffic Safety Administration (NHTSA). Traffic Safety Facts: 2015, 2015.

[NC17] Victor-emil Neagoe and Serban-vasile Carata. Drunkenness Diagnosis Using a Neural Network-Based Approach for Analysis of Facial Images in the Thermal Infrared Spectrum. In *E-Health and Bioengineering Conference (EHB)*, pages 165–168. IEEE, 2017.

[NKCL14] Hung Nguyen, Kazunori Kotani, Fan Chen, and Bac Le. LNCS 8333 - A Thermal Facial Emotion Database and Its Analysis. In Reinhard Klette, Mariano Rivera, and Shin'ichi Satoh, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8333 LNCS, pages 397–408. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[NMGKBM17] Le Thanh Nguyen-Meidine, Eric Granger, Madhu Kiran, and Louis-Antoine Blais-Morin. A comparison of CNN-based face and head detectors for real-time video surveillance applications. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–7. IEEE, 2017.

[NNGM14] Olegs Nikisins, Kamal Nasrollahi, Modris Greitans, and Thomas B. Moeslund. RGB-D-T Based Face Recognition. *2014 22nd International Conference on Pattern Recognition*, pages 1716–1721, 2014.

[NRM09] Mohammad Norouzi, Mani Ranjbar, and Greg Mori. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, 2009 IEEE(3):2735–2742, 2009.

[Nul17] Jan Null. No Heat Stroke: Deaths by State. http://www.noheatstroke.org/, 2017.

[OM99] David Opitz and Richard Maclin. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligent Research*, 11:169–198, 1999.

[Ots79] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[Peš13] M. Pešek. The temperature fields measurement of air in the car cabin by infrared camera. *EPJ Web of Conferences*, 45:01073, 2013.

[PMP13] Stavros Petridis, Brais Martinez, and Maja Pantic. The MAHNOB Laughter database. *Image and Vision Computing*, 31(2):186–202, 2013.

[Por] Portable pulse oximeter, pulse and oxygen saturation. https://www.quirumed.com/en/portable-pulse-oximeter-pulse-and-oxygen-saturation.html?sid=46315.

[PYC$^+$15] Carina Barbosa Pereira, Xinchi Yu, Michael Czaplik, Rolf Rossaint, Vladimir Blazek, and Steffen Leonhardt. Remote monitoring of breathing dynamics using infrared thermography. *Biomedical Optics Express*, 6(11):4378, 2015.

[Que] Question about FLIR One for Android. http://www.eevblog.com/forum/thermal-imaging/question-about-flir-one-for-android/.

[RDGF15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2015.

[Ren08] Xiaofeng Ren. Finding people in archive films through tracking. *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2, 2008.

[RF17] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:6517–6525, 2017.

[RF18] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[RFN17] Ricardo F Ribeiro, Maria Fernandes, and J R Neves. Face Detection on Infrared Thermal Image. *Signal 2017*, pages 38–42, 2017.

[RGST17] R.I. Ramos-Garcia, E. Sazonov, and S. Tiffany. Recognizing cigarette smoke inhalations using hidden Markov models. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 1242–1245, 2017.

[RHGS17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[RMGE09] Nathan D. Rasmussen, Bryan S. Morse, Michael A. Goodrich, and Dennis Eggett. Fused visible and infrared video for use in wilderness search and rescue. *2009 Workshop on Applications of Computer Vision, WACV 2009*, 2009.

[SCN$^+$16] Marc Oliu Simón, Ciprian Corneanu, Kamal Nasrollahi, Olegs Nikisins, Sergio Escalera, Yunlian Sun, Haiqing Li, Zhenan Sun, Thomas B. Moeslund, and Modris Greitans. Improved RGB-D-T based face recognition. *IET Biometrics*, 5(4):297–303, 2016.

[SF16] Connor Schenck and Dieter Fox. Detection and Tracking of Liquids with Fully Convolutional Networks. *arXiv preprint arXiv:1606.06266*, 2016.

[SL09] Chen Siyue and H Leung. An EM-CI based approach to fusion of IR and visual images. *Information Fusion, 2009. FUSION '09. 12th International Conference on*, pages 1325–1330, 2009.

[SLMT13] Edward Sazonov, Paulo Lopez-Meyer, and Stephen Tiffany. A Wearable Sensor System for Monitoring Cigarette Smoking. *Journal of Studies on Alcohol and Drugs*, 74(6):956–964, 2013.

[SLO11] Paterne Sissinto and Jumoke Ladeji-Osias. Fusion of infrared and visible images using empirical mode decomposition and spatial opponent processing. *Proceedings - Applied Imagery Pattern Recognition Workshop*, 2011.

[SMD10] PARUL SHAH, S. N. MERCHANT, and U. B. DESAI. Fusion of Surveillance Images in Infrared and Visible Band Using Curvelet, Wavelet and Wavelet Packet Transform. *International Journal of Wavelets, Multiresolution and Information Processing*, 08(02):271–292, 2010.

[Sou12] Soufradir EC Resource Center. Uncooled Infrared Imaging : Higher Performance , Lower Costs. https://blog.lk-shop.com/wp-content/uploads/2015/06/WP-Uncooled_IR_Imaging-web.pdf, 2012.

[TCTW13] Hsin Chun Tsai, Chi Hung Chuang, Shin Pang Tseng, and Jhing Fa Wang. The optical flow-based analysis of human behavior-specific system. *ICOT 2013 - 1st International Conference on Orange Technologies*, pages 214–218, 2013.

[TOHH05] L. Trujillo, G. Olague, R. Hammoud, and B. Hernandez. Automatic Feature Localization in Thermal Images for Facial Expression Recognition. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 3:14–14, 2005.

[TWZJ07] Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, 40(11):3195–3208, 2007.

[Ume91] Shinji Umeyama. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

[USB17] USB 3.0 Promoter Group. Universal Serial Bus 3.2 Specification, 2017.

[VJ04] Paul Viola and Michael J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[VKP$^+$17] Michael Vollmer, MÃ Klaus-Peter, et al. *Infrared Thermal Imaging: Fundamentals, Research and Applications*. John Wiley & Sons, 2017.

[WC11] Wen-Chuan Wu and Chun-Yang Chen. Detection System of Smoking Behavior Based on Face Analysis. *2011 Fifth International Conference on Genetic and Evolutionary Computing*, pages 184–187, 2011.

[WGT$^+$18] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65, 2018.

[WHC$^+$10] Pin Wu, Jun Wei Hsieh, Jiun Cheng Cheng, Shyi Chyi Cheng, and Shau Yin Tseng. Human smoking event detection using visual interaction clues. In *Proceedings - International Conference on Pattern Recognition*, pages 4344–4347, 2010.

[WHD$^+$12] Wai Kit Wong, Joe How Hui, Jalil Bin Md Desa, Nur Izzati Nadiah Binti Ishak, Azlan Bin Sulaiman, and Yante Binti Mohd Nor. Face detection in thermal imaging using head curve geometry. *2012 5th International Congress on Image and Signal Processing, CISP 2012*, 47(Cisp):881–884, 2012.

[WHG$^+$14] Shangfei Wang, Menghua He, Zhen Gao, Shan He, and Qiang Ji. Emotion recognition from thermal infrared images using deep Boltzmann machine. *Frontiers of Computer Science*, 8(4):609–618, 2014.

[WJ18] Yue Wu and Qiang Ji. Facial Landmark Detection: A Literature Survey. *International Journal of Computer Vision*, pages 1–28, 2018.

[WLL$^+$10] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010.

[WLSJ13] Shangfei Wang, Zhilei Liu, Peijia Shen, and Qiang Ji. Eye localization from thermal infrared images. *Pattern Recognition*, 46(10):2613–2621, 2013.

[WMR14] S. J. Warden and S. M. Mantila Roosa. Physical activity completed when young has residual bone benefits at 94 years of age: A within-subject controlled case study. *Journal of Musculoskeletal Neuronal Interactions*, 14(2):239–243, 2014.

[WRS$^+$12] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics*, 31(4):1–8, 2012.

[WYG$^+$09] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[YA16] Akihiko Yamaguchi and Christopher G. Atkeson. Stereo vision of liquid and particle flow for robot pouring. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots*, pages 1173–1180. IEEE, 2016.

[YCBL14] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 3320–3328, Montreal, Canada, 2014. MIT Press.

[YHG$^+$14] S. Yue, K. Harmer, K. Guo, K. Adams, and A. Hunter. Automatic blush detection in 'concealed information' test using visual stimuli. *International Journal of Data Mining, Modelling and Management*, 6(2):187–201, 2014.

[ZZF04] Ming Zhao, Jian Zhang, and Renato Figueiredo. Distributed file system support for Virtual Machines in Grid computing. *IEEE International Symposium on High Performance Distributed Computing, Proceedings*, 130(x):202–211, 2004.

Appendix A

Use cases

There are multiple applications of using thermal cameras to capture the interior of vehicles. Some of these can be achieved through other means, such as weight sensors or RGB cameras, but the heat-detecting properties of thermal imaging can improve the accuracy or even replace other sensors.

Occupancy detection. Count and locate vehicle occupants, depending on the camera positioning. An example of application is the possibility of triggering a warning when a young child is on the front seat instead of a back seat or even when a young person is left alone inside the vehicle subject to very hot temperatures; a situation also known as pediatric vehicular heatstroke, which causes the death of an average of 37 children in the U.S., according to data registered since 1998 [Nul17]. To combat this, a senate bill entitled "HOT CARS Act" [Blu17] has been introduced suggesting to enforce all engine passenger vehicles to have a safety alert system for child left alone in vehicles. This is just an example of application of an occupancy detection system, but there are others.

Smoking. Being the activity of smoking banned inside vehicles in many countries, it is useful to be able to register this infraction and possibly display a warning to the driver or to the other passengers. Smoking while driving can cause multiple distractions and it increases the likelihood of an accident by 50%, according to a study from 1990 [Bri90]. This kind of activity not only limits the driver's attention to the road and readiness to react to dangers, but also causes cognitive distraction, in the sense that the brain will be focused in the act of smoking instead of solely in driving.

Facial expressions. Infer the emotional state of the occupants. In the context of passenger transport, one possible application is to understand the degree of satisfaction of the customers. Even though the facial expressions may not be directly related to passenger satisfaction, there might exist a small correlation that can be exploited when a large number of data is available or when specific expressions are registered. For example, if the vehicle occupants express fear frequently, it might be an indicator that the driving style is aggressive. Another useful application is the possibility of the vehicle reacting according to the mood of people, for instance by choosing to play an uplifting music when the occupants express sad emotion.

Breath. Recognize the breath patterns of a person. This might help recording apnea and trigger an alarm for medical help, or detecting stress, which may help understanding the mood of an occupant or even identifying suspects potentially engaged in illegal activities.

Vehicle interior health. Detect some defects in the vehicle habitable. When a defect is registered, the on-board system may trigger a warning to let the driver or the responsible for the vehicle know that there is a problem that requires attention. Multiple types of defects can be detected such as a failure in the rear window defogger, a problem with the front window heater and air leaks.

Liquid spilling. In a professional transportation system, when a passenger spills a liquid on a surface of the vehicle interior, the occurrence can be registered and the company informed that the seats might require cleaning. This use case requires a good positioning of the cameras or appropriate field of views.

Temperatures of the habitable. Automatically adjust the air conditioning system to compensate for temperature differences observed inside the vehicle cabin.

Abnormally high temperatures. Detect and respond to fires inside the cabin. When such an anomaly is detected, having a system that automatically triggers an alert might help solving it quicker in order to avoid structural damage or even saving the lives of the occupants.

Seatbelt usage. Although there are reliable ways of knowing if a passenger is wearing the seatbelt (sensors), it is possible to use a camera in order to detect if the seatbelt is being correctly used (in the right position). A proper use of the seatbelt implies positioning it over the shoulder and across the chest, not under both arms or behind the back. With an optical camera, an improper seatbelt positioning could be identified and the vehicular on-board system could trigger a warning for the occupant to reposition it, avoiding injuries in case of an accident.

Talking on the phone. On the one hand, if a driver is talking on the phone without proper hands-free equipment, the system might register an infraction. On the other hand, if a passenger is making a phone call, the vehicle may respond by turning down the music volume.

Drunk driving. In 2014, in the U.S., 31% of the motor vehicle fatal crashes involved drivers with a blood alcohol content of 0.08g/dL or above, resulting in 9967 fatalities in alcohol-impaired-driving situations [Nat15]. Being able to detect the presence of drunk drivers, opens a door for multiple actions to be taken, according to the context. For example, an alert may be triggered to remember the driver that he should not be driving while drinking. Also, if the driver is working professionally, his company may register an infraction and act accordingly.

Appendix B

Facial landmark detector optimization

nu	**depth**	**MSE** ($*10^{-4}$)
0.002	4	3.682
0.002	5	3.498
0.002	6	3.569
0.002	7	3.587
0.003	4	3.778
0.003	5	3.588
0.003	6	3.975
0.003	7	3.785
0.004	4	3.985
0.004	5	4.172
0.004	6	4.257
0.004	7	3.894

Table B.1: Grid-search optimization of the facial landmark detector using a model for images with glasses and a different one for images without glasses, varying *nu* and tree depth, tested in a randomly picked validation set.

nu	**depth**	**MSE** ($*10^{-4}$)
0.002	4	3.878
0.002	5	3.778
0.002	6	3.724
0.002	7	3.747
0.003	4	3.782
0.003	5	3.735
0.003	6	3.719
0.003	7	3.798
0.004	4	3.744
0.004	5	3.773
0.004	6	3.766
0.004	7	3.811

Table B.2: Grid-search optimization of the facial landmark detector using a single model for both images with and without glasses, varying *nu* and tree depth, tested in a randomly picked validation set.

Appendix C

Respiratory rate estimation results

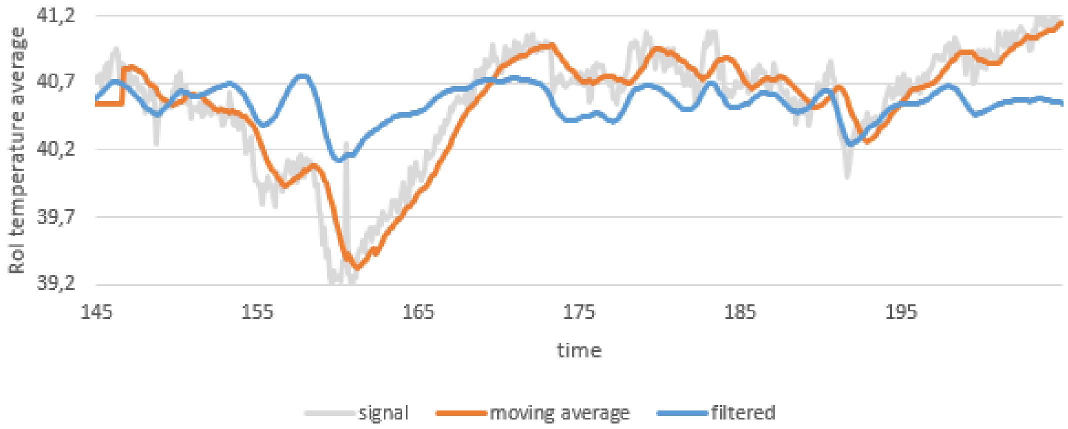


Figure C.1: Temperature of the philtrum RoI during one minute of subject #31 from our database.

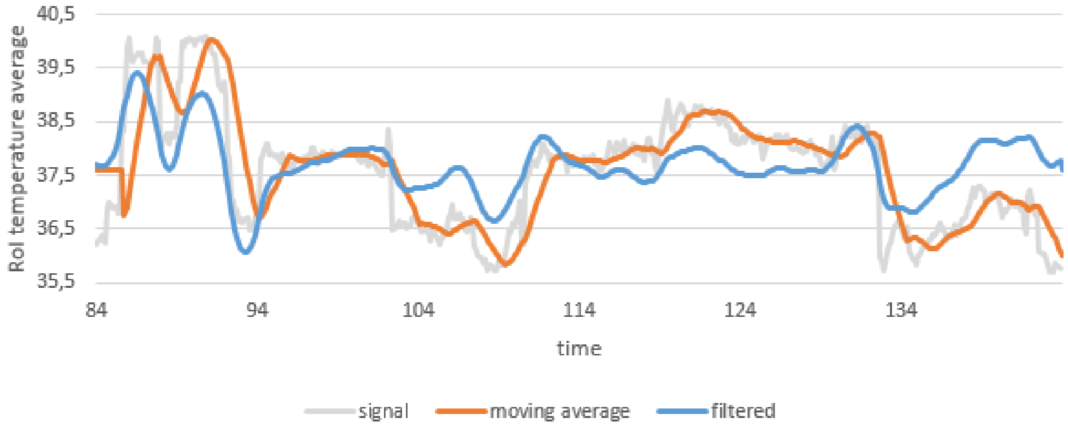


Figure C.2: Temperature of the philtrum RoI during one minute of subject #32 from our database.

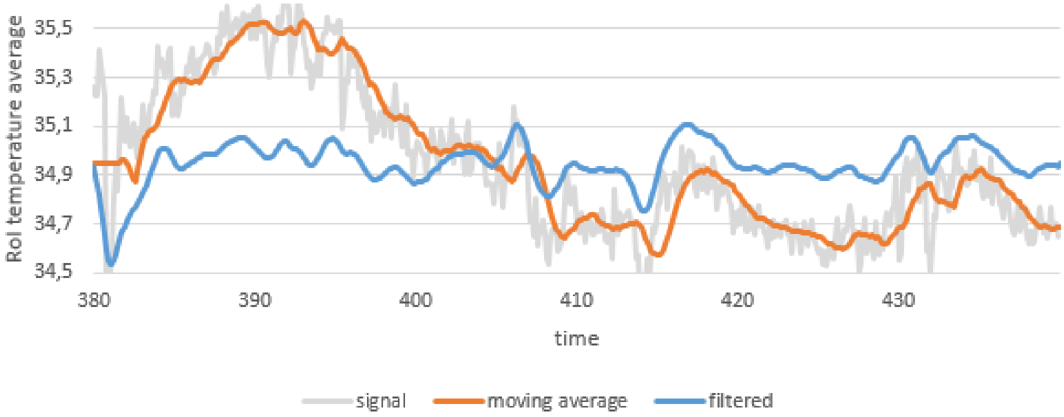


Figure C.3: Temperature of the philtrum RoI during one minute of subject #33 from our database.

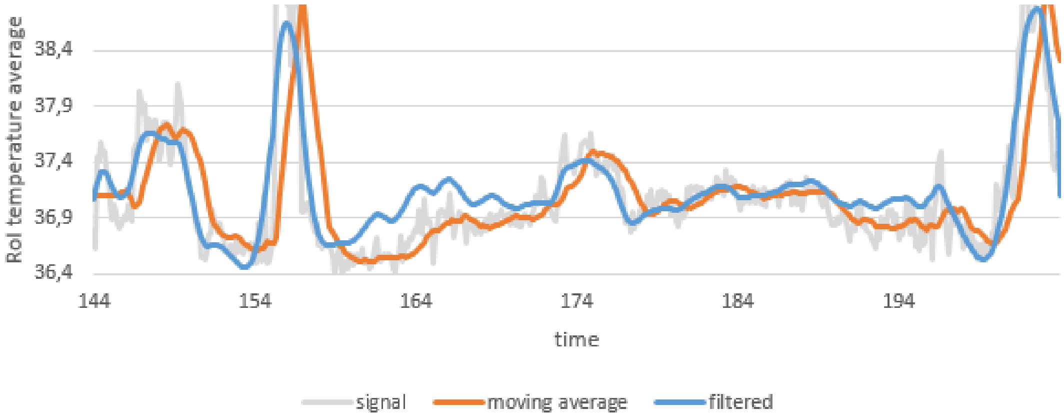


Figure C.4: Temperature of the philtrum RoI during one minute of subject #34 from our database.

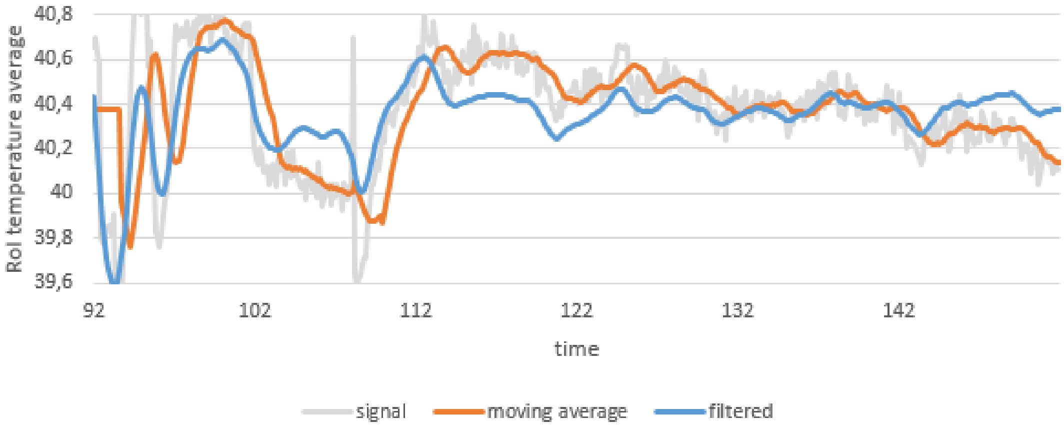


Figure C.5: Temperature of the philtrum RoI during one minute of subject #36 from our database.

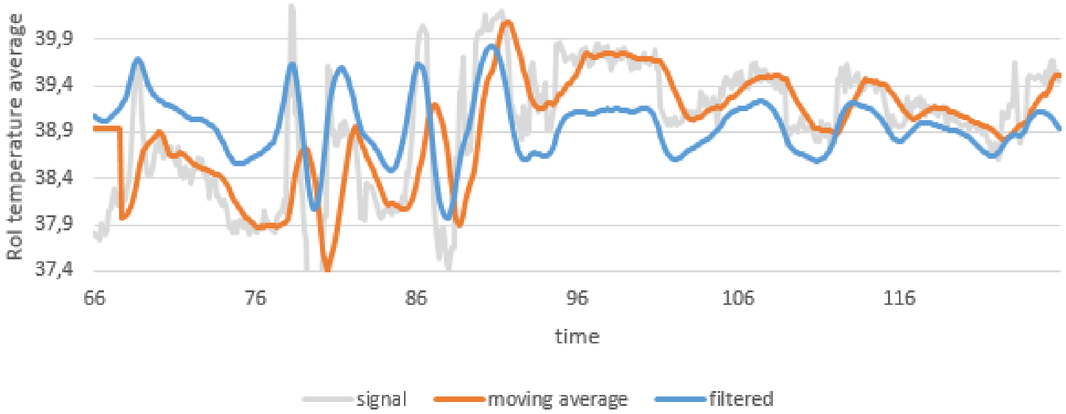


Figure C.6: Temperature of the philtrum RoI during one minute of subject #37 from our database.

Appendix D

Facial expression recognition

iteration	kappa	acc	F1-score	validation loss
0	0,235	58,6%	0,465	0,5146
1	0,545	83,1%	0,653	0,4553
2	0,421	78,8%	0,557	0,4608
3	0,469	80,1%	0,596	0,4259
4	0,540	85,7%	0,628	0,5293
5	0,387	77,7%	0,529	0,525
6	0,542	86,6%	0,621	0,4992
7	0,499	84,8%	0,591	0,4754
8	0,487	84,4%	0,582	0,4929
9	0,413	77,2%	0,557	0,4618

Table D.1: Results of each iteration of the cross-validation with 10 runs to obtain the best model for the valence predictor from thermal images of NVIE spontaneous database. The kappa, accuracy and F1-score metrics are calculated on the test set, but only validation loss was considered for final model selection.

		prediction	
		low	**high**
truth	**low**	389	111
	high	38	94

Table D.2: Confusion matrix of the thermal valence predictor in NVIE posed database.

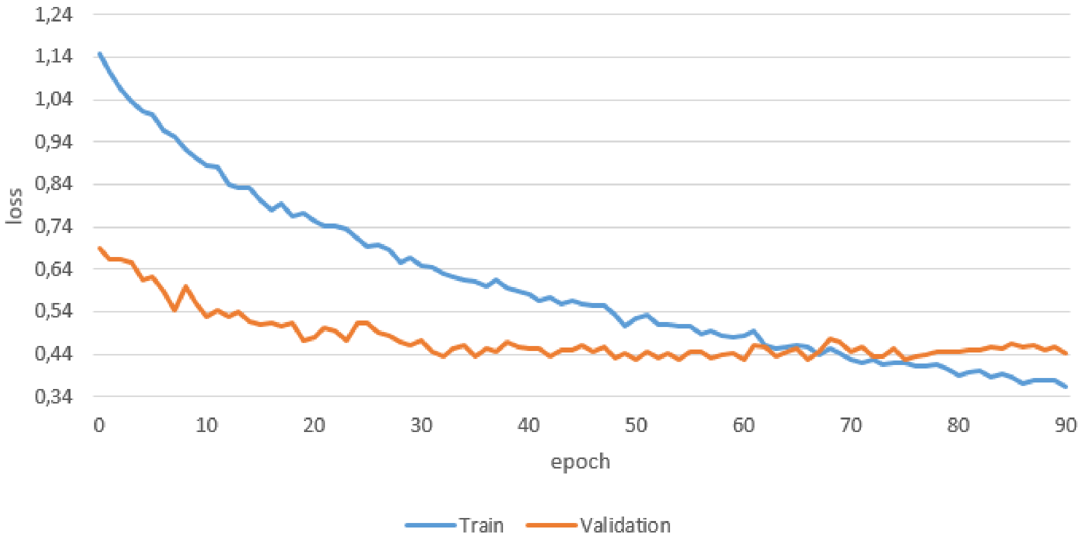


Figure D.1: History of the training session of the model that obtained lowest validation loss in cross-validation for predicting valence in infrared images from NVIE spontaneous database. The train took 75 epochs and reached a validation loss of 0.4259.

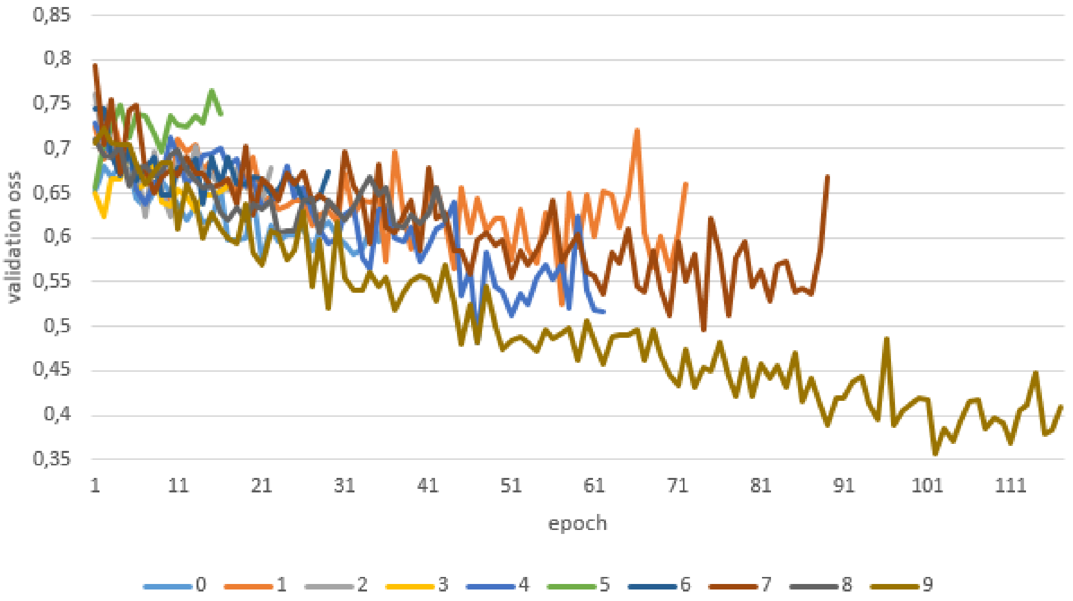


Figure D.2: History of the training sessions of 10 models (for cross-validation) for valence prediction in thermal images, pre-trained in NVIE and fine-tuned in our dataset. There is a high variance in the results due to the small number of images containing labeled facial expressions in our dataset. The best model trained for 101 epochs and reached a validation loss of 0.3563.

		prediction						
		neutral	**happiness**	**disgust**	**fear**	**surprise**	**anger**	**sadness**
truth	**neutral**	3	2	7	0	10	20	4
	happiness	0	107	9	3	6	7	0
	disgust	0	20	15	3	11	43	28
	fear	1	12	14	5	38	41	15
	surprise	0	24	8	4	29	39	14
	anger	0	5	21	2	16	56	27
	sadness	3	5	19	1	12	63	24

Table D.3: Confusion matrix of the RGB+T ensemble predicting facial expressions in NVIE posed database.

		prediction						
		neutral	**happiness**	**disgust**	**fear**	**surprise**	**anger**	**sadness**
truth	**neutral**	48	3	2	11	14	1	17
	happiness	3	38	0	4	7	0	1
	disgust	4	2	2	1	0	0	0
	fear	2	0	1	1	1	0	2
	surprise	1	0	0	2	5	0	1
	anger	0	0	2	1	3	1	1
	sadness	1	0	0	2	2	0	4

Table D.4: Confusion matrix of the RGB+T ensemble predicting facial expressions in our custom-made database inside a vehicle.

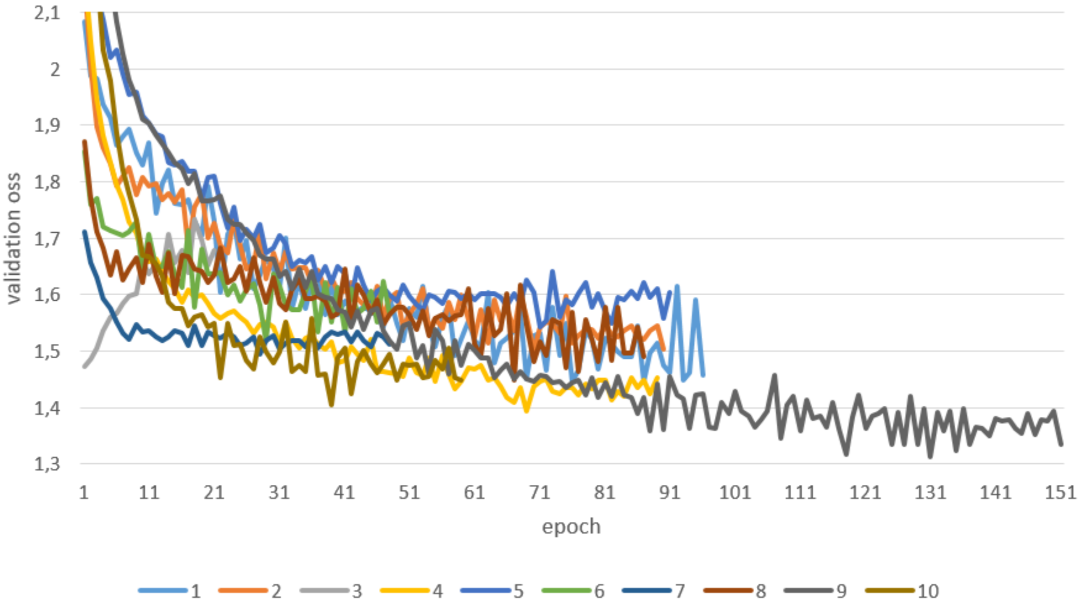


Figure D.3: History of the training sessions of 10 models (for cross-validation) for facial expression recognition using a RGB and thermal ensemble, pre-trained in NVIE and fine-tuned in our dataset. There is a high variance in the results due to the small number of images containing labeled facial expressions in our dataset. The best model trained for 130 epochs and reached a validation loss of 1.314.