UNIVERSITÀ DI PARMA

UNIVERSITÀ DEGLI STUDI PARMA

Dottorato di Ricerca in Tecnologie dell'Informazione

XXXV Ciclo

Content Extraction from Video Signals with Applications to the Automotive and Healthcare Sectors

Coordinatore:

Chiar.mo Prof. Marco Locatelli

Tutor:

Chiar.mo Prof. Riccardo Raheli

Dottoranda: *Veronica Mattioli*

Anni 2019-2022

To my parents
To Davide

Abstract

Content extraction from video signals represents a topic of great research interest thanks to the promising characteristics of video processing techniques, which are mainly related to the use of digital cameras as data acquisition sensors.

The use of digital cameras to generate and record video signals, to be furtherly processed, provides several advantages regarding the cost and deployment of video-based systems developed to extract information about real world scenarios. These systems are non-invasive, as camera sensors do not require a direct contact with the subject to be framed. Also, their cost is moderate, thanks to the ubiquitous diffusion of digital cameras that makes them accessible and user-friendly devices. Thanks to these aspects, video-processing techniques are versatile and may be adopted in a wide range of application scenarios.

Other advantages are related to the characteristics of video signals. Being defined as temporal sequences of still digital images, video signals contain time-related information of framed objects and scenes, that may be associated with relevant evolutionary changes.

In this thesis, video-based algorithms to extract properties of dynamic systems are proposed. In particular, specific applications to the automotive and healthcare sectors are presented and innovative video-based solutions are employed in the tasks of motion analysis and human monitoring.

Robust motion estimation algorithms are needed to extract various characteristics of dynamic objects related to their motion, e.g., speed and periodicity. Speed estimation, for instance, plays an important role in the context of automotive safety. In this thesis, the topic of estimating the speed of framed objects in video sequences is addressed and a method to deal with geometrical transformations superimposed to the shift of the object under analysis in the camera plane is proposed.

Algorithms to extract the periodicity, typical of some movements, are also presented. The respiration act is an example of a periodic movement, that is worth to be investigated in medical applications as it provides important information related to the health status of a subject. As the Maximum Likelihood (ML) principle represents a reliable tool to derive estimators of unknown parameters of interest, it is exploited in this work to implement speed and Respiratory Rate (RR) estimation algorithms based on video processing techniques, adopted to enhance the considered motion signals.

The topic of human monitoring in automotive scenarios represents, instead, a cross-sectoral application concerning both healthcare and automotive safety. In this work, a system to assess the stress status of a driver is proposed by combining information extracted from video signals and from physiological sensors. In particular, thermography is exploited to retrieve skin temperature variations, possibly caused by stressful events, from thermal images acquired inside a vehicle during driving performance.

Contents

List of Acronyms

2D	two-dimensional	2, 5, 50, 82
3D	three-dimensional	2, 5
ABS	Anti-lock Braking System	72
ACC	Adaptive Cruise Control	72
ADAS	Advanced Driver-Assistance System	72
ANS	Autonomic Nervous System	71
CMYK	Cyan, Magenta, Yellow, Black	3
DFD	Displaced Frame Difference	9
DFT	Discrete Fourier Transform	15
ECG	ElectroCardioGram	43, 73, 88
ECU	Electronic Control Unit	75
EVM	Eulerian Video Magnification	45
Faster-RCNN	Faster Region-based Convolutional Neural Network	9
FT	Fourier Transform	9, 52
GA	Genie-Aided	68

GMM Gaussian Mixture Model . . . 12

HR Heart Rate . . . 4, 71, 88

HRV Heart Rate Variability . . . 74

HSV Hue, Saturation, Value . . . 3, 44

i.i.d. independent and identically distributed . . . 13, 56

IIR Infinite Impulse Response . . . 48

IoT Internet of Things . . . 84

IR Infra Red . . . 44, 88

JDVS Joint Driver Vehicle Status . . . 78

LED light-emitting diode . . . 43

MAD Mean Absolute Difference . . . 8

MAP Maximum A Posteriori . . . 9

ML Maximum Likelihood . . . ii, 4, 6, 42, 87

MPEG Moving Picture Experts Group . . . 8

MQTT Message Queue Telemetry Transport . . . 78

MRI Magnetic Resonance Imaging . . . 2

MSE Mean Square Error . . . 8

PDF Power Spectral Density . . . 9

PET Positron Emission Tomography . . . 2

PPG PhotoPlethysmoGraphy . . . 43

RGB Red, Green and Blue . . . 2, 10, 44, 74

RMS Root Mean Squared 34, 60

RMSE Root Mean Squared Error 34, 42

ROI Region of Interest 42, 83

RR Respiratory Rate . . . . . ii, 4, 41, 73, 88

SEM Sensor Electronics Module 75

SIFT Scale Invariant Feature Transform 9

SNR Signal to Noise Ratio 35

SURF Speeded Up Robust Features 9

TCP Transmission Control Protocol 79

UAV Unmanned Aerial Vehicle 8

Chapter 1

Introduction

This thesis is aimed to develop and analyse video processing systems to extract relevant information content in various application scenarios. The topic of content extraction from video signals is part of the broader field of image processing and computer vision, which is gaining increasing attention due to its ubiquity and wide range of applications. Computer vision primarily deals with digital images that can also be in the form of sequences, i.e., videos. Video processing is, indeed, a particular case of image processing. However, richer information is stored in video signals rather than static images. In particular, video signals are able to capture motion, a fundamental property of dynamic systems that is associated with temporal changes and evolutions.

The main goal of computer vision is to emulate the human visual system function of understanding the world that surrounds us by recovering informative characteristics, i.e., features, of objects and scenes through signal-processing techniques. Image understanding may provide an important support in automating specific tasks in various contexts. This thesis focuses on two main fields of application: automotive safety and healthcare. In particular, some of the considered tasks include motion detection, object tracking and the monitoring of physiological parameters of subjects both in automotive and medical scenarios. Other important fields of application are video surveillance, machine inspection in industry and support to military operations.

The rapid growth of the amount of multimedia data, i.e., images and videos, has contributed to significantly boost the research in the computer vision field. Taking pictures and recording videos is nowadays easy and cheap. Digital cameras are, indeed, accessible, affordable and often integrated into other commonly used devices, such as smartphones and computers. Furthermore, various types of more sophisticated sensors exist and allow to produce different types of images, that may be exploited in different contexts according to their characteristics. For instance, Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) are imaging techniques commonly used for medical examinations as they can provide information about the status of internal structures of the human body and some of its physiological functions. Thermal imaging is, instead, an example of infrared imaging that allows to measure the temperature of objects. This property may be exploited in medical applications and for various other purposes that range from surveillance to the detection of heat leaks in buildings. In this work, the considered sensors are Red, Green and Blue (RGB) and thermal cameras.

For any application, however, a solid knowledge of the employed device is fundamental. The task of extracting information, namely interpreting images, is, indeed, quite challenging as during the process of image formation pieces of information are lost. Having a good understanding of how images are formed and represented is, hence, necessary to properly model objects and scenes and to choose the most appropriate processing method.

1.1 Preliminary Definitions

In this work, digital images are conventionally defined as two-dimensional (2D) representations of the three-dimensional (3D) world. Grayscale images and videos are considered for the sake of simplicity and because color images are specified by 3 grayscale images associated with the RGB color components. In particular, a grayscale image of size $M_1 \times M_2$ can be described as a function $x[\mathbf{m}]$ of a 2D discrete variable, whose elements, i.e., pixels, are intensity values identified by the pair of spatial coordinates $\mathbf{m} = (m_1, m_2)$, where $0 \ \leq m_1 \ \leq M_1 - 1$ and $0 \leq m_2 \leq M_2 - 1$. The intensities of a digital image are represented as discrete values as a result of the quantization process [1]. As common practice, a digital image is usually described in the form of a matrix with M_1 rows and M_2 columns and $x[\mathbf{m}]$ can be referred to as a 2D discrete function for brevity.

A grayscale video signal is composed by a sequence of grayscale digital images, which are referred to as frames, whose spatial-intensity pattern may vary over time. It is hence a time varying multidimensional signal and can be represented as a discrete function $x[\mathbf{m}, n]$, where n is the discrete time index corresponding to the frame number. Frames are sampled at instants nT_s, where T_s is the sampling time, with $f_s = 1/T_s$ being the sampling frequency, or frame rate, of the camera. The frame size is obtained as $M_1 \times M_2$, being M_1 and M_2 its height and width, respectively.

In this work, as customary, the dynamic range of the pixel intensity is limited to the interval $[0, 1]$.

In the case of colored images and videos, the pixel intensity values are defined for each color channel of the considered color space. Popular color spaces include RGB, Cyan, Magenta, Yellow, Black (CMYK) and Hue, Saturation, Value (HSV) representations [2].

1.2 Thesis Organization

In this thesis novel video processing-based algorithms to extract content in three different application scenarios are presented. The dissertation is, hence, organized into three main chapters as follows.

- *Motion analysis.* The topic of motion estimation is tackled in the first chapter. After providing an overview on existing techniques, a novel method to estimate the speed of objects framed in video sequences is presented. The application of a fundamental theoretical estimation principle, i.e., the Maximum Likelihood (ML) criterion, that allows to estimate different unknowns according to the goal to be pursued, is investigated. The effectiveness of the proposed method will be demonstrated by a set of results obtained through numerical simulations.

- *Breathing monitoring.* In the second chapter, novel video processing techniques for the visualization and contactless analysis of respiration are discussed. In particular, two motion magnification methods to enhance subtle respiration-related movements and a method to finally estimate the Respiratory Rate (RR) are analysed. The performance and the accuracy of the proposed systems are evaluated on numerical results obtained by testing video sequences of various subjects in a number of scenarios. The performance of the considered methods is also compared against reference data.

- *Driver monitoring for automotive applications.* The last chapter is devoted to the monitoring of the status of drivers in automotive scenarios. In particular, a novel system able to collect physiological parameters of interest is presented and its feasibility is discussed on the basis of experimental evaluations. The proposed system is composed by heterogeneous sensors that provide for different types of data. A wearable sensor and a thermal camera are used to respectively extract physiological information, such as the Heart Rate (HR), and the temperature of the subject. In this case, the temperature information extracted by the considered video sequences is correlated to the other collected data to provide a useful indication on the status of the driver.

Chapter 2

Motion Analysis

Video processing solutions for motion analysis are key tasks in many computer vision applications, ranging from human activity recognition to object detection. In particular, speed estimation algorithms may be relevant in contexts such as street traffic monitoring and surveillance for safety purposes.

In most realistic scenarios, the motion of foreground objects is superposed to other dynamic modifications that mainly arise from periodic behaviours (e.g., typical of some human movements such as walking and running) or directly result from the process of image acquisition [1, 3]. A video frame can be indeed defined as a digital image generated by the projection of a three-dimensional (3D) real-world scene onto a two-dimensional (2D) camera plane. For this reason, perspectival effects are likely to affect the image of the framed objects of interest. Hence the analysis of their motion, i.e., speed estimation and periodic feature extraction, may be challenging in some scenarios and requires advanced speed estimation techniques based on robust algorithms for object detection that are able to deal with potential geometrical modifications.

In this chapter, a novel video processing method to estimate the speed of foreground objects is presented. The proposed algorithm is composed of a sequence of pre-processing operations, that aim to reduce or neglect perspectival effects affecting the objects of interest, followed by the speed estimation phase based on the Maximum Likelihood (ML) principle. The ML estimation method represents, indeed, a consolidated statistical tool that may be exploited to obtain an estimator of the speed of the objects of interest achieving reliable results. As a matter of fact, the literature on the extraction of information content from video signals is mainly based on heuristic ad-hoc solutions and little or no attempts to employ sound approaches from estimation theory have been pursued. As exceptions to this general trend, [4, 5] need to be mentioned as previous contributions in which the ML criterion was successfully employed in the context of video processing for the extraction of periodic features.

The chapter is organized as follows. In Section 2.1, an overview on existing motion estimation techniques is introduced. In Section 2.2, the proposed method is presented. The preliminary video processing operations are described along with the mathematical formulation of the dynamic motion model, whereas the ML estimation approach is detailed in Section 2.3. Results are presented in Section 2.4, where the performance of the proposed algorithm is evaluated on sets of synthetic and real videos and compared with a reference method, i.e., the block-matching approach [6, Ch. 4]. Section 2.5 is dedicated to final remarks.

2.1 Motion Estimation Techniques

Motion estimation can be defined as the process that allows to measure how objects move in a video sequence [7]. When a video is recorded, a 3D real-world scene is projected onto a 2D surface, i.e., the camera plane, resulting in a sequence of 2D digital images, i.e., frames. Hence, a motion in a video stream can be considered as the consequence of the video acquisition process and contains important spatio-temporal information about the captured 3D scene. In practice, the motion of an object correspond to variations in the pixel intensity values in the 2D image plane.

Recovering motion-related information, such as the speed of objects, plays a fundamental role in contexts such as traffic control and road monitoring for automotive safety. Thanks to the increasing deployment of cameras for surveillance purposes, a relevant amount of street-related information is available and may be exploited to build non-intrusive video-based solutions for object speed estimation.

Various motion estimation algorithms are widely documented in the literature, as reported, e.g., in [6, Ch. 4] where a broad classification is proposed. In particular, differential and matching methods represent two wide employed techniques, whereas other categories include optimization and transform-domain methods. Feature- and learning-based approaches are also being employed to track objects and retrieve their motion. A brief overview of these approaches is provided hereafter.

Differential methods

Differential methods are strictly related to the concept of optical flow, defined as the apparent motion perceived from the variations in the pixel intensity patterns in the 2D image. These changes may be due either to a true motion in the 3D space or to illumination changes [8]. For this reason, the optical flow is very sensitive to brightness variations that cause intensity oscillations despite a possible absence of motion.

Differential methods aim to estimate the optical flow by means of gradient techniques that exploit the properties of spatial and temporal partial derivatives. The most popular solutions in this category are the Lucas-Kanade [9] and the Horn-Schunck methods [10], that rely on strong initial assumptions. In the Lucas-Kanade method, the brightness constancy model is adopted and the optical flow is considered constant in a sufficiently small region surrounding the pixel under analysis. On the other hand, the Horn-Schunck approach is based on the hypothesis that the optical flow variations are smooth in the considered image. Due to these specific assumptions, the two methods may fail in properly describing motion in some realistic scenarios. Despite this main limitation, both methods continue to be widely employed, as in the following examples.

The works proposed in [11] and [12] are examples of speed estimation algorithms where the Lucas-Kanade algorithm is applied. In particular, in [11], the average traffic speed is extracted from Unmanned Aerial Vehicle (UAV) videos and the optical flow is computed to track points of interest between pairs of consecutive frames. The result of this procedure is a number of motion vectors that are subsequently grouped in clusters of vehicle that move with a similar speed. The motion vectors related to the centers of the clusters define the average speed of the corresponding clusters. In [12], a similar approach is proposed to estimate the instantaneous speed of a vehicle. Corners are detected as points of interest and tracked throughout the video sequence by means of an optimized version of the Lucas-Kanade algorithm. The speed of the corners belonging to a considered vehicle are averaged to obtain its speed.

Matching methods

Block matching approaches are the most popular matching methods. They provide for partitioning a considered video frame into several blocks of pixels, that may or may not overlap and may have fixed or variable size. Each block is associated with a motion vector that can be obtained according to different criteria. In particular, the location of a block in a present frame is searched within a frame which is considered as reference and different searching strategies may be adopted, as detailed in [6, Ch. 4]. The criteria employed to match two blocks in different frames are usually based on error functions, e.g., Mean Square Error (MSE) and Mean Absolute Difference (MAD), that need to be minimized.

Due to their easy implementation, block matching approaches are widely employed, especially in video compression applications. In this category, algorithms to estimate the average speed of a group of vehicles from Moving Picture Experts Group (MPEG) video streams are presented in [13] and [14], where motion-related information is directly extracted from the considered stream. In fact, the MPEG standard allows to encode and store motion vectors obtained by the block-matching algorithm, that can be easily extracted through an MPEG parser.

Block matching criteria, despite being straightforward and basic, are prone to errors, especially in the presence of noise. Their performance is also highly sensitive to the parameter setting, such as the block size, and may be strongly affected by the characteristics of the considered video frames. Motion vectors extracted from low texture scenes are often unreliable as the blocks may be easily mismatched in case of high similarity between non corresponding blocks.

Other methods

Among other motion estimation techniques, optimization and frequency-domain methods need to be mentioned. Pel-recursive and Bayesian approaches [6, Ch. 4] are two popular solutions in the first category. In particular, in pel-recursive algorithms, a specific error function, i.e., the Displaced Frame Difference (DFD), is minimized, whereas Bayesian methods exploit the Maximum A Posteriori (MAP) estimation criterion to maximize the Power Spectral Density (PDF) of the motion field.

Phase correlation [6, Ch. 4] is, instead, a method to estimate a displacement in the Fourier domain, which provides for computing the phase difference between the Fourier Transform (FT) of two considered frames.

More modern motion estimation approaches provide for detecting and tracking the objects of interest through feature- or learning-based methods. In [15], Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) algorithms are exploited to extract feature points of moving vehicles. The detected points of interest are matched throughout the video sequence according to a specific matching function and the distance, in pixels, between two matched feature points is finally computed to obtain the speed of the vehicles. Examples of learning-based methods can be found in [16] and [17], where the Faster Region-based Convolutional Neural Network (Faster-RCNN) framework is exploited to extract the locations of the framed vehicles and compute their speed.

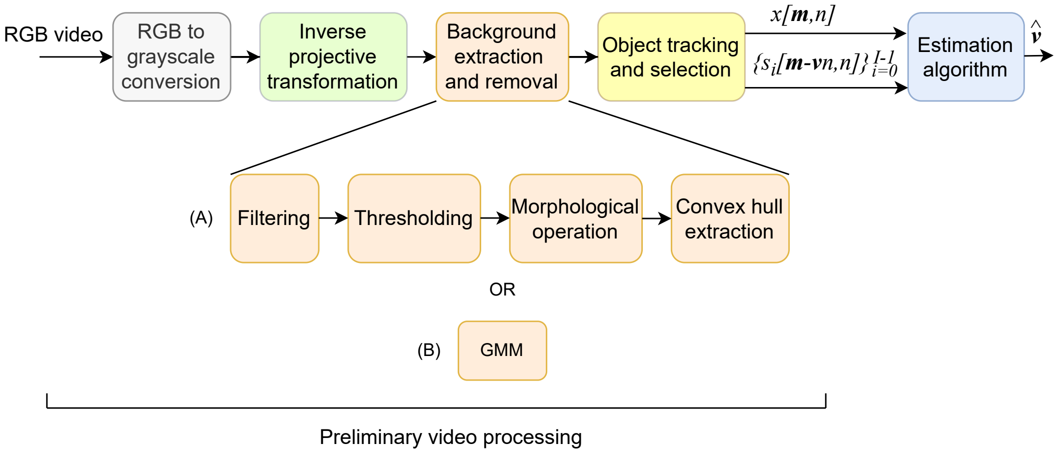

RGB video
RGB to grayscale conversion
Inverse projective transformation
Background extraction and removal
Object tracking and selection
x[m,n]
{s_i[m-vn,n]} I-1 i=0
Estimation algorithm
v̂
(A)
Filtering
Thresholding
Morphological operation
Convex hull extraction
OR
(B)
GMM
Preliminary video processing

Figure 2.1: Overview of the proposed speed estimation method.

2.2 Observation Model

2.2.1 Preliminary processing for object detection

The extraction of foreground objects in a video sequence is at the basis of the proposed solution. An overview of the proposed method is shown in Figure 2.1. In particular, an initial grayscale conversion from Red, Green and Blue (RGB) input sequences is performed to obtain a grayscale video signal with frame period T_s. This conversion is motivated by the significant simplification it entails in the following processing operations and is highlighted in the first block of the diagram shown in Figure 2.1. Extension to processing methods that operate on color videos are straightforward and will not be pursued. As already mentioned, in many realistic scenarios, foreground moving objects are likely to be subject to perspective transformations arising from the projective mapping of 3D real world points to corresponding points on the 2D image plane [1]. This mapping has also an immediate effect on the object speed: a constant speed in the real world may correspond to a non-constant speed in the image plane. However, it is possible to remove perspectival distortions by exploiting some image processing techniques that also allow to approximately recover the original shape and size of the objects of interest [1, 18]. To this purpose, as highlighted in the second block of the diagram in Figure 2.1, inverse projective transformations may be applied to each frame of the considered video sequence in order to compensate for the non-constant speed.

Once the original shape and speed of the framed objects is restored and can be considered constant in the 2D image plane, they can be detected by removing the background, which is assumed to be static. The background removal operation is highlighted in the third block of the diagram in Figure 2.1 and may be implemented by means of different techniques. A first solution consists in a basic approach composed of four main steps, i.e., filtering, thresholding, morphological operation and convex hull extraction, as shown in the respective blocks (A) in Figure 2.1. In this case, a basic image filtering operation based on the absolute difference of each frame and the background frame, i.e., reference frame, is initially performed and the result is thresholded to obtain a binary image where pixels belonging to the foreground have intensity equal to 1 (white) and those belonging to the background have intensity equal to 0 (black). When the considered video sequence is affected by noise, such as Gaussian noise, a spatial averaging filter can be applied to each frame to smooth the noise effect [1]. Also, a temporal average operation may be performed on the whole sequence to estimate the background in the presence of noise or if the video sequence does not contain any background frame, i.e., a frame where the background is totally visible and not partially occluded by any foreground object. A cascade of morphological operations [1], i.e., erosion followed by dilation with different structuring elements, is then applied to reduce the misclassification of isolated objects and the convex hull of the foreground objects is extracted to fill any remaining hole. Erosion is indeed necessary to remove small isolated objects, such as noisy pixels and non-removed parts of background. However, as erosion has also the effect of reducing the size of objects, included the objects of interest, dilation is needed to broaden and thicken eroded regions [1].

As an alternative solution, the Gaussian Mixture Model (GMM) algorithm [19], which aims to adaptively model each pixel belonging to the background as a mixture of Gaussian distributions, may be exploited to remove the background. However, this method may be less robust against noise and may have a high computational cost, as discussed in [20]. For this reason, more basic and robust approaches, such as the one shown in the blocks (A) in Figure 2.1, may be preferable. The GMM is denoted as the block (B) in Figure 2.1.

As a last step of this preliminary processing, object tracking and selection may be performed, as shown in Figure 2.1, based on standard feature extraction techniques [3] to select the objects of interest, whereas the rest of the scene can be considered as background. Once the regions of interest have been selected, a feature detection algorithm can, indeed, be exploited to extract the locations of corners or other points of interest of the foreground objects, e.g., vehicles, at a specific frame. In particular, the minimum eigenvalue algorithm [21] can be employed to detect points exhibiting good texture information. Regions with the two minimum eigenvalues above a fixed threshold, determined according to noise requirements, are normally associated to patterns which are reliable to track. The locations of the detected features may be then searched in subsequent frames through feature tracking algorithms such as the Kanade-Lucas-Tomasi algorithm [22].

Once a particular object or a group of objects of interest is correctly detected over the whole sequence, its speed can be estimated by the proposed algorithm. The detection and tracking procedure may be repeated for each group of objects framed in the video sequence. However, when different objects of interest, e.g., vehicles, are not simultaneously framed and/or move in distinct regions of the considered sequence of frames, e.g., opposite road lanes, feature detection may not be necessary. In this case, basic operations of trimming and cropping may be performed to obtain new video sequences framing a single foreground object at a time.

The definitions of the notation $x[\mathbf{m}, n]$, $\{s_i[\mathbf{m} - \mathbf{v}n, n]\}_{i=0}^{I-1}$ and $\hat{\mathbf{v}}$ in Figure 2.1 are provided in Sections 2.2.2 and 2.3.

2.2.2 Dynamic motion model

Consider a generic grayscale video sequence defined by the 2D discrete function $x[\mathbf{m}, n]$, as described in Chapter 1.

Considering I framed objects subject to perspectival distortion, or any other dynamic change, the image of the i-th object can be denoted as $s_i[\mathbf{m}, n]$, with $0 \leq i \leq I-1$. A time dependent displacement vector $\boldsymbol{\delta}_i[n] = (\delta_{i,1}[n], \delta_{i,2}[n])^\mathsf{T}$, where $\delta_{i,1}[n]$ and $\delta_{i,2}[n]$ represent the horizontal and vertical components, respectively, can also be considered. Following the approach in [20], and defining $s_i[\mathbf{m} - \boldsymbol{\delta}_i[n], n]$ as a shift that affects the i-th object in the 2D image plane at the n-th frame, the pixel intensities of the grayscale video signal can be modeled as

$$x[\mathbf{m}, n] = b[\mathbf{m}] + \sum_{i=0}^{I-1} s_i[\mathbf{m} - \boldsymbol{\delta}_i[n], n] + v_b[\mathbf{m}, n] + w[\mathbf{m}, n] \tag{2.1}$$

where $b[\mathbf{m}]$ is the static background, whose partial occlusion/un-occlusion due to the motion of the objects is taken into account by the term $v_b[\mathbf{m}, n]$, and $w[\mathbf{m}, n]$ represents samples of independent and identically distributed (i.i.d.) zero-mean Gaussian noise. When inverse projective transformations are applied and the shape and size of the framed objects can be considered constant, their images and shifts in the 2D image plane can be simplified as $s_i[\mathbf{m}]$ and $s_i[\mathbf{m} - \boldsymbol{\delta}_i[n]]$, respectively. The dynamic motion model in (2.1) represents a suitable model to describe a video frame, to which the ML estimation principle can be applied, with proper mathematical manipulation. A schematic illustration of the model in (2.1) is shown in Figure 2.2, where the motion of an object from the n_1-th to the n_2-th frame is represented.

The I objects are now assumed to be moving with comparable and almost constant speed. This assumption is a basic requirement for any speed estimation problem, as the speed must be almost constant over a window of

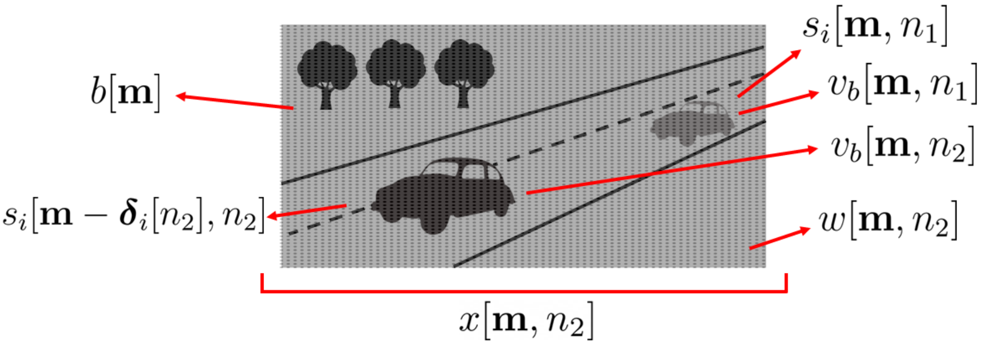


Figure 2.2: Schematic illustration of the motion model in (2.1).

consecutive frames of sufficient duration in order to enable its correct estimation. Hence, the common displacement term can be expressed as $\boldsymbol{\delta}[n] = \mathbf{v}n$, where $\mathbf{v} = (v_1, v_2)^\mathsf{T}$ is the vector of the common uniform speed, measured in pixel/frame, to be estimated. Accordingly, the model in (2.1) can be written as

$$x[\mathbf{m}, n] = \; b[\mathbf{m}] + \sum_{i=0}^{I-1} s_i[\mathbf{m} - \mathbf{v}n, n] + v_b[\mathbf{m}, n] + w[\mathbf{m}, n]. \tag{2.2}$$

By the preliminary background removal operation discussed in Section 2.2.1 and corresponding to the third block of the diagram shown in Figure 2.1, it is possible to further simplify the observation model in (2.2) as

$$x[\mathbf{m}, n] = \sum_{i=0}^{I-1} s_i[\mathbf{m} - \mathbf{v}n, n] + w[\mathbf{m}, n] \tag{2.3}$$

where the background-related terms $b[\mathbf{m}]$ and $v_b[\mathbf{m}, n]$ have been neglected and the observation sequence $x[\mathbf{m}, n]$ is obtained after the background removal operation.

Assume first that $\mathbf{v}$ has integer components in pixel/frame. Further processing can be conveniently performed in the FT domain, which provides the advantage of expressing a displacement as a linear phase term thanks to the shift theorem, as also discussed in [20]. Hence, applying the definition of the Discrete Fourier Transform (DFT) of a generic 2D discrete function [23], the frequency domain equivalent of the model in (2.3) can be expressed as

$$X[\mathbf{k}, n] = \sum_{i=0}^{I-1} S_i[\mathbf{k}, n] e^{-j2\pi \mathbf{u_k}^{\mathsf{T}} \mathbf{v} n} + W[\mathbf{k}, n] \tag{2.4}$$

where $\mathbf{k} = (k_1, k_2)^{\mathsf{T}}$ is the vector of the two discrete indices of the 2D DFT, with $0 \leq k_l \leq M_l - 1$, $l = 1, 2$, $\mathbf{u_k} = (k_1/M_1, k_2/M_2)^{\mathsf{T}}$ is the vector of the normalized spatial frequencies and uppercase letters denote the DFTs of the corresponding signals in (2.3).

Consider now the case of a fractional value of the speed vector $\mathbf{v}$, denote the number of sub-pixel quantization levels by the integer F and assume $\mathbf{v}$ is quantized accordingly. The displacement vector can be written as

$$\mathbf{v}n = \mathbf{d}[\mathbf{v}, n] + \frac{\mathbf{f}[\mathbf{v}, n]}{F} = \left(d_1[v_1, n] + \frac{f_1[v_1, n]}{F}, d_2[v_2, n] + \frac{f_2[v_2, n]}{F} \right)^{\mathsf{T}} \tag{2.5}$$

where

$$\mathbf{d}[\mathbf{v}, n] = (d_1[v_1, n], d_2[v_2, n])^{\mathsf{T}} = \left\lfloor \mathbf{v}n \right\rfloor \tag{2.6}$$

$$\frac{\mathbf{f}[\mathbf{v}, n]}{F} = \frac{(f_1[v_1, n], f_2[v_2, n])^{\mathsf{T}}}{F} = \{\mathbf{v}n\} \tag{2.7}$$

represent the integer and fractional parts of the vector, respectively, with $f_i[v_i, n] \in \{0, 1, 2, \dots, F-1\}$, $i = 1, 2$, $\lfloor \cdot \rfloor$ denotes the floor function and $\{x\} = x - \lfloor x \rfloor$.

The model in (2.3) can be now extended to the general case where the foreground objects may shift with a fractional speed in both directions. The fractional sub-pixel translation of an image can be defined as the bilinear interpolation of the neighbouring 4 pixels at the vertices of the sub-pixel position of interest. Hence, considering an image $s[\mathbf{m}]$ to be shifted with fractional displacement $\left(\frac{f_1}{F}, \frac{f_2}{F}\right)$, $f_i \in \{0, 1, 2, \dots, F-1\}$, $i = 1, 2$, its fractional sub-pixel translation $y[\mathbf{m}]$ can be defined as

$$\begin{aligned} y[\mathbf{m}] =& \left(1 - \frac{f_1}{F}\right)\left(1 - \frac{f_2}{F}\right) s[\mathbf{m}] \\ &+ \frac{f_1}{F}\left(1 - \frac{f_2}{F}\right) s\big[\mathbf{m} - \mathbf{h}_1\big] \\ &+ \left(1 - \frac{f_1}{F}\right)\frac{f_2}{F} s\big[\mathbf{m} - \mathbf{h}_2\big] \\ &+ \frac{f_1}{F}\frac{f_2}{F} s\big[\mathbf{m} - \mathbf{h}_1 - \mathbf{h}_2\big] \end{aligned} \tag{2.8}$$

where $\mathbf{h}_1 = (1,0)^\mathsf{T}$ and $\mathbf{h}_2 = (0,1)^\mathsf{T}$ are the unitary vectors related to the two components. The four terms in expression (2.8) represent the contribution of each vertex to the fractional sub-pixel translation. The model in (2.3) can be thus expanded as

$$\begin{aligned} x[\mathbf{m},n] =& \left(1 - \frac{f_1[v_1,n]}{F}\right)\left(1 - \frac{f_2[v_2,n]}{F}\right)\sum_{i=0}^{I-1} s_i\big[\mathbf{m} - \mathbf{d}[\mathbf{v},n],n\big] \\ &+ \frac{f_1[v_1,n]}{F}\left(1 - \frac{f_2[v_2,n]}{F}\right)\sum_{i=0}^{I-1} s_i\big[\mathbf{m} - \mathbf{d}[\mathbf{v},n] - \mathbf{h}_1,n\big] \\ &+ \left(1 - \frac{f_1[v_1,n]}{F}\right)\frac{f_2[v_2,n]}{F}\sum_{i=0}^{I-1} s_i\big[\mathbf{m} - \mathbf{d}[\mathbf{v},n] - \mathbf{h}_2,n\big] \\ &+ \frac{f_1[v_1,n]}{F}\frac{f_2[v_2,n]}{F}\sum_{i=0}^{I-1} s_i\big[\mathbf{m} - \mathbf{d}[\mathbf{v},n] - \mathbf{h}_1 - \mathbf{h}_2,n\big] + w[\mathbf{m},n]. \end{aligned} \tag{2.9}$$

Taking the 2D DFT of (2.9), an equivalent observation model in the frequency domain can be obtained. Using again the shift theorem, this model can be formulated as

$$X[\mathbf{k},n] = \sum_{i=0}^{I-1} S_i[\mathbf{k},n] e^{-j2\pi \mathbf{u_k}^\mathsf{T}\mathbf{d}[\mathbf{v},n]} a[\mathbf{v},n] + W[\mathbf{k},n] \tag{2.10}$$

where

$$
\begin{aligned}
a[\mathbf{v},n] =& \left(1-\frac{f_1[v_1,n]}{F}\right)\left(1-\frac{f_2[v_2,n]}{F}\right) \\
&+ \frac{f_1[v_1,n]}{F}\left(1-\frac{f_2[v_2,n]}{F}\right)e^{-j2\pi\frac{k_1}{M_1}} \\
&+ \left(1-\frac{f_1[v_1,n]}{F}\right)\frac{f_2[v_2,n]}{F}e^{-j2\pi\frac{k_2}{M_2}} \\
&+ \frac{f_1[v_1,n]}{F}\frac{f_2[v_2,n]}{F}e^{-j2\pi\left(\frac{k_1}{M_1}+\frac{k_2}{M_2}\right)}.
\end{aligned}
\tag{2.11}
$$

2.3 Maximum Likelihood Speed Estimation

Observing now that the model in (2.10) describes Gaussian observations that are independent in the spatial and discrete frequency domains, ML estimation can be used to derive an expression of the estimator $\hat{\mathbf{v}}$ of the unknown speed vector [23]. The dependence of (2.10) on the speed vector is through the terms $\mathbf{d}[\mathbf{v},n]$ and $a[\mathbf{v},n]$.

Considering an observation window of N frames, the relevant likelihood function of the model in (2.10) is

$$
\begin{aligned}
&p\big(X[\mathbf{k},0]\cdots X[\mathbf{k},N-1];\mathbf{v}\big) \\
&\quad= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{M_1M_2N}{2}} \cdot \exp\Bigg\{-\frac{1}{2\sigma^2}\sum_{k_1=0}^{M_1-1}\sum_{k_2=0}^{M_2-1}\sum_{n=0}^{N-1}\Bigg|X[\mathbf{k},n] \\
&\qquad -\sum_{i=0}^{I-1}S_i[\mathbf{k},n]e^{-j2\pi\mathbf{u_k}^{\mathsf{T}}\mathbf{d}[\mathbf{v},n]}a[\mathbf{v},n]\Bigg|^2\Bigg\}
\end{aligned}
\tag{2.12}
$$

where $p\big(X[\mathbf{k},0]\cdots X[\mathbf{k},N-1];\mathbf{v}\big)$ denotes the joint PDF of the noisy observation sequence for a trial speed vector $\mathbf{v}$ and σ is the standard deviation of the additive Gaussian noise elements.

The log-likelihood function can also be derived from (2.12) as

$$\ln\big(p(X[\mathbf{k},0]\cdots X[\mathbf{k},N-1];\mathbf{v})\big) = \underbrace{-\frac{M_1M_2N}{2}\ln(2\pi\sigma^2)}_{\text{(a)}} \\ -\frac{1}{2\sigma^2}\underbrace{\sum_{k_1=0}^{M_1-1}\sum_{k_2=0}^{M_2-1}\sum_{n=0}^{N-1}\left|X[\mathbf{k},n]-\sum_{i=0}^{I-1}S_i[\mathbf{k},n]e^{-j2\pi\mathbf{u_k}^{\mathsf{T}}\mathbf{d}[\mathbf{v},n]}a[\mathbf{v},n]\right|^2}_{\text{(b)}} \tag{2.13}$$

where some terms are highlighted. In particular, given that (a) is a constant term and the multiplicative coefficient $-\frac{1}{2\sigma^2}$ is a constant factor, in the sense that they do not depend on the trial speed value $\mathbf{v}$, they are irrelevant for the estimation problem and can be discarded. The term (b) can be equivalently minimized with respect to the value of $\mathbf{v}$. Hence, an equivalent likelihood function to be minimized is

$$\sum_{k_1=0}^{M_1-1}\sum_{k_2=0}^{M_2-1}\sum_{n=0}^{N-1}\left|X[\mathbf{k},n]-\sum_{i=0}^{I-1}S_i[\mathbf{k},n]e^{-j2\pi\mathbf{u_k}^{\mathsf{T}}\mathbf{d}[\mathbf{v},n]}a[\mathbf{v},n]\right|^2. \tag{2.14}$$

The expression in (2.14) can be made more explicit as

$$\sum_{k_1=0}^{M_1-1}\sum_{k_2=0}^{M_2-1}\sum_{n=0}^{N-1}\Bigg\{|X[\mathbf{k},n]|^2+\left|\sum_{i=0}^{I-1}S_i[\mathbf{k},n]a[\mathbf{v},n]\right|^2 \\ -2\sum_{i=0}^{I-1}\mathrm{Re}\left\{X[\mathbf{k},n]S_i^*[\mathbf{k},n]e^{j2\pi\mathbf{u_k}^{\mathsf{T}}\mathbf{d}[\mathbf{v},n]}a^*[\mathbf{v},n]\right\}\Bigg\} \tag{2.15}$$

where $\mathrm{Re}\{\cdot\}$ and $(\cdot)^*$ are the real part and the complex conjugate operators, respectively. The quadratic terms in (2.15) are irrelevant or practically so. In fact, the term $|X[\mathbf{k},n]|^2$ is independent of $\mathbf{v}$ and is irrelevant. The term $\left|\sum_{i=0}^{I-1}S_i[\mathbf{k},n]a[\mathbf{v},n]\right|^2$ depends on $\mathbf{v}$ through the factor $a[\mathbf{v},n]$, that depends only on the fractional part of the speed vector. In particular, if a grid search with a resolution of $\frac{1}{F}$ pixel/frame is implemented, then F^2 possible values of $a[\mathbf{v},n]$ are obtained according to (2.11), which repeat periodically over the grid of possible values of $\mathbf{v}$. It turns out that these values are subject to very small variations if compared against the mixed term. As consequence, the term $\left|\sum_{i=0}^{I-1} S_i[\mathbf{k},n]a[\mathbf{v},n]\right|^2$ can be considered almost constant, hence practically also irrelevant, and (2.15) can be minimized by maximizing the following approximate likelihood function

$$\sum_{k_1=0}^{M_1-1}\sum_{k_2=0}^{M_2-1}\sum_{n=0}^{N-1}\text{Re}\left\{\sum_{i=0}^{I-1} X[\mathbf{k},n]S_i^*[\mathbf{k},n]e^{j2\pi\mathbf{u_k}^\mathsf{T}\mathbf{d}[\mathbf{v},n]}a^*[\mathbf{v},n]\right\} \tag{2.16}$$

where the linearity of Re$\{\cdot\}$ and sum operators has been exploited. The accuracy of this approximate likelihood function will be discussed and numerically demonstrated in the next section.

Finally, the following expression for the (quasi) ML speed estimator is obtained:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\text{argmax}}\sum_{k_1=0}^{M_1-1}\sum_{k_2=0}^{M_2-1}\sum_{n=0}^{N-1}\text{Re}\left\{\sum_{i=0}^{I-1} X[\mathbf{k},n]\cdot S_i^*[\mathbf{k},n]e^{j2\pi\mathbf{u_k}^\mathsf{T}\mathbf{d}[\mathbf{v},n]}a^*[\mathbf{v},n]\right\}. \tag{2.17}$$

In the case of integer values of displacement components, $\mathbf{d}[\mathbf{v},n] = \mathbf{v}n$ and $a[\mathbf{v},n] = 1$ in (2.11)-(2.17), thus a simplified version of the proposed solution is obtained. In this particular case, it is possible to express (2.17) in the following compact form:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\text{argmax}}\sum_{k_1=0}^{M_1-1}\sum_{k_2=0}^{M_2-1}\text{Re}\left\{Y\left[\mathbf{k},-\frac{\mathbf{u_k}^\mathsf{T}\mathbf{v}}{T_s}\right]\right\} \tag{2.18}$$

where

$$Y[\mathbf{k},q] = \sum_{n=0}^{N-1}\sum_{i=0}^{I-1} X[\mathbf{k},n]S_i^*[\mathbf{k},n]e^{-j2\pi qT_s n} \tag{2.19}$$

is the continuous-frequency FT of the temporal sequence $\left\{\sum_{i=0}^{I-1} X[\mathbf{k},n]S_i^*[\mathbf{k},n]\right\}$ in the continuous-frequency variable q. The function (2.19) can be used with $q = -\frac{\mathbf{u_k}^\mathsf{T}\mathbf{v}}{T_s}$ to obtain (2.18).

The estimated speed vector $\hat{\mathbf{v}}$ is specified by the coordinates of the maximum of the log-likelihood function defined in agreement with (2.17) as follows

$$J(\mathbf{v}) = \frac{1}{M_1 M_2} \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} \sum_{n=0}^{N-1} \mathrm{Re}\left\{ \sum_{i=0}^{I-1} X[\mathbf{k}, n] \cdot S_i^*[\mathbf{k}, n] e^{j2\pi \mathbf{u_k}^\mathsf{T} \mathbf{d}[\mathbf{v}, n]} a^*[\mathbf{v}, n] \right\} \tag{2.20}$$

where the normalization coefficient $1/M_1M_2$ is introduced to reduce the dynamic range of the log-likelihood function by several orders of magnitude without any impact on the final estimate.

As described in Section 2.2.2, inverse projective transformations can be applied to recover the original shape and size of the objects of interest, whose images can thus be considered constant over time. If the image of the objects of interest can be considered constant with respect to n, the definition in (2.20) can be simplified as follows:

$$J(\mathbf{v}) = \frac{1}{M_1 M_2} \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} \sum_{n=0}^{N-1} \mathrm{Re}\left\{ \sum_{i=0}^{I-1} X[\mathbf{k}, n] \cdot S_i^*[\mathbf{k}] e^{j2\pi \mathbf{u_k}^\mathsf{T} \mathbf{d}[\mathbf{v}, n]} a^*[\mathbf{v}, n] \right\}. \tag{2.21}$$

2.4 Numerical Results

In this section, the performance of the proposed algorithm is discussed on the basis of some experimental results directly obtained by maximizing (2.20) or (2.21). A set of synthetic and software generated videos is considered to preliminarily test the effectiveness of the method in a controlled environment, then a number of real-world videos, specifically recorded for this purpose, is analysed. As reference method, among the motion estimation techniques described in Section 2.1, the block matching algorithm [6, Ch. 4] has been selected for comparison due to its effectiveness, implementation simplicity and previous use in the literature, e.g., in [13] and [14].

Since reasonable assumptions about the range of values of the correct speed can be made, a simple grid search can be implemented to find the optimal

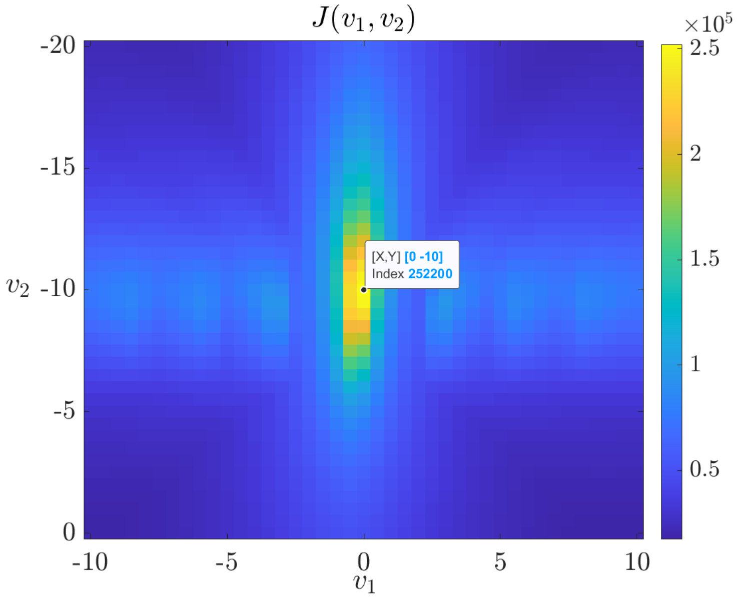
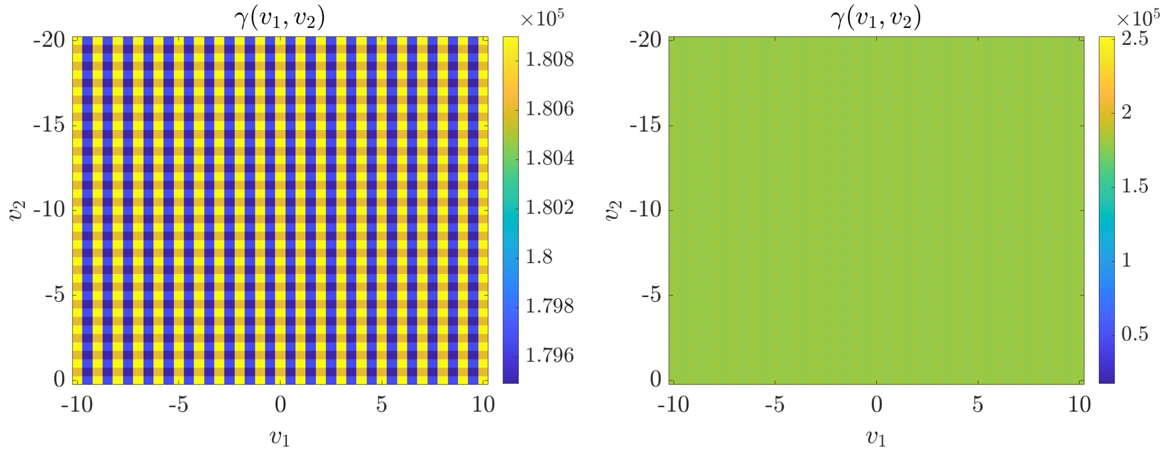


Figure 2.3: Example of the log-likelihood function (2.20) plotted versus speed components.

value of $\hat{\mathbf{v}}$. Other optimization methods, such as iterative gradient-search approaches, could also be considered to expedite the numerical solution [23].

For the sake of simplicity, scenes framing a single moving object are considered in this section, i.e., $I = 1$ in (2.20) and (2.21). This assumption is instrumental for the development of the proposed method. However, if multiple objects moving at different speeds are contained in the considered video sequence, the operations of object tracking and selection described in Section 2.2 can be exploited to extract the objects of interest, one at a time, and the proposed estimation method could be independently applied to each object.

As an illustrative example, the log-likelihood function in (2.20), for one of the studied cases, is displayed in Figure 2.3 versus the components of the speed vector $\mathbf{v}$, with a grid resolution of 0.5 pixel/frame for both components (i.e., $F = 2$ in (2.11)). The peak of the function, whose coordinates indicate



(a) (b)

Figure 2.4: Term in (2.22) plotted versus speed components for two different visualization scale ranges: (a): magnified range, (b) range equal to that in Fig. 2.3.

the estimated speed value, is highlighted.

For completeness, the log-likelihood function in Figure 2.3 is now compared with the quadratic term in (2.15), that has been neglected to derive the approximate likelihood function (2.20). Accounting for the proper scaling factor 1/2, this term is:

$$\gamma(\mathbf{v}) = \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} \sum_{n=0}^{N-1} \frac{1}{2} \left| \sum_{i=0}^{I-1} S_i[\mathbf{k}, n] a[\mathbf{v}, n] \right|^2 . \tag{2.22}$$

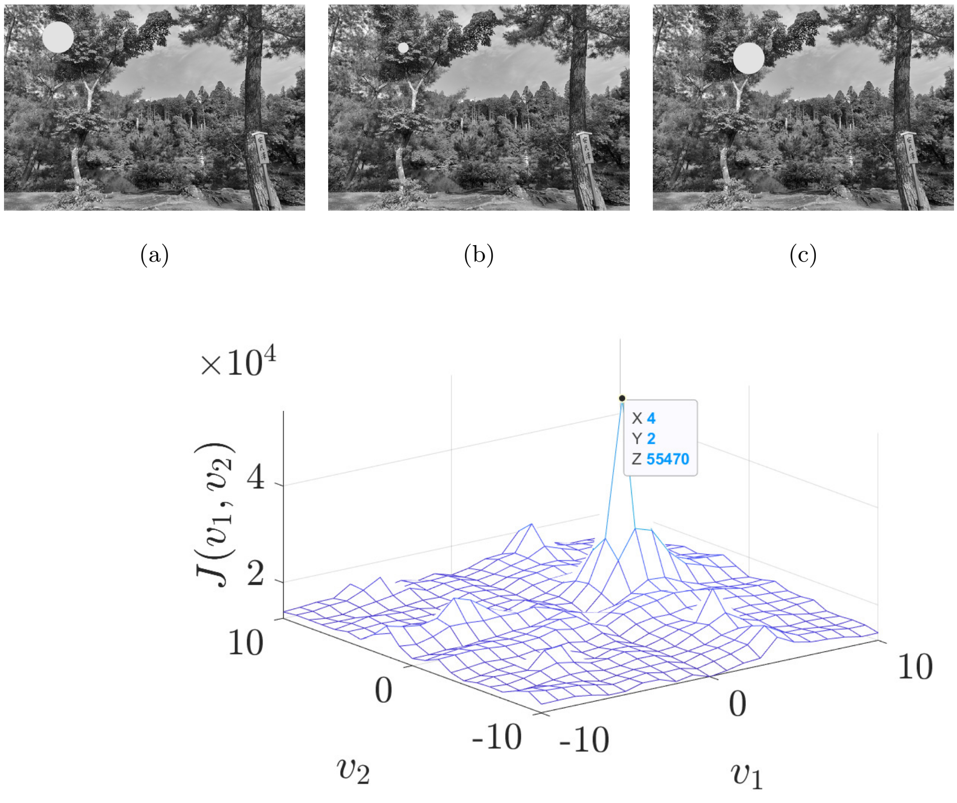
The term in (2.22), obtained in agreement with the log-likelihood function shown in Figure 2.3, is plotted in Figure 2.4, where two different scale ranges are considered for the sake of visualization. Note that, for the considered resolution of 0.5 pixel/frame, only four values of $a[\mathbf{v}, n]$ and $\gamma(\mathbf{v})$ are obtained in expressions (2.11) and (2.22). In particular, these values are subject to very small variations with respect to the approximate log-likelihood function in (2.20). For this reason, the variations of the term $\gamma(\mathbf{v})$ in (2.22), are better visible when a limited visualization scale range is selected, as in Figure 2.4(a), where the four values can be easily observed. On the other hand, when the broad scale range of the function in (2.20), shown in Figure 2.3, is used to visualize the small variations of $\gamma(\mathbf{v})$, as in Figure 2.4(b), the practically constant behaviour of this term is clearly visible. This comparison demonstrates the accuracy of the presented approximate ML estimator.

2.4.1 Synthetic video sequences

The performance of the proposed algorithm is firstly assessed on synthetic video sequences framing geometrical objects subject to periodic and non-periodic modifications superimposed to their translation. These videos were software-generated by setting a number of parameters including the frame size $M_1 \times M_2$, the frame rate, set as $f_s = 15$ Hz, and the video duration D [s]. The number of frames was derived as $N = Df_s$. Gaussian noise with spatially and temporally i.i.d. elements was superimposed to each of the considered sequences in order to emulate a more realistic scenario. Since the pixel intensity values belong to the interval $[0, 1]$, a noise variance $\sigma^2 = 0.025$ is set as an illustrative value in the following examples. The speed vector components $\mathbf{v} = (v_1, v_2)$ were also set and integer values of displacement components were considered; hence, (2.18) was applied to estimate the speed, regardless of the superimposed geometrical transformation. Real pictures were set as the static background which was removed by the GMM before applying (2.18) in all the presented examples.

Linear and periodic motion

An example of superimposed linear and periodic motion is first introduced. Three sample frames of a considered video sequence are shown in Figures 2.5(a), 2.5(b) and 2.5(c), which correspond to the 41st, 50th and 60th frames of the sequence, respectively. The frames have size 432×576 pixels. The video has a duration of about 7 s. The number of observed frames is $N = 75$, which



(a)

(b)

(c)

(d)

Figure 2.5: Results for linear with superimposed periodic motion: (a) frame at $n = 41$, (b) frame at $n = 50$, (c) frame at $n = 60$, (d) log-likelihood function where the estimated speed components are highlighted at the coordinates of the maximum peak at $(4, 2)$.

follow 40 background frames, exploited by the GMM algorithm. In the sequence, the geometrical moving object, a circle, is subject to linear motion with superimposed periodic transformation of its radius with sinusoidal behaviour between 10 and 30 pixels and period of 20 frames, and is shifting with a speed $\mathbf{v} = (4, 2)$ pixel/frame.

The result obtained by applying (2.18) to estimate the speed of the linear component of the motion is shown in Figure 2.5(d), where the log-likelihood function (2.20) is plotted against the speed components v_1 and v_2. The estimated components are highlighted at the coordinates of the maximum peak of the function at $(4, 2)$ and correspond to the values set for the video generation, thus the estimate of the vector $\mathbf{v}$ corresponds exactly to the speed simulated in the video stream.

Moving pendulum

As a second experiment, another case of object subject to linear motion that shows a periodic behaviour is considered. Three samples of a considered video are shown in Figures 2.6(a), 2.6(b) and 2.6(c). The frame size is again 432×576 pixels. The video duration is about 7 s and the first 40 background frames are followed by $N = 75$ observed frames. The object of interest is represented by a parallelogram, which is rotating with an approximate angle between -50 and 50 degrees around a selected vertex with a period of 10 frames and simultaneously shifting with a speed $\mathbf{v} = (5, 1)$ pixel/frame.

The estimated values of the speed components are highlighted at the coordinates of the maximum peak of the log-likelihood function (2.20), which is plotted in Figure 2.6(d) against v_1 and v_2. The peak of the function is at $(5, 1)$, which correspond to the correct values of the speed components.

Linear and non-periodic motion

As a third example, a more general scenario is analysed. Three frames of a considered video sequence are depicted in Figures 2.7(a), 2.7(b) and 2.7(c). In this case, the framed object, a parallelogram, is subject to a linear shift with superimposed non-periodic scaling in order to emulate the perspectival transformation of a vehicle moving on a road in the camera direction. The frame size of the considered video is 375×500 pixels and the video duration is about 5 s.The number of background frames is 40, as in the previous examples.

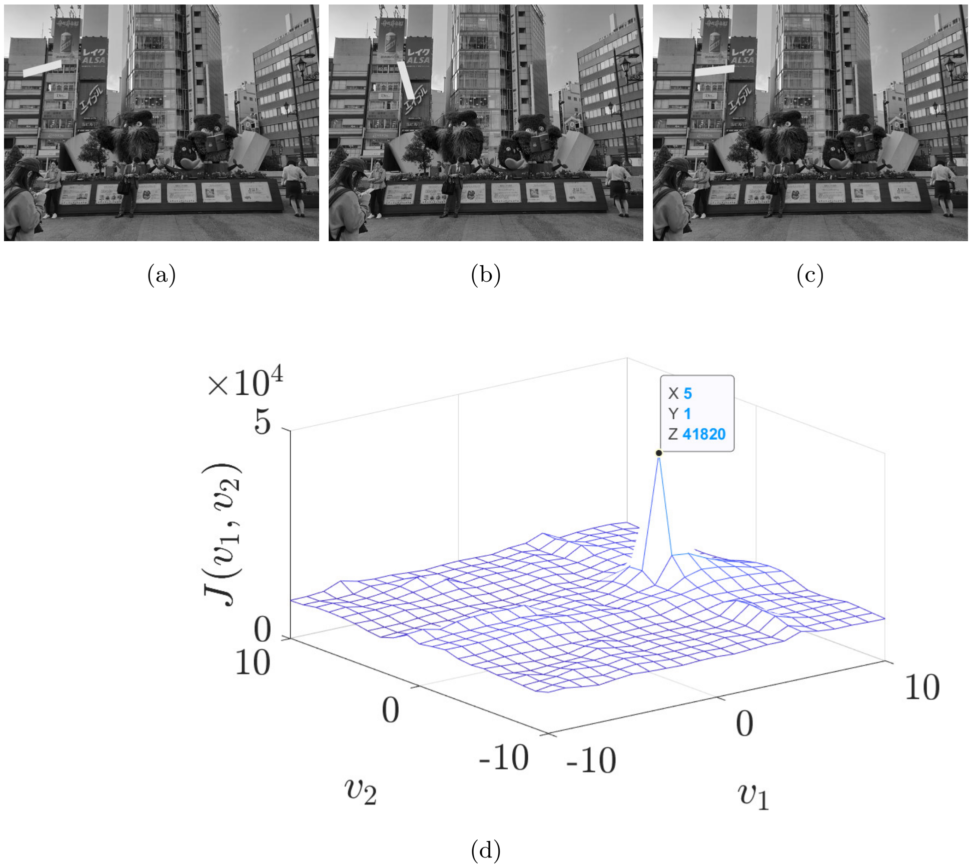
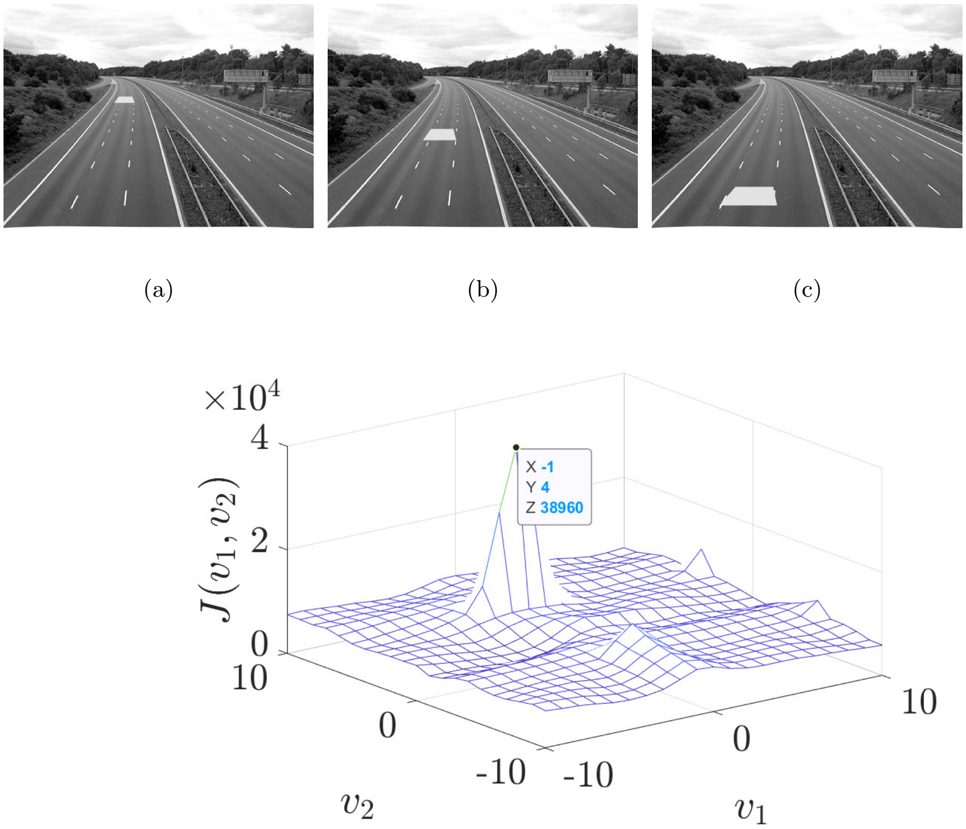


Figure 2.6: Results for moving pendulum: (a) frame at $n = 41$, (b) frame at $n = 45$, (c) frame at $n = 50$, (d) log-likelihood function where the estimated speed components are highlighted at the coordinates of the maximum peak at $(5, 1)$.

The number of observed frames is $N = 45$ and the object is translating with a speed $\mathbf{v} = (-1, 4)$ pixel/frame.

The log-likelihood function (2.20) is plotted against the speed components v_1 and v_2 in Figure 2.7(d) and the estimated result is highlighted at the coordinates of its maximum peak at $(-1, 4)$. This result confirms the effectiveness



(a) (b) (c)

(d)

Figure 2.7: Results for linear shift with superimposed non-periodic motion: (a) frame at $n = 41$, (b) frame at $n = 55$, (c) frame at $n = 80$, (d) log-likelihood function where the estimated speed components are highlighted at the coordinates of the maximum peak at $(-1, 4)$.

of the proposed algorithm if the shifting object is undergoing a generic time-variant transformation.

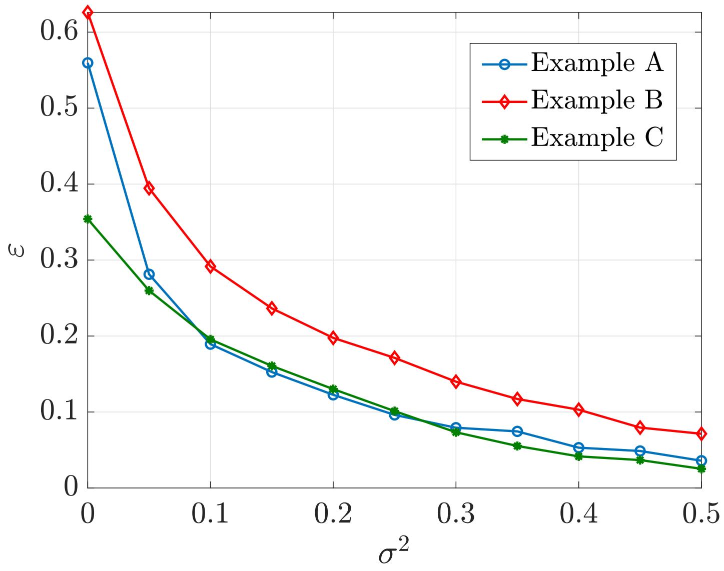


Figure 2.8: Normalized noise margin versus noise variance.

Performance analysis

As a first indication of the effectiveness of the proposed algorithm, its performance is analysed for increasing value of noise variance σ^2. In particular, a normalized noise margin is defined in terms of the log-likelihood function (2.20) as:

$$\varepsilon = \frac{J(\mathbf{v}) - J(\tilde{\mathbf{v}})}{J(\mathbf{v})} \tag{2.23}$$

where $\tilde{\mathbf{v}}$ is the speed value corresponding to the largest peak of the log-likelihood function in (2.20) excluding the peak at the correct $\mathbf{v}$. The value of ε in (2.23) is plotted in Figure (2.8) against the noise variance σ^2 for the three examples presented in the previous sections. Examples A, B and C in Figure (2.8) refer to the linear and periodic motion, moving pendulum and linear and non-periodic motion examples, respectively. As shown by the decreasing trend of the three curves in Figure (2.8), the difference between the correct peak $J(\mathbf{v})$ and the peak $J(\tilde{\mathbf{v}})$ is lower for higher values of noise variance. Nevertheless, this value is always positive for all the considered cases. This confirms that the peak $J(\tilde{\mathbf{v}})$ is lower than the correct one, hence the estimated speed values are correct also for high noise variance, despite with a reduced margin against noise.

2.4.2 Real videos

To further analyse the performance of the proposed estimation algorithm, 12 more sequences extracted from 6 real videos specifically recorded, i.e., 6 sets of videos, were tested. As in the previous examples, Gaussian noise with spatially and temporally i.i.d. elements is superimposed to each video sequence to test the robustness of the algorithm against noise. In particular, 10 different noise realizations are considered. To smooth the noise effect, a spatially averaging filter with size 7×7 pixel is applied.[1] Various camera angles and locations are examined in order to assess the robustness of the proposed method in different perspectival conditions. The main characteristics of the dataset are summarized in Table 2.1, which reports the camera height h, the camera pointing direction, the number of sequences extracted from each video set and the correct speed per sequence. The camera pointing direction is defined by the horizontal angle α between the camera and the road axes and the vertical inclination β of the camera axis with respect to the horizontal plane. Side and top views of the camera positioning are shown in Figure 2.9, where the height h and the angles α and β are highlighted. Reference speeds are expressed in pixel/frame and computed after the application of the inverse projective transformations.

The preprocessing operations described in Section 2.2.1 need to be calibrated for each set of videos recorded with the same camera setting, position and location.

The proper inverse projective transformation is computed for each set of

[1] The size of 7×7 pixel of the averaging filter was selected as compromise by trial and error.

	h [m]	α	β	# sequences	$\mathbf{v}$ [pixel/frame]
Set 1	7	0°	10°	4	$(0, 22.2)$ $(0, 22.2)$ $(0, -20.6)$ $(0, -19.7)$
Set 2	6.20	0°	10°	2	$(0, 3.5)$ $(0, -3.3)$
Set 3	15.20	30°	20°	2	$(0, 3.5)$ $(0, 5.4)$
Set 4	15.20	20°	15°	2	$(0, -9.6)$ $(0, 9.3)$
Set 5	11.70	45°	30°	1	$(-16.4, 0)$
Set 6	9	50°	20°	1	$(3.6, 0)$

Table 2.1: Dataset parameters.

videos and background removal is performed according to the sequence of operations in the block (A) of Figure 2.1, as it was observed that this method is more robust against noise with respect to the GMM previously employed in Section 2.4.1. The shape and dimension of the structuring elements of the morphological operations is set by trial and error. In particular, the structuring elements of the erosion and dilation operations are respectively defined as squares of size 8 and 9 pixel, and diamonds of size 5 and 6 pixel. Multiple cars moving at different speeds are captured in some of the recorded videos. However, since they are not simultaneously framed or move in distinct regions of the road plane, it is sufficient to trim and crop the sequences to focus on a single vehicle at a time, i.e, object tracking and selection is not necessary. The duration of the trimmed video sequences is variable and ranges from 25 to 165 frames. Also, the size and the frame rate depend on the setting of the employed recording device. In particular, here the frame rate is set as $f_s = 25$

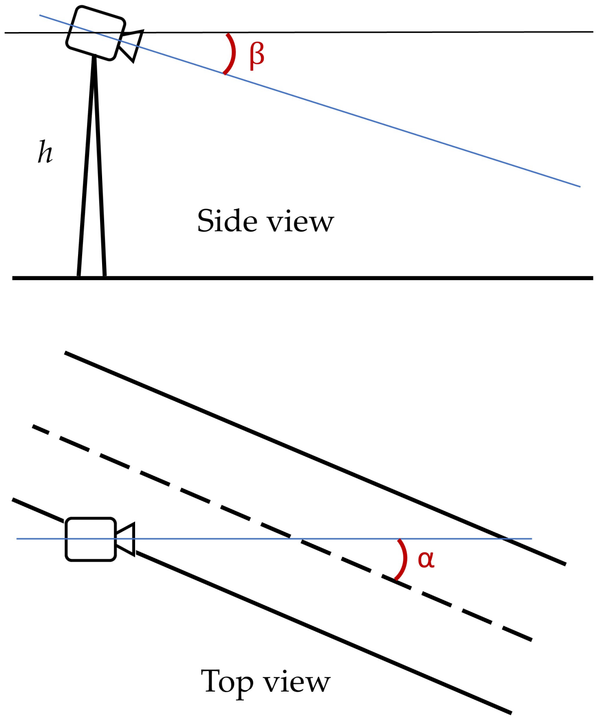


Figure 2.9: Side and top views of the camera positioning, where the height h and the angles α and β are highlighted.

or 30 Hz, whereas the frame size for all videos is converted to a fixed size of 800×300 or 300×800 pixel after the inverse projective transformation.

In Figures 2.10 and 2.11, a few illustrative examples of original and processed frames of two considered sequences are shown. In particular, columns correspond to different frame indices and rows (a), (b) and (c) indicate the original sequences, the processed sequences after the inverse projective transformation and those after convex hull extraction, respectively. The two sequences are referred to as Sequence 1 (Figure 2.10) and Sequence 2 (Figure 2.11), for brevity, respectively belonging to Sets 3 and 6 of Table 2.1.

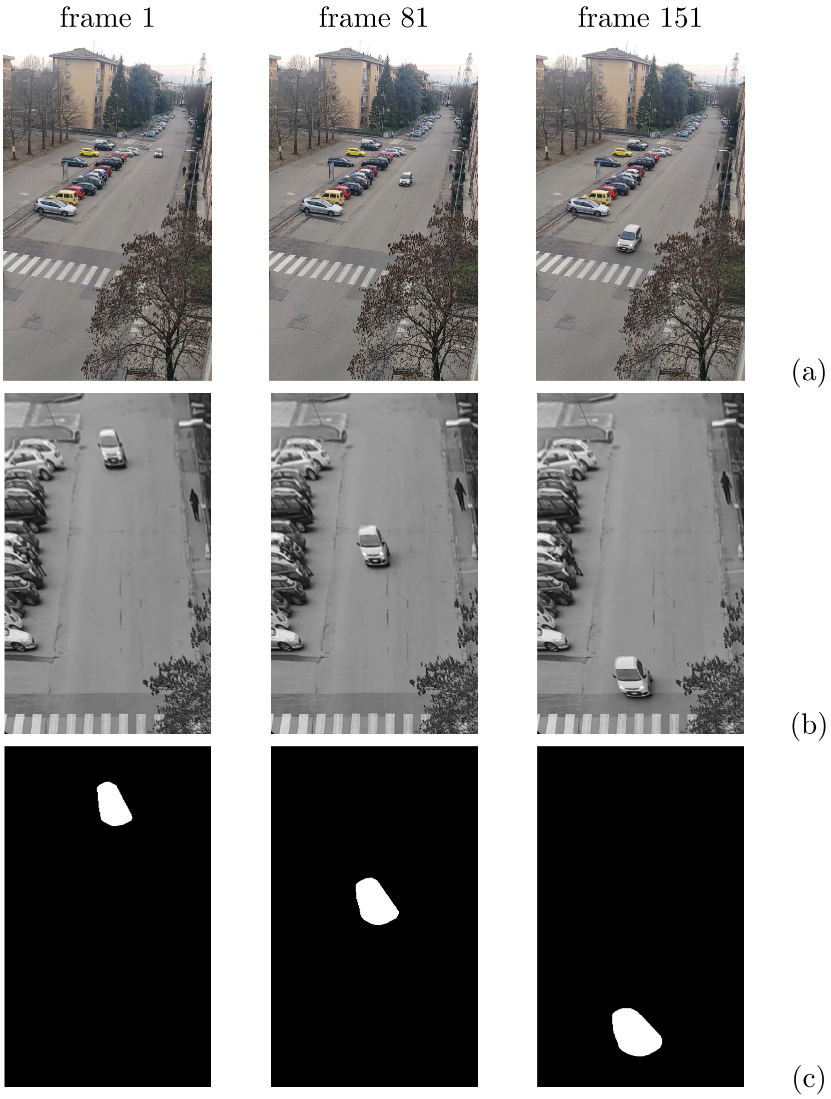


Figure 2.10: Sample frames of: (a) original Sequence 1, (b) processed sequence after the inverse projective transformation, (c) processed sequence after background removal and convex hull extraction.

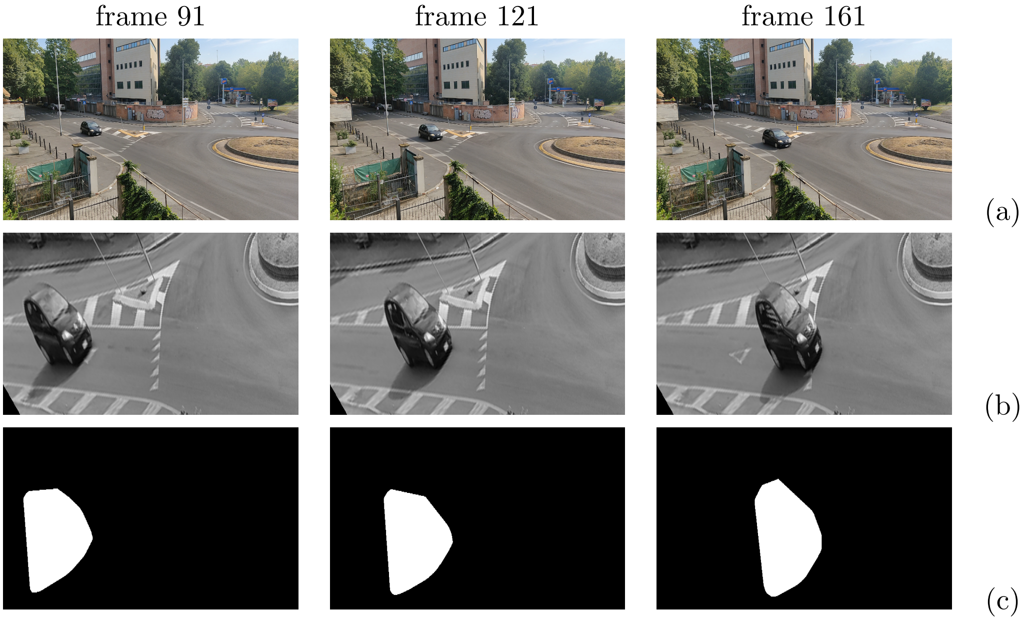


Figure 2.11: Sample frames of: (a) original Sequence 2, (b) processed sequence after the inverse projective transformation, (c) processed sequence after background removal and convex hull extraction.

The estimated speed values are expressed in pixel/frame and can be converted to real world measurement units, such as km/h, by camera calibration and a suitable conversion rule, e.g.,

$$\hat{\mathbf{v}}_{km/h} = \hat{\mathbf{v}} \cdot \frac{l_m}{l_{px}} \cdot f_s \cdot 3.6 \qquad (2.24)$$

where l_m and l_{px} represent a reference length in the real world, expressed in meters, and its projection onto the 2D scene, expressed in pixels, respectively. The reference length l_m could be any known element of the real world scene, e.g., the road length or width.

Performance analysis

The performance of the estimation method is now analysed in terms of Root Mean Squared Error (RMSE) between the estimated speed vector and the correct speed, which is manually measured in pixel/frame by video inspection. Considering R different noise realizations and J analysed videos, the RMSE normalized to the Root Mean Squared (RMS) value of the correct speed can be obtained by averaging over the noise realizations and video sequences as

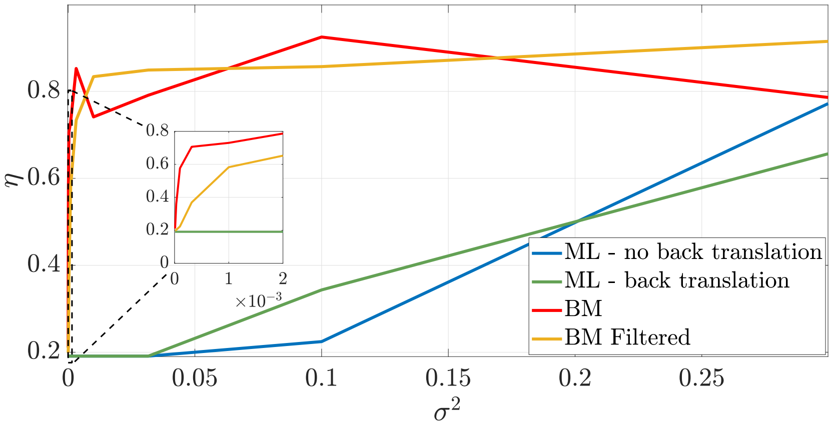
$$\eta = \sqrt{\frac{\sum\limits_{r=1}^{R}\sum\limits_{j=1}^{J}\left|\hat{\mathbf{v}}_{r,j}-\mathbf{v}_j\right|^2}{R\sum\limits_{j=1}^{J}|\mathbf{v}_j|^2}} \tag{2.25}$$

where $\mathbf{v}_j$ and $\hat{\mathbf{v}}_{r,j}$ are the correct speed components for the j-th video and the estimated speed components for the j-th video and the r-th noise realization, respectively. All speeds are measured in pixel/frame.[2]

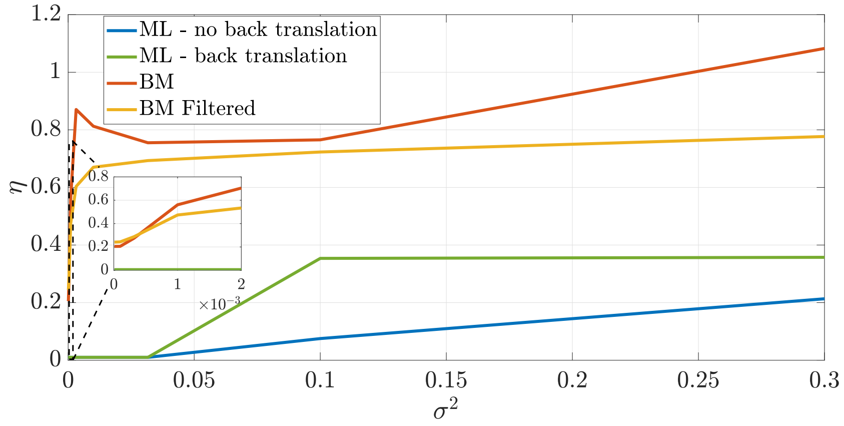
The obtained results are compared with the performance of the reference block-matching approach [6, Ch. 4]. The parameters for the block matching method also need to be set. In particular, the block size is set to 75 × 75 or 105 × 105 pixel by trial and error depending on the object size. Unlike the proposed ML estimation algorithm, the block matching approach is applied to the considered processed video sequences where the background extraction and removal operations are not performed. The presence of background provides, indeed, texture information about the diversity of blocks that helps the block matching function to avoid undesired mismatches.

At first, the normalized RMSE η in (2.25) is computed for both methods for increasing values of the noise variance σ^2 for single video sequences and a small set of video sequences in which the same camera viewpoint, position and location are preserved. The overall performance is finally evaluated on all considered scenarios for increasing values of the peak signal to average noise

[2] As the RMSE in (2.25) represents a relative error, the selected unit of measurement does not affect the results.



(a)



(b)

Figure 2.12: Performance of the assessed estimation methods in terms of RMSE vs. noise variance for: (a) sample Sequence 1 (Figure 2.10) and (b) sample Sequence 2 (Figure 2.11).

power ratio, or Signal to Noise Ratio (SNR) for brevity, hence for decreasing values of the noise variance σ^2.

In Figures 2.12(a) and 2.12(b) the normalized RMSE η in (2.25) is shown against increasing values of noise variance for the two sample sequences shown in Figures 2.10 and 2.11, respectively. The RMSE η in (2.25) is hence computed with $J = 1$ and $R = 10$. The image of the foreground moving object can be considered constant in both examples, as can be observed in the rows (c) of both Figures 2.10 and 2.11, where the processed sequences are shown. The ML estimation method is tested by directly maximizing (2.20) or (2.21) and estimated speed components are searched over a quantization grid with fractional values of 0.5 pixel/frame.

When the image of the object of interest can not be considered constant with respect to n, (2.20) holds and a dedicated processing operation is necessary to obtain the sequence $S_i[\mathbf{k}, n]$. To this purpose, the moving object of interest is translated back to its original position by shifting each frame of the sequence $X[\mathbf{k}, n]$ by the object centroid computed at the n-th frame after the convex hull extraction operation depicted in Figure 2.1. However, if the image of the foreground object can be considered constant, as in this case, implementing expression (2.20) may be unnecessary and computationally expensive. In Figure 2.12, results obtained by applying (2.20) and (2.21) are referred to as “ML - back translation” and “ML - no back translation”, respectively. Both options are here analysed for the sake of completeness and the trend of the respective RMSE curves plotted in Figure 2.12 confirms that their performance is equivalent, especially for low values of noise variance.

On the other hand, the block matching method is tested with and without the application of the spatial 7 × 7 pixel average filter. These variations of the algorithm are indicated in Figure 2.12 as “BM Filtered” and “BM”, respectively. According to the results in Figure 2.12, this filter has a positive effect for values of interest of the noise variance, i.e., those considered in the figure insets. It can also be observed that the RMSE obtained with the block-matching method does not reach zero even in the absence of noise, conversely to the ML curves in Figure 2.12(b). In this specific case, the RMSE curves for the ML-based approaches are always far better than the ones obtained by the block-matching method. For the sake of visualization, the rapidly increasing

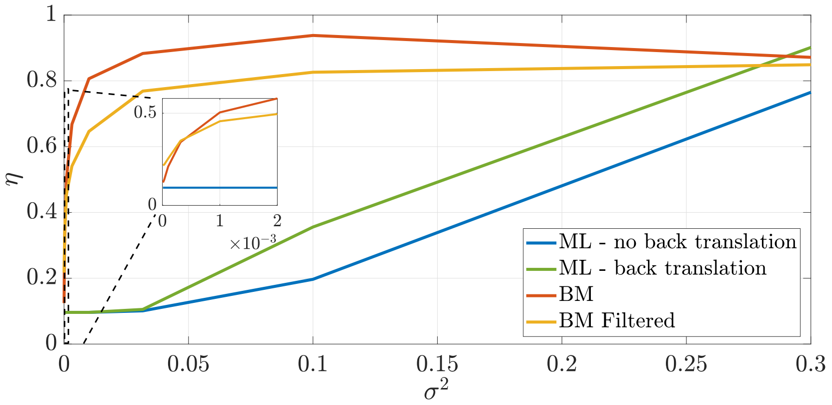


Figure 2.13: Performance of the assessed estimation methods in terms of RMSE vs. noise variance for a set of two video sequences where the same camera angle, position and location are set.

trend of the curves obtained with the block-matching method for increasing noise variance σ^2 can be better observed in the insets depicted in Figures 2.12(a) and 2.12(b).

As a further analysis, the normalized RMSE η in (2.25) is also computed for increasing noise variance σ^2 on a set of two video sequences where the same perspectival conditions (i.e., camera viewpoint, position and location) are set. In this case, $J = 2$ and $R = 10$ in (2.25). The obtained results are shown in Figure 2.13 and confirm the performance observed in the previous Figure 2.12.

In Figure 2.14, the normalized RMSE η in (2.25) is finally shown against increasing values of the SNR, defined as $1/\sigma^2$, for all considered scenarios, i.e., $J = 12$ and $R = 10$. Both ML curves tend to stabilize around 0.07 because of the error introduced by the used quantization level. This value is in agreement with an estimate of normalized RMSE obtained by assuming uniformly distributed quantization error over the range $[-0.5, 0.5)$ in each dimension which, for the given video sequences, is about 0.04.

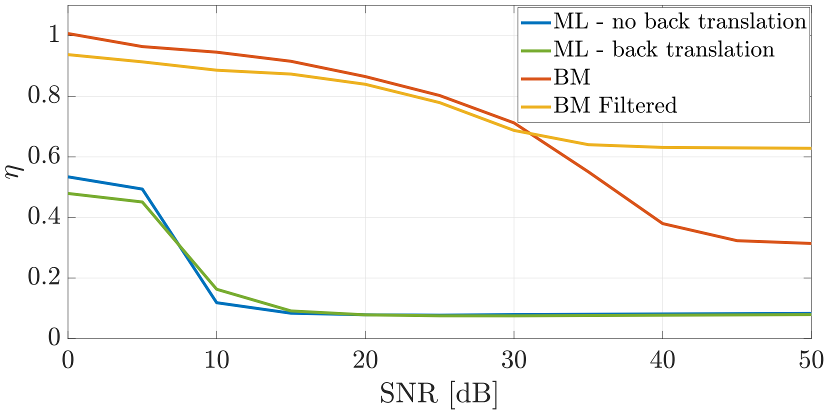


Figure 2.14: Performance of the assessed estimation methods in terms of RMSE vs. SNR.

Observing the curves obtained for the block-matching method, the mentioned effect of the spatial average filter is confirmed: it is slightly positive for low values of SNR, but excessively smoothing at high ones. The performance of the block-matching approach in both cases is far below the proposed ML estimation method. The presence of noise, even when comparatively low, prevents indeed the block-matching algorithm from correctly detecting and matching blocks. The block size and repetitive patterns, such as road lines, which are present in some of the analysed scenarios, are critical aspects which impair significantly the overall performance of the block-matching algorithm.

The effectiveness of the proposed ML estimation method with respect to the block-matching approach is thus demonstrated in the considered heterogeneous set of realistic videos accounting for different perspectival views. Thanks to sound pre-processing operations, the presented method is robust against noise, achieving low values of RMSE also for low values of SNR and high values of noise variance σ^2.

2.5 Conclusions

In this chapter a novel method to estimate the speed of foreground objects in video signals is proposed. A model to describe the motion of objects undergoing dynamic changes, such as perspective transformations, is derived, proper pre-processing operations are defined and the ML principle is applied to obtain an estimator of the speed of the framed objects.

The proposed method is composed of robust video pre-processing stages followed by the speed estimation algorithm. Its performance is evaluated on synthetic and real video sequences also in the presence of noise and a comparison with the well-known block matching approach, that is subject to some major limitations, is presented. The tested video sequences differ in camera viewpoint, position and location in order to include in the analysis scenarios affected by different perspective transformations. The effectiveness of the proposed method is finally analysed on a number of experimental videos demonstrating its good and robust performance.

Chapter 3

Breathing Monitoring

Breathing monitoring is a fundamental diagnostic tool to assess the physiological status of a patient. In particular, the Respiratory Rate (RR) is a main indicator of potential dysfunctions of the human respiratory system that may be caused by critical medical conditions. Typical values of the RR in healthy adults at rest lie between 12 and 20 breaths per minute and may vary with age. The RR in newborns and children is usually higher. Abnormal values of the RR may be a sign of severe issues arising from respiratory disorders or complications. For instance, diseases such as chronic obstructive pulmonary disease, asthma, anaemia and epileptic seizures may cause oxygen levels in the blood to significantly drop, potentially leading to cyanosis, cerebral palsy or cardiac arrest and ischaemic events [24].

In this chapter, two video processing techniques for contactless estimation of the RR of framed subjects are presented. Due to the modest extent of movements related to respiration in both infants and adults, specific algorithms to efficiently detect breathing are needed. For this reason, motion-related variations in video signals are exploited to identify respiration of the monitored patient and simultaneously estimate the RR over time. The proposed methods rely on two motion magnification algorithms that are exploited to enhance the subtle respiration-related movements. In particular, amplitude- and phase-based algorithms for motion magnification are considered to extract reliable motion signals. The proposed estimation systems perform both spatial decomposition of the video frames combined with proper temporal filtering to extract breathing information. After periodic, or quasi periodic, respiratory signals are extracted and jointly analysed, the Maximum Likelihood (ML) criterion is applied to estimate the fundamental frequency, corresponding to the RR. The performance of the presented methods is first assessed by comparison with reference data evaluated on videos framing different subjects, i.e., newborns and adults. Finally, the accuracy of both methods is measured and analysed in terms of normalized Root Mean Squared Error (RMSE).

The chapter is organized as follows. In Section 3.1, an overview on traditional breathing monitoring techniques is provided. In Section 3.2, amplitude- and phase- based procedures to extract amplified motion signals are presented. In Section 3.3, the RR estimation approach is detailed along with a method to select specific Regions of Interest (ROIs). In Section 3.4, the performance of the considered methods is discussed on the basis of numerical evaluations performed on real videos and is compared against reference data. Finally, in Section 3.5 conclusions are drawn.

3.1 Breathing Monitoring Systems

Respiration monitoring systems are typically classified into two main categories: contact-based and contactless [24]. Contact-based techniques mainly include expensive conventional instrumentation typically deployed in clinical settings and hardly adaptable to domestic environments. On the other hand, innovative approaches based on video processing techniques represent contactless solutions of great interest due to their attainability. An overview of existing breathing monitoring methods belonging to both categories is presented hereafter.

3.1.1 Contact-based techniques

Traditional contact-based methods for respiration monitoring require the direct contact of a sensor with the body of the patient, being often moderately intrusive and uncomfortable, for both adults and newborns. The pneumography [25] and phlebotomy [26] are examples of invasive procedures that respectively allow to measure the thoracic movements, related to the respiration act, and to sample arterial, capillary or venous blood gas to analyse the pulmonary activity. Besides its high accuracy, phlebotomy may be painful and difficult to perform, especially in children and newborns, and may lead to complications such as thrombosis, haemorrhage and aneurysm formation [26]. Pneumography is, instead, performed by means of the Pneumogram, that is composed by wired electrodes to be directly attached to the chest of the patient [25].

Recording the cardiac activity may also be useful to monitor some respiratory disorders. In particular, the polysomnography technique [27] may be adopted for sleep monitoring in order to detect apneas or seizure events. A Polysomnograph is composed by several systems, including the ElectroCardioGram (ECG), that allows to measure the heart electrical activity through electrodes attached on the chest of the patient. The main limitation of these instruments is their deployment, as it is mainly limited to clinical settings, being not suitable for home care.

Another example of contact-based devices is the pulse oximeter, that has become very popular nowadays as it allows to easily measure the oxygen saturation in the blood, also in domestic environments. The pulse oximeter is usually clipped to the fingertip of a patient and measures the changes in the transmission or reflection of the light emitted by a light-emitting diode (LED) hitting the skin of the subject. This working principle is referred to as PhotoPlethysmoGraphy (PPG) and has also inspired some contactless video-based monitoring methodologies, as discussed in the following.

3.1.2 Contactless techniques

Video processing solutions as contactless techniques for vital signs monitoring present several advantages made available by the use of specific sensors. In particular, contactless devices include Red, Green and Blue (RGB) and Infra Red (IR) thermal cameras, among other sensors [28], whose cost is significantly lower than sophisticated equipments, usually deployed in hospital environments. These instruments are also non-invasive, hence more comfortable, as they do not require a direct contact with the body of the patient.

In [29], RGB sensors are considered to extract the respiratory frequency from selected PPG signals computed on a region that surrounds the pit of the neck of the subject. The RR is estimated in frequency and time domain and the performance of different RGB camera sensors is analysed. The PPG principle is also exploited in [30], but the hue channel of the Hue, Saturation, Value (HSV) colour space is considered for the analysis. Also the optical flow principle may be exploited to detect and track breathing related movements in video sequences, as in [31, 32]. However, since these movements may be subtle and difficult to detect, especially in newborns, motion magnification techniques may be applied to enhance them, as in [33, 34].

A novel promising method for the RR extraction based on thermography is, instead, proposed in [35]. Thermal cameras allow to monitor the air temperature at a specific region considered near the nostrils/mouth of a patient. A difference in terms of temperature can indeed be observed in the inhaled and exhaled air. Temperature changes induce pixel intensity changes in thermal images, hence breathing information, such as RR, can be obtained by analysing pixel intensity variations in a sequence of thermal images.

Other contactless devices, whose employment is analysed in [28], include depth and radar sensors.

3.2 Motion Signal Extraction

In this section, two novel video processing methods to extract respiratory signals based on motion magnification algorithms are presented. Amplitude- and phase-based techniques, respectively inspired by the works in [36, 37], are considered. In particular, in [36] spatial and temporal processing is combined to amplify the variations of the pixel intensities for frequency bands of interest. In [37] an approximation of the Riesz transform is proposed to perform phase amplification of motion signals. The two methods presented in this section are based on preliminary works appearing in [38–40].

3.2.1 Amplitude-based motion magnification

Amplitude-based techniques for motion magnification aim at linearly amplifying variations of each pixel intensity over time. The method proposed in [36], called Eulerian Video Magnification (EVM), performs temporal processing on different spatial frequency bands obtained by decomposing each frame of the input video into a set of subimages. The processed and unprocessed video subsignals so obtained are finally recombined to obtain the amplified output video. In this section, a spatio-temporal approach to extract motion signals inspired by the EVM algorithm in [36] is presented. Related preliminary works appeared in [38, 39], where the final signal recombination is not performed because not of interest for the purpose of breathing monitoring. An illustrative overview of the method is shown in Figure 3.1, where each processing step is associated to a diagram block and is detailed hereafter.

Spatial decomposition Consider a generic grayscale video sequence $f[\mathbf{u}, n]$. As a first step, each frame of the video $f[\mathbf{u}, n]$ is decomposed into a set of M subimages with scaled resolutions, each representing a different spatial frequency band. The M scaled subimages, referred to as "levels", are obtained by computing a Laplacian pyramid [41] and are sorted with decreasing resolution. A Laplacian pyramid is formed as follows. Firstly, a Gaussian

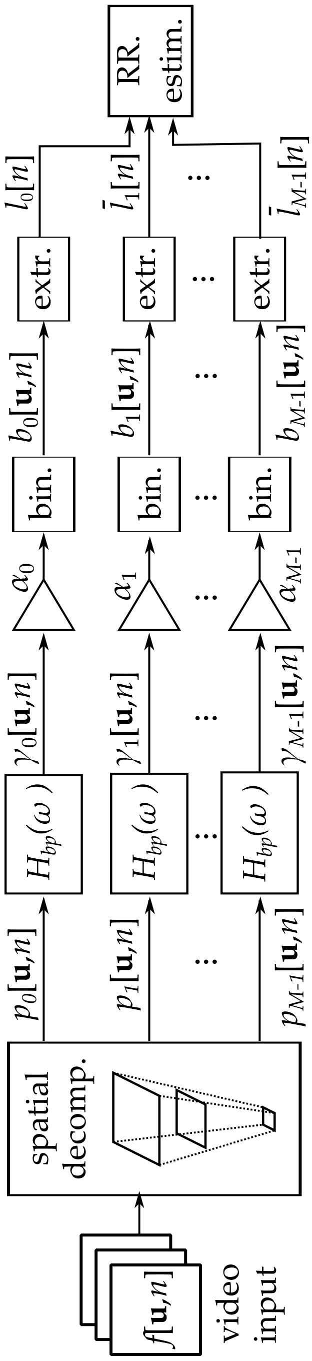


Figure 3.1: Amplitude-based (spatio-temporal) RR estimation algorithm.

pyramid [41] is derived, where $g_0[\mathbf{u}, n] = f[\mathbf{u}, n]$ is set as the bottom level that corresponds to the highest spatial frequency band and is characterized by the highest resolution. Upper levels, representing lower spatial frequency bands and characterized by lower resolution, are recursively computed according to a “reduce” function defined as

$$g_m[\mathbf{u}, n] = \sum_{k_1=-R_M}^{+R_M} \sum_{k_2=-R_M}^{+R_M} w[k_1, k_2] g_{m-1}[2u_1 - k_1, 2u_2 - k_2, n] \tag{3.1}$$

where $m = 1, \dots M - 1$ denotes the m-th pyramid level, $w[k_1, k_2]$ is a proper truncated Gaussian low-pass filter that is designed according to specific constraints described in [41], and R_M is a positive integer that specifies the size of this filter as $(2R_M + 1) \times (2R_M + 1)$. An “expand” function can also be defined as

$$\hat{g}_m[\mathbf{u}, n] = 4 \sum_{k_1=-R_M}^{+R_M} \sum_{k_2=-R_M}^{+R_M} w[k_1, k_2] g_{m+1}\left[\frac{u_1 - k_1}{2}, \frac{u_2 - k_2}{2}, n\right] \tag{3.2}$$

to obtain a specific level by expanding the dimensions of the lower one by interpolation. The filter mask $w[k_1, k_2]$ is the same in (3.1) and (3.2).

The Laplacian pyramid levels are derived from (3.1) and (3.2) as

$$p_m[\mathbf{u}, n] = \begin{cases} g_m[\mathbf{u}, n] - \hat{g}_m[\mathbf{u}, n] & m = 1, \dots, M - 2 \\ g_m[\mathbf{u}, n] & m = M - 1 \end{cases} \tag{3.3}$$

where $p_{M-1}[\mathbf{u}, n] = g_{M-1}[\mathbf{u}, n]$ is set as the highest-index level and describes the lowest spatial frequency band. The expression in (3.3) represents the error image between a level of the Gaussian pyramid g_m and the same level $\hat{g}_m$ obtained by expanding the upper one according to the function in (3.2).

The operation of spatial decomposition is highlighted in the first block of the diagram in Figure 3.1.

Temporal filtering Once the spatial processing is performed and a spatial decomposition based on the Laplacian pyramid is obtained, each level is pixel-wise temporally filtered to extract a frequency band that corresponds to a

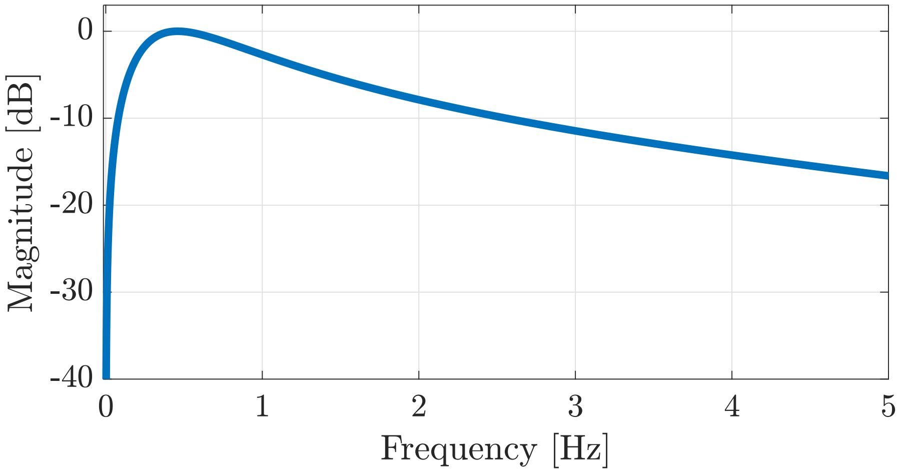


Figure 3.2: Frequency response of the IIR filter employed for adults.

typical range of RR. A Butterworth filter of the second order with Infinite Impulse Response (IIR) can be selected as a proper temporal digital band-pass filter. Its transfer function can be defined as

$$H_{bp}(z) = K\frac{(1+z^{-1})(1-z^{-1})}{(1-pz^{-1})(1-p^{*}z^{-1})} \tag{3.4}$$

where the scale factor K and the complex conjugates poles p and p^{*} can be computed following the filter design rules to fit the requirements for the lower and upper 3-dB cut-off frequencies $f_{\mathrm{L}}^{\mathrm{co}}$ and $f_{\mathrm{H}}^{\mathrm{co}}$ [42]. The cut-off frequencies of the filter are set according to to the framed subject: for adults $f_{\mathrm{L}}^{\mathrm{co}} = 0.19$ Hz and $f_{\mathrm{H}}^{\mathrm{co}} = 0.9$ Hz, corresponding to a range of $11 - 54$ breaths per minute, whereas for newborns $f_{\mathrm{L}}^{\mathrm{co}} = 0.3$ Hz and $f_{\mathrm{H}}^{\mathrm{co}} = 1.1$ Hz, corresponding to a range of $18 - 66$ breaths per minute. The frequency response of the IIR filter employed for adults is shown in Figure 3.2.

The temporal processing is represented as a filter bank in Figure 3.1 and the obtained filtered levels are denoted as $\{\gamma_m[\mathbf{u}, n]\}_{m=0}^{M-1}$.

Signal amplification Each filtered level $\gamma_m[\mathbf{u}, n]$, $m = 0, \ldots, M-1$, is multiplied by a proper amplification factor to linearly amplify motions related to the respiration. The amplification coefficients are denoted $\{\alpha_m\}_{m=0}^{M-1}$ in Figure 3.1 and are properly set according to [36] to avoid noise amplification or motion artefacts. Values much larger than 1, e.g., 12, are used at high-index levels and are linearly attenuated for low-index ones. The amplification coefficient for the lowest-index level is set as $\alpha_0 = 1$ and increasing values of amplification are used for higher-index levels, up to $\alpha_{M-2} = 12$. As the highest-index level has too low resolution to provide useful information, α_{M-1} is set to 0.

Binarization Binarization is performed pixel-wise on the amplified signals $\{\gamma_m[\mathbf{u}, n]\alpha_m\}_{m=0}^{M-1}$ to reduce the computational complexity (blocks labelled "bin." in Figure 3.1). This operation allows to highlight the respiration movements by setting to 1 the pixel intensity values whose variation is due to respiratory movements, whereas the rest of the framed scene is set to 0. This operation yields the following binarized levels, also highlighted in Figure 3.1

$$b_m[\mathbf{u}, n] = \begin{cases} 1, & \text{if } |\{\gamma_m[\mathbf{u}, n]\alpha_m\}| \geq \mathit{\Gamma}_{\text{th}} \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

where $\mathit{\Gamma}_{\text{th}}$ is a proper binarization threshold heuristically set to adjust the sensitivity to motion.[1]

Signal extraction As a last step, the motion signals are extracted by spatial averaging each binarized level of the pyramid as

$$\bar{l}_m[n] = \frac{1}{c_m r_m} \sum_{u_1=1}^{c_m} \sum_{u_2=1}^{r_m} b_m[\mathbf{u}, n] \tag{3.6}$$

where $\{c_m\}_{m=0}^{M-1}$ $\{r_m\}_{m=0}^{M-1}$ are the widths and heights of the binarized frames (blocks labelled "extr." in Figure 3.1).

[1] The set of coefficients $\{\alpha_m\}_{m=0}^{M-1}$ and the threshold $\mathit{\Gamma}_{\text{th}}$ in (3.5) can be scaled by a common factor without affecting the binarized frames. This lability is overcome by setting $\alpha_0 = 1$.

3.2.2 Phase-based motion magnification

Amplitude-based motion magnification presents some limitations directly linked to the linear amplification operation. When the analysed motion is small, it can be approximated as pixel intensity variations by a first-order Taylor series expansion [43]. If the small motion condition is not verified, or the amplification factor α_m is too large, the approximation is not accurate and the magnification may cause undesired artefacts. Furthermore, for $\alpha_m > 1$, noise is also amplified.

A solution to overcome problems related to linear amplification is provided by phase-based magnification methods, that aim at amplifying the phase of each pyramidal subsignal [37]. In this section, an algorithm for motion magnification inspired by [37] is described. Related preliminary work appeared in [40]. An illustrative overview of the method is shown in Figure 3.3, in which each processing step is associated with a diagram block and will be detailed hereafter.

Spatial decomposition Similarly to the amplitude-based (spatio-temporal) method described in Section 3.2.1, the first step to extract amplified motion signals, as also highlighted in the first block of Figure 3.3, consists in decomposing each frame of the input video sequence $f[\mathbf{u}, n]$ into a set of M scaled levels by computing a Laplacian pyramid [41] according to (3.1)-(3.3). An efficient representation of the signals, where amplitudes and phases are highlighted, can now be adopted by computing the Riesz transform [44] of all the pyramid levels $\{p_m[\mathbf{u}, n]\}_{m=0}^{M-1}$. The Riesz transform can be defined as a two-dimensional (2D) generalization of the Hilbert transform and its 2D frequency response in the Fourier domain can be expressed as [45]

$$H(\boldsymbol{\omega}) = \begin{pmatrix} H_1(\omega_1) \\ H_2(\omega_2) \end{pmatrix} = \begin{pmatrix} -j\omega_1/||\boldsymbol{\omega}|| \\ -j\omega_2/||\boldsymbol{\omega}|| \end{pmatrix} \tag{3.7}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2)$ is the 2D vector of normalized angular frequencies and $||\cdot||$ is the euclidean norm operator. The following operation, shown in the second

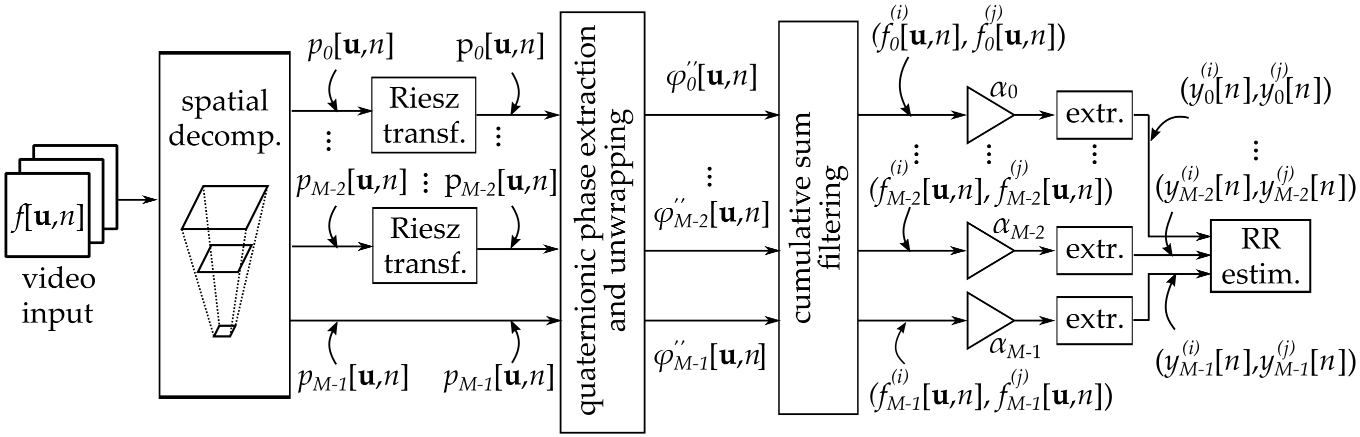

$f[\mathbf{u},n]$
video input
spatial decomp.
$p_0[\mathbf{u},n]$
Riesz transf.
$\mathrm{p}_0[\mathbf{u},n]$
$p_{M\text{-}2}[\mathbf{u},n]$
$\mathrm{p}_{M\text{-}2}[\mathbf{u},n]$
Riesz transf.
$p_{M\text{-}1}[\mathbf{u},n]$
$p_{M\text{-}1}[\mathbf{u},n]$
quaternionic phase extraction and unwrapping
$\varphi_0''[\mathbf{u},n]$
$\varphi_{M\text{-}2}''[\mathbf{u},n]$
$\varphi_{M\text{-}1}''[\mathbf{u},n]$
cumulative sum filtering
$(f_0^{(i)}[\mathbf{u},n], f_0^{(j)}[\mathbf{u},n])$
$(f_{M\text{-}2}^{(i)}[\mathbf{u},n], f_{M\text{-}2}^{(j)}[\mathbf{u},n])$
$(f_{M\text{-}1}^{(i)}[\mathbf{u},n], f_{M\text{-}1}^{(j)}[\mathbf{u},n])$
α_0
$\alpha_{M\text{-}2}$
$\alpha_{M\text{-}1}$
extr.
extr.
extr.
$(y_0^{(i)}[n], y_0^{(j)}[n])$
$(y_{M\text{-}2}^{(i)}[n], y_{M\text{-}2}^{(j)}[n])$
$(y_{M\text{-}1}^{(i)}[n], y_{M\text{-}1}^{(j)}[n])$
RR estim.

Figure 3.3: Amplitude-based (spatio-temporal) RR estimation algorithm.

bank of blocks in Figure 3.3, is hence performed:

$$\mathcal{R}\{p_m[\mathbf{u},n]\} = \begin{pmatrix} r_{1,m}[\mathbf{u},n] \\ r_{2,m}[\mathbf{u},n] \end{pmatrix} = \begin{pmatrix} h_1[\mathbf{u}] * p_m[\mathbf{u},n] \\ h_2[\mathbf{u}] * p_m[\mathbf{u},n] \end{pmatrix} \tag{3.8}$$

where $\mathcal{R}\{\cdot\}$ represents the Riesz transform operator, $h_i[\mathbf{u}] = \mathcal{F}^{-1}(H_i(\boldsymbol{\omega}))$, $i = 1, 2$, $*$ denotes the 2D convolution operator and $\mathcal{F}^{-1}(\cdot)$ is the inverse 2D Fourier Transform (FT) operator.

The following triple of elements

$$\mathrm{p}_m[\mathbf{u},n] = (p_m[\mathbf{u},n], r_{1,m}[\mathbf{u},n], r_{2,m}[\mathbf{u},n]) \tag{3.9}$$

is known as the monogenic signal of the m-th level. In particular, the combination of $\{\mathrm{p}_m[\mathbf{u},n]\}_{m=0}^{M-2}$ with the last level of the Laplacian pyramid $p_{M-1}[\mathbf{u},n]$ forms a Riesz pyramid.

It may be convenient to represent the monogenic signal in (3.9) as the quaternion [46]

$$\mathrm{q}_m[\mathbf{u},n] = p_m[\mathbf{u},n] + ir_{1,m}[\mathbf{u},n] + jr_{2,m}[\mathbf{u},n] + k \cdot 0 \tag{3.10}$$

where i, j and k are the imaginary units. Following the quaternionic algebra in [46], the norm and the logarithm of the quaternion in (3.10) can be respectively defined as

$$\|\mathrm{q}_m\| = \sqrt{p_m[\mathbf{u},n]^2 + r_{1,m}[\mathbf{u},n]^2 + r_{2,m}[\mathbf{u},n]^2} \tag{3.11}$$

$$\log(\mathrm{q}_m) = \log(\|\mathrm{q}_m\|) + \frac{ir_{1,m}[\mathbf{u},n] + jr_{2,m}[\mathbf{u},n]}{\|ir_{1,m}[\mathbf{u},n] + jr_{2,m}[\mathbf{u},n]\|} \arccos \frac{p_m[\mathbf{u},n]}{\|p_m[\mathbf{u},n]\|}. \tag{3.12}$$

The amplitude and the quaternionic phase of (3.10) can now be computed as

$$A_m[\mathbf{u},n] = \|\mathrm{q}_m[\mathbf{u},n]\| \tag{3.13}$$

$$i\varphi_m[\mathbf{u},n]\cos(\vartheta_m[\mathbf{u},n]) + j\varphi_m[\mathbf{u},n]\sin(\vartheta_m[\mathbf{u},n]) = \log(\mathrm{q}_m[\mathbf{u},n]/\|\mathrm{q}_m[\mathbf{u},n]\|) \tag{3.14}$$

where

$$\varphi_m[\mathbf{u}, n] = \arctan\left(\left(\sqrt{r_{1,m}[\mathbf{u}, n]^2 + r_{2,m}[\mathbf{u}, n]^2}\right) / p_m[\mathbf{u}, n]\right) \tag{3.15}$$

$$\vartheta_m[\mathbf{u}, n] = \arctan\left(r_{2,m}[\mathbf{u}, n] / r_{1,m}[\mathbf{u}, n]\right) \tag{3.16}$$

are the m-th phase and orientation, respectively. The main advantage of this signal representation is that the quaternionic phase in (3.14) is invariant to the signs of the phase and orientation in (3.15) and (3.16) [46].

Temporal filtering As a second step of the proposed phase amplification method, temporal filtering is again necessary to select a range of frequencies of interest. An IIR band-pass Butterworth filter of the second order with lower and higher cut-off frequencies $f_{\mathrm{L}}^{\mathrm{co}}$ and $f_{\mathrm{H}}^{\mathrm{co}}$ can be employed to filter the phases of each level of the Riesz pyramid. As discussed in [46], the quaternionic phases in (3.14) are first unwrapped and their cumulative sum is subsequently filtered. To this purpose, the quaternionic logarithm of the m-th ($m = 0, \ldots, M-1$) normalized Riesz pyramid coefficient is computed as

$$\begin{cases} \log(\bar{\mathrm{q}}_m[\mathbf{u}, n]) & \text{for } n = 0 \\ \log(\bar{\mathrm{q}}_m[\mathbf{u}, n]\bar{\mathrm{q}}_m^{-1}[\mathbf{u}, n-1]) & \text{for } n = 1, 2, \ldots \end{cases} \tag{3.17}$$

where $\bar{\mathrm{q}}_m[\mathbf{u}, n] = \frac{\mathrm{q}_m[\mathbf{u},n]}{\|\mathrm{q}_m[\mathbf{u},n]\|}$ is the normalized quaternion [46] and the following definitions of the inverse and conjugate quaternion are recalled considering (3.10)

$$\mathrm{q}_m^{-1} = \frac{\mathrm{q}_m^*[\mathbf{u}, n]}{\|\mathrm{q}_m\|^2} \tag{3.18}$$

$$\mathrm{q}_m^* = p_m[\mathbf{u}, n] - i r_{1,m}[\mathbf{u}, n] - j r_{2,m}[\mathbf{u}, n]. \tag{3.19}$$

Assuming that the orientations are approximately constant in time, the elements in (3.17) for $n = 1, 2, \ldots$ can be written as

$$i(\varphi_m^{'}[\mathbf{u}, n]) \cos(\vartheta_m[\mathbf{u}]) + j(\varphi_m^{'}[\mathbf{u}, n]) \sin(\vartheta_m[\mathbf{u}]) \tag{3.20}$$

where the term

$$\varphi_m^{'}[\mathbf{u},n] = \varphi_m[\mathbf{u},n] - \varphi_m[\mathbf{u},n-1] \tag{3.21}$$

is the phase difference. Defining now the unwrapped phase as

$$\varphi_m^{''}[\mathbf{u},n] = \varphi_m[\mathbf{u},0] + \sum_{k=1}^{n} \varphi_m^{'}[\mathbf{u},k] \quad \text{for } n = 1, 2, \dots \tag{3.22}$$

the following cumulative sum can be computed

$$i\varphi_m^{''}[\mathbf{u},n]\cos(\vartheta_m[\mathbf{u}]) + j\varphi_m^{''}[\mathbf{u},n]\sin(\vartheta_m[\mathbf{u}]). \tag{3.23}$$

Filtering the quantity in (3.23) in time, leads to two imaginary quaternionic components

$$\begin{aligned} f_m^{(i)}[\mathbf{u},n] &= \delta_m[\mathbf{u},n]\cos(\vartheta_m[\mathbf{u}]) \\ f_m^{(j)}[\mathbf{u},n] &= \delta_m[\mathbf{u},n]\sin(\vartheta_m[\mathbf{u}]) \end{aligned} \tag{3.24}$$

that define the spatial translation due to a framed motion. In Figure 3.3, the quaternionic phase extraction and unwrapping operations are associated with a single block that is followed by the cumulative sum filtering block.

Signal amplification Following the approach presented in [47], in order to enhance a motion of interest, the two filtered quaternionic components in (3.24) at each pyramid level $m \in \{0, \dots, M-1\}$ can be multiplied by the amplification factor α_m, $m \in \{0, \dots, M-1\}$, as shown in Figure 3.3, obtaining $\{\alpha_m f_m^{(i)}[\mathbf{u},n], \alpha_m f_m^{(j)}[\mathbf{u},n]\}_{m=0}^{M-1}$.

Signal extraction Motion signals can finally be extracted by spatial averaging the amplified and filtered quaternionic components (blocks labelled "extr." in Figure 3.3). Considering a frame size of $U_1 \times U_2$, the following signals are obtained

$$\begin{aligned} y_m^{(i)}[n] &= \tfrac{1}{U_1U_2} \sum_{u_1=1}^{U_1-1} \sum_{u_2=1}^{U_2-1} \alpha_m f_m^{(i)}[\mathbf{u},n] = \tfrac{1}{U_1U_2} \sum_{u_1=1}^{U_1-1} \sum_{u_2=1}^{U_2-1} \alpha_m \delta_m[\mathbf{u},n]\cos(\vartheta_m[\mathbf{u}]) \\ y_m^{(j)}[n] &= \tfrac{1}{U_1U_2} \sum_{u_1=1}^{U_1-1} \sum_{u_2=1}^{U_2-1} \alpha_m f_m^{(j)}[\mathbf{u},n] = \tfrac{1}{U_1U_2} \sum_{u_1=1}^{U_1-1} \sum_{u_2=1}^{U_2-1} \alpha_m \delta_m[\mathbf{u},n]\sin(\vartheta_m[\mathbf{u}]). \end{aligned} \tag{3.25}$$

3.3 Maximum Likelihood RR estimation

Once the motion signals are extracted at each pyramid level, the RR is estimated according to the ML principle.

The ML approach is indeed a reliable and consolidated tool that allows to estimate unknown parameters of interest. Since respiration is characterized by periodic (or quasi-periodic) movements of the chest and abdomen, i.e., expansion and relaxation, the ML criterion can be exploited to investigate the presence of a fundamental periodic component, corresponding to the RR, and estimate it. In [39, 48], the ML principle is also used to automatically select ROIs where the framed motion is mainly due to breathing. The amplitude- and phase- based method described in Sections 3.2.1 and 3.2.2 may, indeed, be applied to full-frame video sequences or to specific ROIs to reduce the computational complexity. The ROI selection operation, preliminarily presented in [39, 48], will be summarized in Section 3.3.1.

The RR estimation operation is embedded in the last blocks of Figures 3.1 and 3.3. As the motion signals are extracted at each pyramid level for both the presented approaches, a data aggregation method similar to the one proposed in [49] for multiple sensors can be employed.

For the sake of compactness, the motion signals extracted at each pyramid level can be grouped as follows

$$\mathbf{l}[n] = \begin{bmatrix} \bar{l}_0[n] \\ \bar{l}_1[n] \\ \vdots \\ \bar{l}_{M-1}[n] \end{bmatrix} \tag{3.26}$$

$$\mathbf{Y}[n] = \begin{bmatrix} y_0^{(i)}[n] & y_0^{(j)}[n] \\ y_1^{(i)}[n] & y_1^{(j)}[n] \\ \vdots & \\ y_{M-1}^{(i)}[n] & y_{M-1}^{(j)}[n] \end{bmatrix} \tag{3.27}$$

in the case of amplitude (3.26) and phase (3.27) components, respectively. Let us define $\mathbf{X}[n]$ as a generic observation model, that can be written in the form of (3.26) or (3.27) according to the considered method. The generic size of $\mathbf{X}[n]$ is $M \times C$, where M is the number of considered pyramid levels and the number of columns C is equal to 1 and 2 in the case of (3.26) and (3.27), respectively.

Given the nature of the respiration movements of interest, the observation model $\mathbf{X}[n]$ can be defined as

$$\mathbf{X}[n] = \mathbf{B} + \mathbf{A}\cos(2\pi f_0 T_s n + \mathbf{\Phi}) + \mathbf{W}[n] \tag{3.28}$$

where $\mathbf{B}$ are the continuous components, $\mathbf{A}$ and $\mathbf{\Phi}$ are the amplitudes and phases, respectively, and $\mathbf{W}[n]$ are sequences of independent and identically distributed (i.i.d.) zero-mean Gaussian noise samples, all of size $M \times C$. In (3.28), the amplitudes $\mathbf{A}$, the fundamental frequency f_0 and the phases $\mathbf{\Phi}$ are unknown parameters and may be collected as the array of parameters $\mathbf{\Theta} = [\mathbf{A}, f_0, \mathbf{\Phi}]$. Following the standard method presented in [23, p. 193-195] and extending it to the case of multi-dimensional signals, as in [49] and [50], the vector $\mathbf{\Theta}$ can be estimated on a window of N frames by minimizing the likelihood function

$$J(\mathbf{\Theta}) = \sum_{c=0}^{C-1} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \big[x[m,c,n] - a[m,c]\cos(2\pi f_0 T_s n + \phi[m,c])\big]^2 \tag{3.29}$$

where $x[m,c,n]$, $a[m,c]$ and $\phi[m,c]$ are the generic elements of the matrices $\mathbf{X}[n]$, $\mathbf{A}$ and $\mathbf{\Phi}$, respectively. As shown in [23, p. 193-195], if the real frequency f_0 is not close to 0 or $f_s/2$, an approximate expression of the estimator $\hat{f}_0$ of the fundamental frequency can be derived from (3.29) as

$$\hat{f}_0 = \underset{f_{\min} \le f \le f_{\max}}{\operatorname{argmax}} \sum_{c=0}^{C-1} \sum_{m=0}^{M-1} \left| \sum_{n=0}^{N-1} x[m,c,n] e^{-j2\pi f T_s n} \right|^2 \tag{3.30}$$

where the maximization is performed over the limited frequency interval $[f_{\min}, f_{\max}]$, with $f_{\min}$ and $f_{\max}$ being the minimum and the maximum feasible frequencies, respectively, that must be heuristically set.

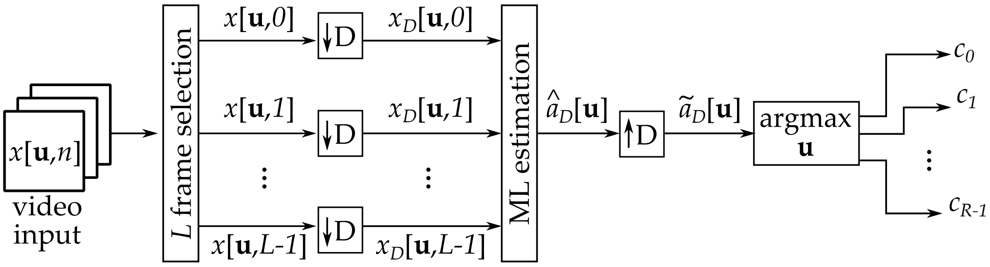


Figure 3.4: ROI selection algorithm.

The amplitudes can be similarly estimated as

$$\hat{a}[m,c] = \frac{2}{N} \sum_{c=0}^{C-1} \sum_{m=0}^{M-1} \left| \sum_{n=0}^{N-1} x[m,c,n] e^{-j2\pi \hat{f}_0 n T_s} \right| \tag{3.31}$$

and the presence of a significant periodic component is declared, according to [50], only if the following condition is verified

$$\frac{N}{MC} \sum_{c=0}^{C-1} \sum_{m=0}^{M-1} \hat{a}^2[m,c] > \eta \tag{3.32}$$

where η is a properly set threshold.

3.3.1 Region of interest selection

To reduce the computational complexity of the proposed algorithms, a method to extract ROIs can be exploited to obtain and process video sequences with reduced frame size. In this section, the automatic ROI selection algorithm preliminary presented in [39, 48] is summarized. The proposed method, whose illustrative overview is shown in Figure 3.4, is based on the above described ML approach, now applied to the considered video sequence.

Given the generic video sequence $x[\mathbf{u}, n]$, the first step for automatically extracting R ROIs consists in selecting L frames where variations are only due to respiration movements. This processing step is associated with the first block of the diagram in Figure 3.4. The L frames $\{x[\mathbf{u}, n]\}_{n=0}^{L-1}$ are first downsampled in space by an integer value D to reduce the computational complexity, obtaining a new block of frames $\{x_D[\mathbf{u}, n]\}_{n=0}^{L-1}$ with a smaller dimension $\lceil U_1/D \rceil \times \lceil U_2/D \rceil$, where $\lceil \cdot \rceil$ represents the ceiling operator. This operation is highlighted by the second bank of blocks of the diagram in Figure 3.4. The ML approach described in Section 3.3, associated with the third block of Figure 3.4, is applied to the downsampled sequences $\{x_D[\mathbf{u}, n]\}_{n=0}^{L-1}$ to estimate the fundamental frequency $\hat{f}_0$ and the amplitudes $\hat{a}_D[\mathbf{u}]$, according to (3.30) and (3.31), respectively, where $x[m, c, n]$ and $M \times C$ are replaced now by $x_D[\mathbf{u}, n]$. i.e., the intensity of the pixel at position $\mathbf{u}$, and $\lceil U_1/D \rceil \times \lceil U_2/D \rceil$, i.e., the size of the frames.

To compute the centres of the selected R ROIs, the matrix of the amplitudes $\hat{a}_D[\mathbf{u}]$, estimated for the reduced frames, is interpolated at the original frame size $U_1 \times U_2$ to estimate the amplitudes $\tilde{a}_D[\mathbf{u}]$ in the original block of frames. The centres $\{c_r\}_{r=0}^{R-1}$ are finally found by selecting the coordinates of the pixels that correspond to the maximum values of $\tilde{a}_D[\mathbf{u}]$. The interpolation and the selection of the ROIs centres are the operations embedded in the fourth and fifth (last) blocks of the diagram in Figure 3.4. This procedure allows to extract R ROIs with a fixed size $W \times W$ and may be repeated over time to deal with changes in the position of the framed subject.

3.3.2 Large motion detection

To discard ROIs where the motion is affected by large movements unrelated with breathing, a further control procedure may be needed, as discussed in [39, 48]. To this purpose, the intensity of the pixel at position $\mathbf{u}$ at the n-th frame of the r-th ROI can be defined as $x_r[\mathbf{u}, n]$ and the pixel-wise difference of consecutive frames can be computed as

$$i[\mathbf{u}, n] = x_r[\mathbf{u}, n] - x_r[\mathbf{u}, n-1]. \tag{3.33}$$

To reduce the computational complexity, the filtered signal in (3.33) could also be binarized according to the following binarization rule

$$i_r[\mathbf{u},n] = \begin{cases} 0 & \text{if } |x_r[\mathbf{u},n] - x_r[\mathbf{u},n-1]| < \gamma_{\text{bin}} \\ 1 & \text{else} \end{cases} \qquad r = 1,2,\ldots,R \tag{3.34}$$

where γ_{bin} is a properly chosen binarization threshold. The average motion signal on the r-th region can now be computed as

$$\bar{i}_r[n] = \frac{1}{W^2} \sum_{u_1=1}^{W-1} \sum_{u_2=1}^{W-1} i_r[\mathbf{u},n]. \tag{3.35}$$

A good decision strategy is such that the r-th ROI is discarded if $\bar{i}_r[n]$ in (3.35) is above a heuristically set threshold, as expressed by the following decision rule:

$$\kappa_r = \begin{cases} 0 & \text{if } \bar{i}_r[n] > \gamma_{\text{th}} \\ 1 & \text{else} \end{cases} \qquad r = 1,2,\ldots,R \tag{3.36}$$

where the binary-valued decision κ_r defines the presence (κ_r=1) or absence (κ_r=0) of large motion inside the r-th ROI and γ_{th} is the selected decision threshold.

Finally, the RR is estimated by maximizing the following likelihood function

$$J(\boldsymbol{\Theta}) = \sum_{r=1}^{R} \kappa_r J_r(\boldsymbol{\Theta}). \tag{3.37}$$

where $J_r(\boldsymbol{\Theta})$ is defined according to (3.29) and refers to the r-th ROI.

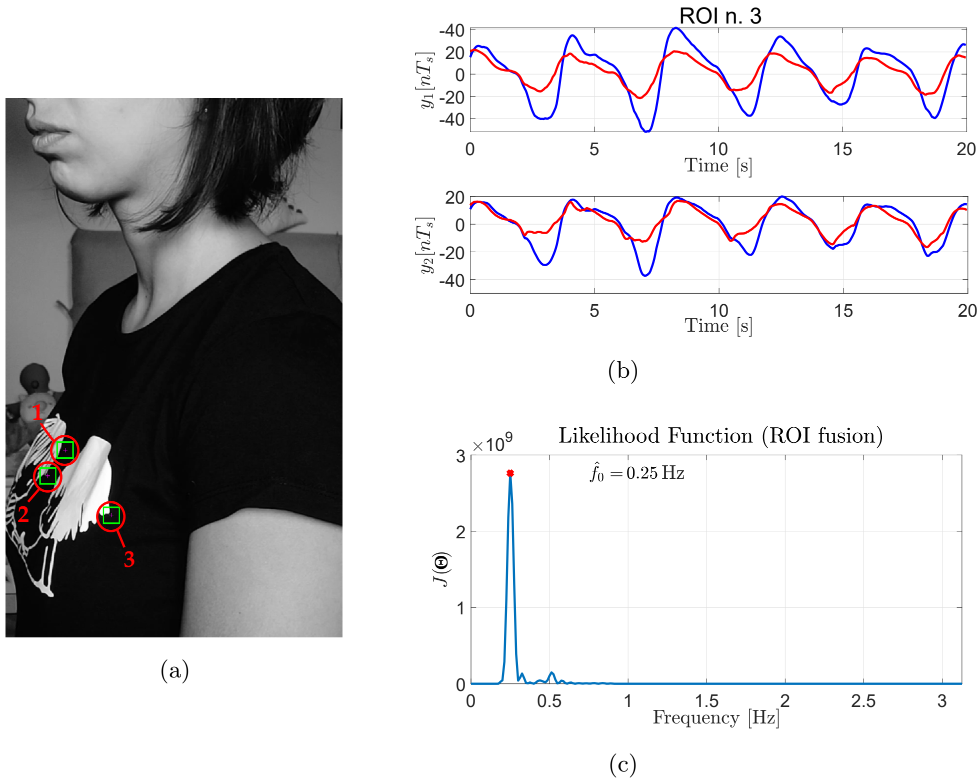
3.4 Applications and Results

The performance of the estimation algorithms detailed in Sections 3.2 and 3.3 is now discussed on the basis of experimental results directly obtained by applying the proposed methods on three sets of videos specifically recorded. In particular, the first set includes 2 videos of a newborn [51], whereas the second and third sets include 4 and 10 videos, respectively, of adults sitting still. All videos are recorded indoor by placing a camera laterally or in front of a steady subject normally breathing and not affected by respiratory disorders. Possible random movements of the subject unrelated to the respiration do not significantly affect the performance of the estimation algorithms, as the large motion detection method described in Section 3.3.2 is employed to discard such movements.

Motion signals are initially extracted and compared with reference data. In the case of the newborn, a pneumogram is used as a gold standard device to acquire the reference respiratory waveform by placing an elastic belt around the chest of the subject. In the case of adults, two wearable sensors, i.e., Shimmer3 by Shimmer Sensing$^{\text{TM}}$ and Equivital EQ02 LifeMonitor by Equivital$^{\text{TM}}$ [52] are used to record, respectively, the reference accelerometric signal and the respiratory waveform. A comparison between the two proposed methods is also presented. Results in terms of the RMSE normalized to the Root Mean Squared (RMS) value of the reference data are finally presented.

In Figure 3.5(a), an illustrative image of a framed subject is shown, highlighting three ROIs as squared regions. The centers of the ROIs are computed according to the procedure detailed in Section 3.3.1 for $R = 3$. In Figure 3.5(b), the corresponding motion information extracted by the phase-based motion magnification estimation method is shown. In particular, the signals $y_m^{(\iota)}[n]$, $\iota \in \{i, j\}$, obtained by applying (3.25), are plotted over a 20 s time window for the third ROI, i.e., $r = 3$, and for two pyramid levels, i.e., $m = 1, 2$. Finally, in Figure 3.5(c), the corresponding likelihood function obtained by applying (3.37) is shown as a function of the frequency. The estimated frequency $\hat{f} = 0.25$ Hz is the one corresponding to the maximum peak of the function.

As illustrative examples, motion signals extracted by the amplitude- and phase-based motion magnification estimation methods are shown in Figures 3.6 and 3.7, respectively, along with the corresponding reference signals. In particular, in Figure 3.6, the motion signal extracted from a video of a newborn by applying (3.6) is plotted over a 20 s time window along with the reference signal, i.e., the pneumogram, for the second level ($m = 1$) of the pro-



(a)

(b)

(c)

Figure 3.5: Example of: (a) image of a framed subject where 3 ROIs are highlighted, (b) extracted motion information for $r = 3$ and $m = 1, 2$ and (c) likelihood function where the estimated RR is highlighted by the argument of the peak at 0.25 Hz.

cessed pyramid. Considering that one period of the pneumogram corresponds to a complete respiratory cycle, that involves two main movements, i.e., inhalation and exhalation, a good correspondence between the two signals can be observed. On the other hand, the average phase variations extracted by two videos, of a newborn and an adult, are plotted over two 20 s time windows and compared with the corresponding pneumogram and accelerometric signal in Figures 3.7(a) and 3.7(b), respectively. In each case, the two pairs of

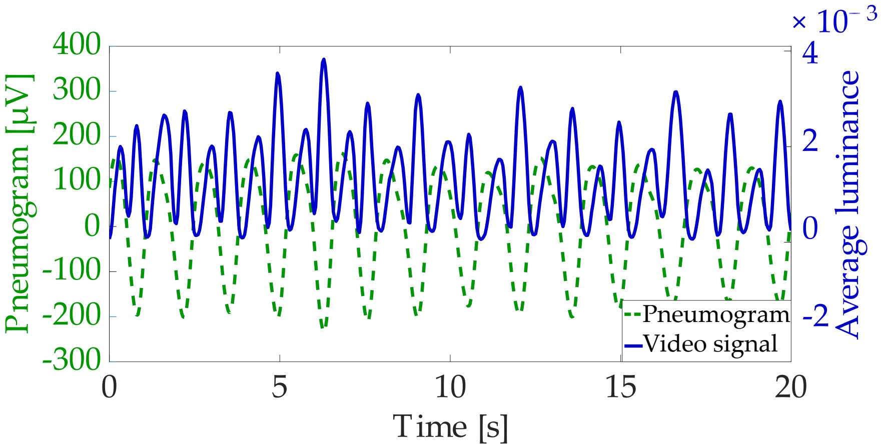
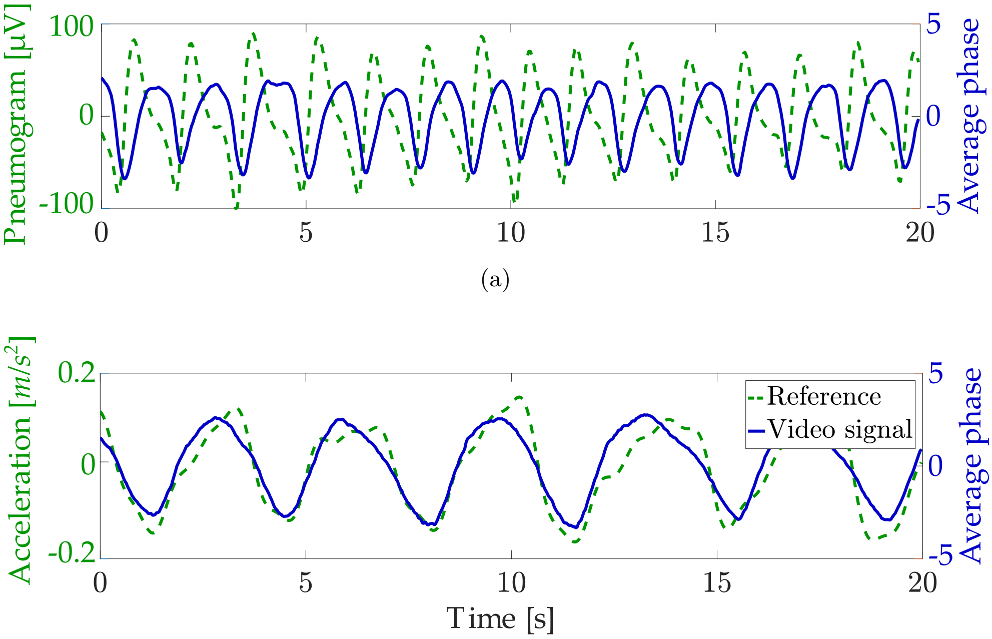


Figure 3.6: Comparison between the motion signal extracted from a video of a newborn by the amplitude-based motion magnification estimation method and the reference signal, i.e., the pneumogram.

signals exhibit a comparable periodicity, whereas the differences between the two reference signals, in particular the RR, depend on the employed sensors and on the age of the subject.

As further investigation, the ML estimation method presented in Section 3.3 is performed on interlaced windows of N frames, each corresponding to NT_s s. In the following results, interlaced windows are considered to track the RR over time with proper resolution and the overlap of consecutive windows is defined by an interlacing factor $\rho \in [0, 1)$. An example of windows of length NT_s s interlaced by a factor $\rho = 0.75$ is shown in Figure 3.8.

In Figure 3.9, the frequencies estimated by the phase-based method on interlaced windows for a video framing an adult sitting still are compared with the reference frequencies estimated by the Equivital EQ02 LifeMonitor. The duration of the considered video is 56 s and the RR estimation is performed on 20 s windows interlaced by 90% (i.e., $\rho = 0.9$). This corresponds to 28 processed windows. The first 9 windows should not be considered in the analysis because processed data are incomplete due to the chosen window overlap,



(a)

(b)

Figure 3.7: Comparison between two motion signals extracted by the phase-based motion magnification estimation method and the reference signals: (a) pneumogram of a newborn, (b) accelerometric signal of an adult.

as shown in Figure 3.8 for $\rho = 0.75$, which exhibits 3 incomplete initial windows. It can be noticed that the RR is estimated with good approximation in all windows, confirming the robustness of the system. Tolerance boundaries highlighted in Figure 3.9 are computed according to the medical practice of considering acceptable a ±15% variation from to the reference frequency.

A comparison of the presented amplitude- and phase-based methods is now proposed in Figure 3.10. In particular, the signal extracted by the amplitude-based motion magnification estimation method from a video of a newborn is shown in Figure 3.10(a) along with the corresponding signals $y_0^{(\iota)}[n]$, $\iota \in \{i, j\}$, locally extracted from a selected ROI of the considered video by the phase-

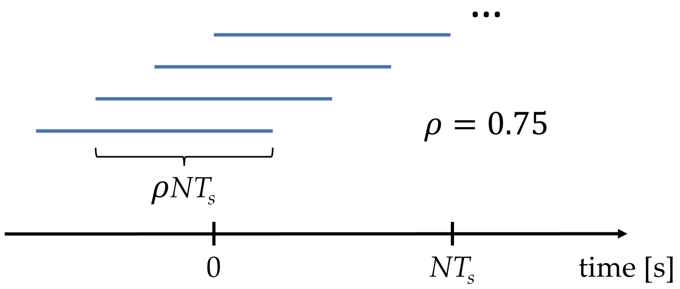


Figure 3.8: Windows of length NT_s s interlaced by a factor $\rho = 0.75$.

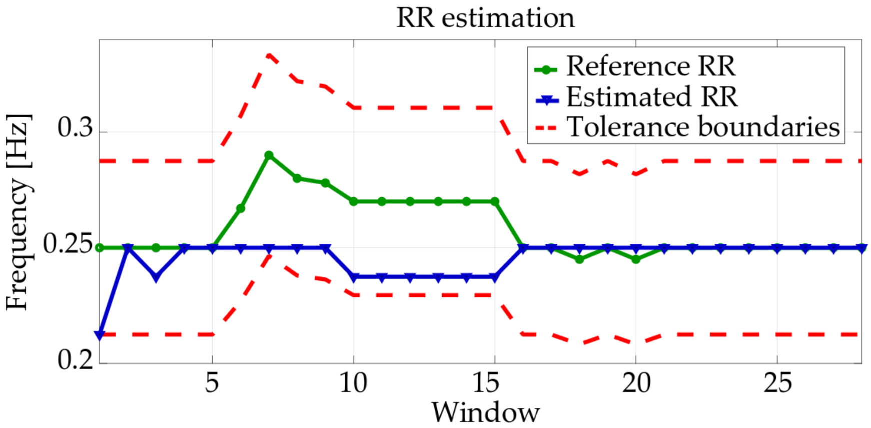
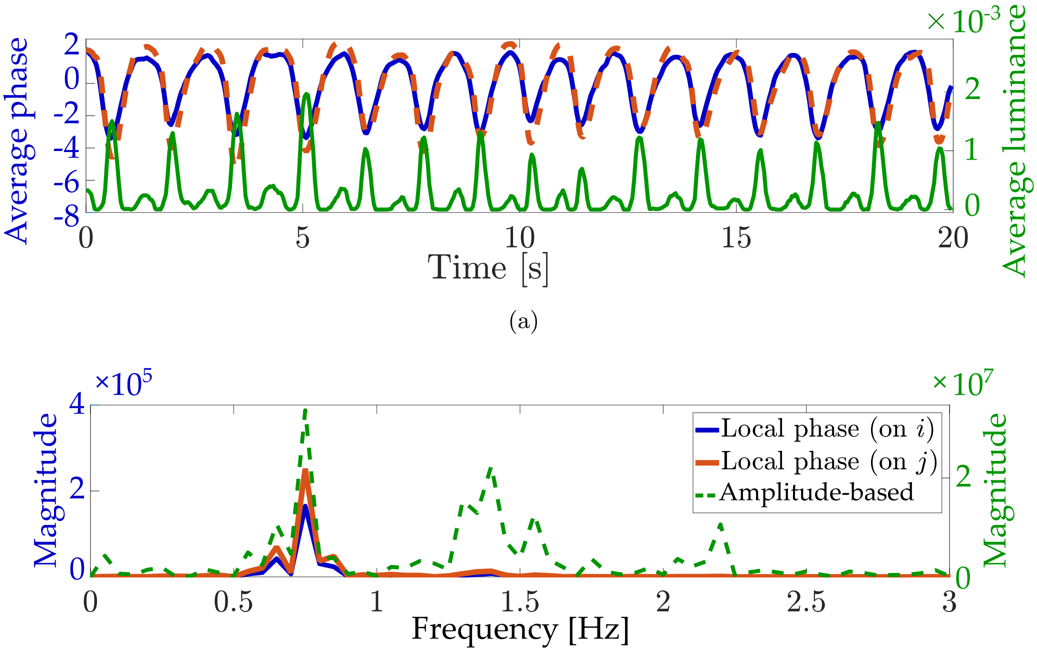


Figure 3.9: Comparison between estimated and reference RR for the phase-based estimation method.

based method. The duration of the considered video signal is 20 s. The signal extracted by the amplitude-based method is always positive as the quantity obtained by applying (3.6) defines the average luminance for each processed frame. For this reason, inhalation and exhalation acts, which are characterized by movements in opposite directions, may not be clearly distinguishable, especially under critical conditions, e.g., poor camera positioning or patient



(a)

(b)

Figure 3.10: Comparison between the presented methods: (a) extracted motion signals and (b) magnitude spectra.

type. The phase-based method allows to overcome this limit, as the extracted signals $y_0^{(\iota)}[n]$, $\iota \in \{i, j\}$ exhibit negative values, too. The different characteristics of the two types of signals are also visible in Figure 3.10(b), where their magnitude frequency spectra are plotted. As the phase-based magnification is performed on a selected ROI, the extracted phases are indicated as "local phases" in the legend of Figure 3.10(b). A peak around 0.75 Hz can be observed for all the considered cases, corresponding to the correctly estimated RR of 45 breath/min. Nevertheless, the shape of the signal extracted by the amplitude-based method causes other peaks, related to higher order harmonics, to appear around 1.4 Hz and 2.2 Hz. Under critical conditions, these secondary peaks may be higher than the fundamental one impairing the RR estimation: for example, a frequency twice the correct one could be estimated. On the other hand, as the signals extracted by the phase-based method are quasi-sinusoidal, due to the direct application of (3.25), peaks related to higher order harmonics are negligible. This leads to a more reliable RR estimation.

3.4.1 Performance analysis

To evaluate the performance of the presented methods, videos framing different subjects in different scenarios were tested. The main characteristics of the considered videos are summarized in Table 3.1, where the parameter setting for the video processing analysis is also reported. The duration of the videos vary between around 1 min 35 s and 5 min. The camera resolution and the sampling frequency vary according to the employed recording device. As a reminder, the parameters M, W and R respectively indicate the number of pyramid levels, the fixed size of the ROIs and the number of ROIs. The cut-off frequencies of the employed Butterworth filter, used to extract the frequency band of interest, are denoted as $f_{\mathrm{L}}^{\mathrm{co}}$ and $f_{\mathrm{H}}^{\mathrm{co}}$, α is the amplification factor, NT_s is the duration of the processed time window and the interlacing factor denotes the overlap between consecutive estimation windows. The device used as reference is also indicated.

The accuracy of the presented methods is now analysed in terms of normalized RMSE for 6 tested videos. The results, expressed in dB, are shown in Figure 3.11, where the type of framed subject is reported. Considering N_w temporal windows where the RR estimation is performed, the RMSE for each video is defined as

$$\xi = \sqrt{\frac{\sum_{n=1}^{N_w} \left|\hat{f}_0[n] - f_0[n]\right|^2}{\sum_{n=1}^{N_w} |f_0[n]|^2}} \tag{3.38}$$

where $\hat{f}_0[n]$ and $f_0[n]$ are the estimated and reference frequency for the n-th window, respectively. The reference frequencies are obtained by means

Video	**No.**	**Camera**	f_s	M	W	R	$[f_{\mathrm{L}}^{\mathrm{co}}, f_{\mathrm{H}}^{\mathrm{co}}]$	α	NT_s	ρ	**Reference**
set	**samples**	**resolution**	[Hz]		[pixel]		[Hz]		[s]		**device**
Newborns	2	360 × 288	25	3	21	4	[0.3 1.1]	25	20	0.5	Pneumograph
Adults	4	800 × 600	30	4	41	3	[0.19 0.9]	20	20	0.5	Accelerometer
Adults	10	1920 × 1080	30	3	16	3	[0.19 0.9]	20	20	0.5	Equivital EQ02 LifeMonitor

Table 3.1: Characteristics of the considered videos and parameter setting.

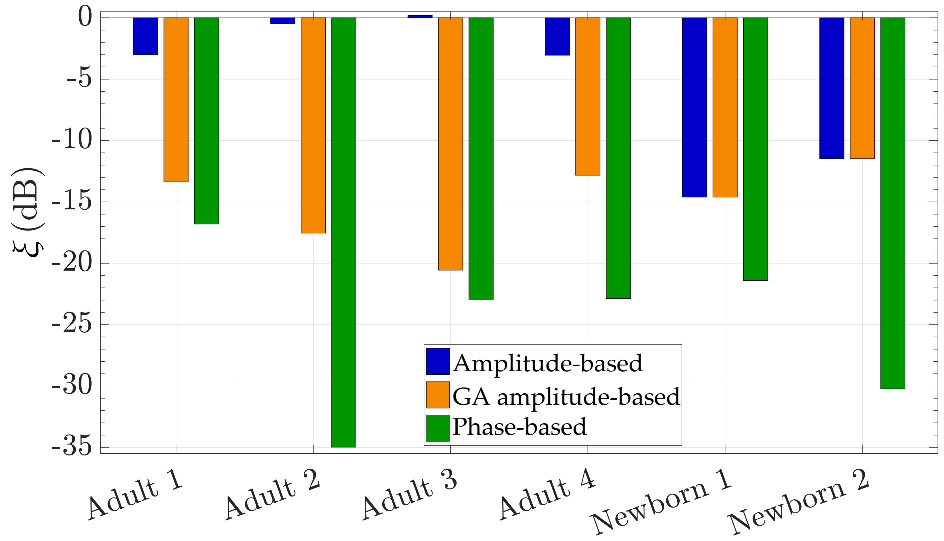


Figure 3.11: Performance of the assessed methods in terms of normalized RMSE for 6 considered videos.

of an accelerometer, for adults, and a pneumogram, for newborns. As the RR estimation obtained by the amplitude-based approach is prone to errors caused by higher order harmonics, as previously discussed, an idealized Genie-Aided (GA) version of this method is also considered as a benchmark. The GA method automatically corrects estimated double frequencies. Despite this adjustment, the phase-based method exhibits better performance for all the considered videos. Estimates are indeed more reliable due to the characteristics of the motion signals in (3.25) of inherently distinguishing motions in the different directions associated with inhalation and expiration.

In order to further analyse the performance of the more efficient phase-based method, 10 more videos, all framing adults sitting still, were tested and the normalized RMSE ξ was computed according to (3.38). Various subjects, scenarios and camera angles were considered and the Equivital EQ02 Life

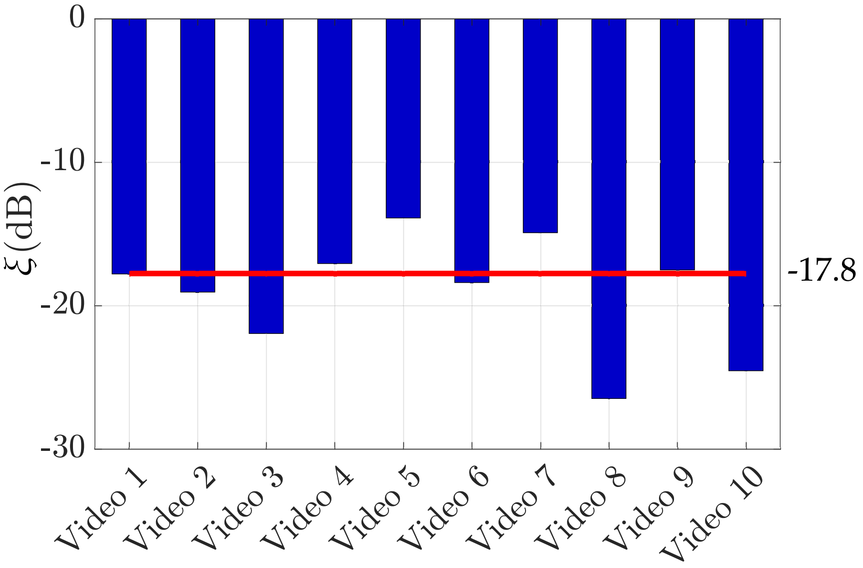


Figure 3.12: Performance of the phase-based method in terms of the normalized RMSE ξ for 10 considered videos.

Monitor was used as the reference device. The results, expressed in dB, are presented in Figure 3.12 and show a good agreement with the RMSE values in Figure 3.11, thus confirming the robustness of the considered method. The average error over all the videos is also highlighted as a straight line at -17.8 dB and it can be observed that the RMSE obtained for 6 videos is below or equal to this value.

3.5 Conclusions

In this chapter, two video-based techniques for respiration monitoring are presented along with a brief review of relevant state-of-the art methods. As the act of respiration induces periodic movements of the chest and abdomen, video cameras can be used to capture motion signals to be properly processed in order to estimate the RR. The proposed methods are based on amplitude and phase motion magnification to amplify subtle respiratory movements and combine spatial and temporal processing techniques to extract reliable motion information. Suitable ROIs, where the motion is mainly due to respiration, may be selected to enhance the estimation. Once the procedure to extract the motion signals is defined, the estimation of the RR is performed by means of the ML approach, which allows to aggregate data from different ROIs and pyramid levels. The performance of the two techniques is evaluated on a number of real videos framing newborns and adults and specifically recorded. In particular, a comparison with reference data is presented showing good agreement between the estimated signals and the reference ones. Nevertheless, the characteristics of the motion signal obtained by applying the amplitude-based approach may lead to wrong frequency estimates, doubled with respect to the correct one. This issue is overcome by the phase-based method that leads to more reliable estimates due to the regular shape of the extracted motion signals which may resemble quasi-sinusoidal ones. The performance of the two methods is also compared in terms of normalized RMSE showing the better accuracy of the phase-based approach, that achieves lower error for all the tested videos.

Chapter 4

Human Monitoring for Automotive Applications

In the context of automotive safety, driver monitoring is a key task in preventing hazardous events, such as road accidents. To this purpose, driver monitoring systems aim to provide prompt assistance in case of anomalies, such as alterations of the driver's psycho-physiological status. The driving performance may, indeed, be negatively affected by high levels of stress and workload, that might be caused by different factors. In particular, scientific evidence supports the hypothesis of a strong correlation between high stress levels and the traffic, road and weather conditions, as demonstrated in [53] and [54].

In this chapter, a novel in-vehicle system to assess the psycho-physiological status of a driver is proposed. The system is based on heterogeneous techniques to collect and analyse data obtained by various components and aim to estimate the arousal level of the driver. In particular, a wearable sensor and a thermal camera are employed to extract physiological parameters of interest, such as the Heart Rate (HR) and the temperature of the subject. Temperature variations detected on the subject's face, and other bodily functions including the HR, are regulated by the activity of the Autonomic Nervous System (ANS) and represent, indeed, an important indicator of arousal levels.

Thermal imaging techniques and video processing algorithms are jointly employed in this work to record and extract temperature-related information in a non-invasive and contactless way.

The architecture and the data acquisition protocol of the proposed monitoring system are described in this chapter. Results are shown at a preliminary stage, as data fusion techniques need to be developed to aggregate heterogeneous information in order to extract a correlation between the analysed parameters to provide useful indications about the driver's status.

The chapter is organized as follows. In Section 4.1, an overview on existing monitoring systems for automotive applications is introduced. In Section 4.2, the architecture of the proposed systems and a brief description of the each component is provided. The procedures to collect and process data obtained through the considered sensors are detailed in Sections 4.3 and 4.4, respectively. Preliminary results are presented in Section 4.5 and conclusions are finally drawn in Section 4.6.

4.1 Driver Monitoring Systems

The definition and analysis of in-vehicle monitoring systems which are able to simultaneously collect physiological indices and useful data to determine the joint driver-vehicle status is a topic of great interest. Recent technological progress has introduced solutions to provide assistance to the traditional manual driving. In particular, on-board Advanced Driver-Assistance Systems (ADASs) [55], commonly employed in vehicles, e.g., cars, trucks, etc., are aimed at improving safety and security during driving experiences for both drivers and passengers [56].

Popular ADAS mechanisms equipping modern vehicles are Anti-lock Braking System (ABS) [57], Adaptive Cruise Control (ACC) [58] and other technologies that aim to counteract road accidents by compensating for dangerous driving behaviours. Human errors are possibly related to alterations of the driver's psycho-physiological status and may be caused by drowsiness, sudden sickness and distractions. Safety measures to avoid collisions and fatalities include the activation of alarms or, when necessary, the take-over of the vehicle itself.

Looking at the interaction level of all these paradigms, the majority of existing ADASs do not take into account aspects related to the psycho-physiological status of the driver. However, ADASs provided with information about the driver's psycho-physiological condition could take more contextualized actions, implementing complex decisions that are compatible with the driver's possible reactions, but are very challenging to obtain. A system able to extract and deal with physiological data requires, indeed, a continuous monitoring and understanding of the the driver's status along with an effective communication of the ADAS decisions to the driver. Direct actions on the vehicle may be also necessary in some critical situations. Another aspect that should be considered in the field of in-vehicle monitoring systems is the degree of intrusiveness of the sensors employed for collecting the data of interest, which in fact should not interfere, or minimize as much as possible their interference, with the driving activity, and therefore should be selected accordingly.

4.1.1 Physiological parameters monitoring

One of the uprising unobtrusive ways to monitor individual physiological parameters while driving is through the collection and analysis of biological signals obtained by non-invasive sensors, such as wearables and cameras. Informative indices about the driver's stress condition include the HR and its variability, ElectroCardioGram (ECG) signals, Respiratory Rate (RR) and skin temperature variations. All these functions are governed by the ANS, whose activity is highly affected by stressful events. For this reason, physiological sensors represent an important tool for the accurate assessment of stress levels.

A survey about physical and physiological measures to detect stress can be found in [59]. For instance, changes in the heart electrical activity may be investigated by acquiring and processing ECG signals, whose amplitude fluctuations may be related to high stress levels. In particular, variations in the heartbeat time intervals are measured by the Heart Rate Variability (HRV) index, that can be extracted from ECG signals and is documented to be a reliable measure for stress [60, 61]. Hence, ECG analysis directly performed inside the vehicle cabin may be useful to assess parameters, such as the HRV, related to both mental and physical stress, and workload [62–64]. Additional physiological indicators, such as the RR and skin temperature may also be investigated.

Camera sensors

An important category of non-invasive monitoring technologies is represented by digital cameras. Whereas Red, Green and Blue (RGB) sensors may be sensitive to illumination changes, infrared thermal cameras [65] can detect objects in different environmental conditions, e.g., rain, darkness, in the presence of fog, etc., are unaffected by sun glare, improving situational awareness [66], and are more robust against reflections, shadows and car headlights [67].

The use of thermal imaging for stress analysis has been investigated both in laboratory settings [68] and driving scenarios [69]. Both studies consider the nasal tip to be one of the main facial region to provide relevant stress-related information. In fact, nasal skin temperature is regulated by the ANS and is demonstrated to drop when arousal levels increase [70].

4.2 In-Vehicle IoT-oriented Monitoring Architecture

The proposed system is composed by different sensing devices, which are properly positioned inside the vehicle cabin as shown in Figure 4.1, where the experimental system set-up is illustrated. In particular, the driver's physiological parameters of interest are obtained through the employed wearable sensor, whereas the thermal camera records the temperature variations on the subject's facial skin. Vehicular data are extracted by embedded inertial sensors,

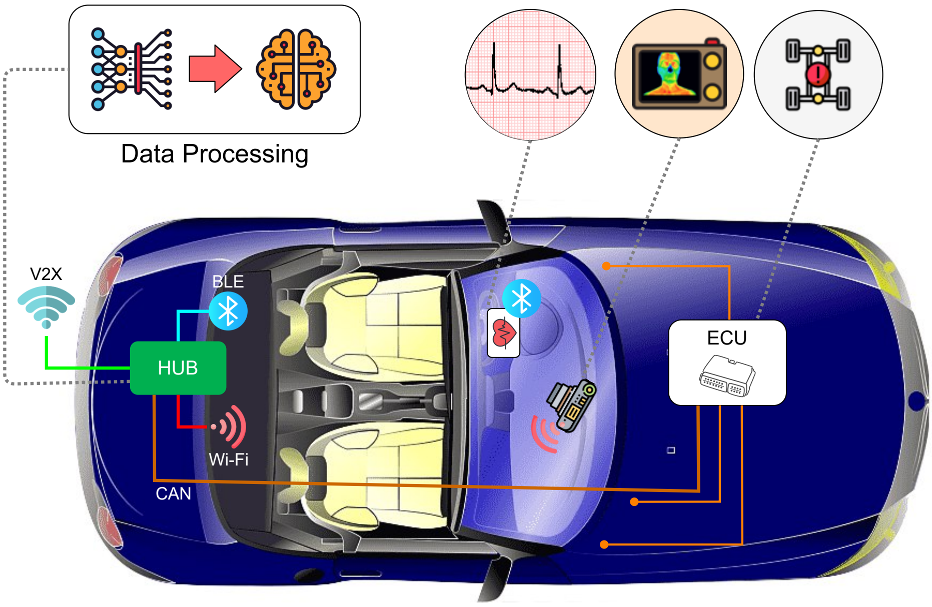
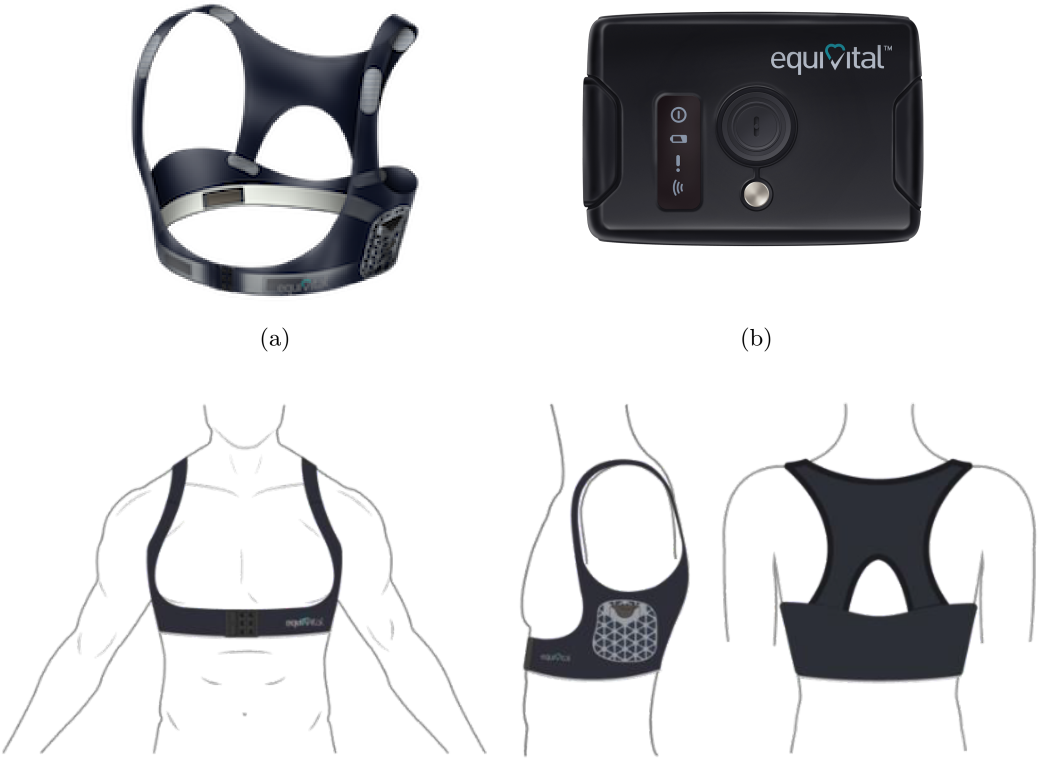


Figure 4.1: Experimental set-up of the proposed monitoring system.

i.e., the Electronic Control Unit (ECU). A brief description of the various system components and their positioning is provided hereafter.

4.2.1 Wearable sensor

In order to collect the driver's physiological data, an Equivital EQ02 Life monitor [52] sensor is adopted. This wearable sensing device is composed by two elements: a chest belt containing fabric electrodes, that need to be placed in good contact with the driver's skin, and a Sensor Electronics Module (SEM). In particular, the SEM records various parameters, including the ECG signal, HR, RR and skin temperature. It also provides indications about the position and motion of the subject, by collecting data from a 3-axis accelerometer. Pictures of the sensor belt and SEM are shown in Figures 4.2(a) and 4.2(b), respectively, whereas their correct positioning is illustrated in Figure 4.2(c).



(a)

(b)

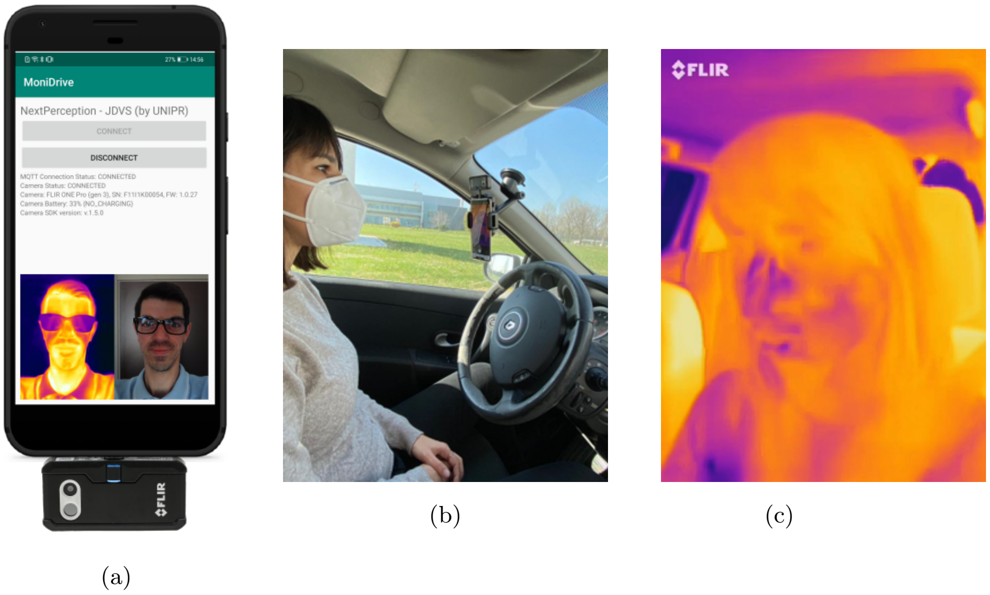
(c)

Figure 4.2: Equivital EQ02 Life monitor sensor: (a) belt, (b) SEM, (c) positioning.

On a practical side, the belt securely holds the SEM on the driver's body through a specific pocket on the left side of the belt itself.

4.2.2 Thermal camera

The second component integrated in the architecture of the proposed monitoring system is a FLIR One Pro LT [71] thermal camera, which allows to collect thermal information from the body (i.e., face) of the driver and the surrounding environment. In detail, this video-capturing device is composed



MQTT Connection Status: CONNECTED
Camera Status: CONNECTED
Camera: FLIR ONE Pro (gen 3), SN: F11I1K00054, FW: 1.0.27
Camera Battery: 33% (NO_CHARGING)
Camera SDK version: v.1.5.0

(a) (b) (c)

Figure 4.3: FLIR One Pro LT thermal camera: (a) device connection, (b) positioning, (c) recorded infrared frame.

by an RGB and an infrared sensor and is intended to work as an external "dongle" to be plugged into a smartphone, that may run either Android or iOS operating systems. The smartphone and the connected thermal camera need to be accurately installed inside the vehicle in order to find the optimal positioning for the data acquisition. A fair trade-off between the quality of the recordings and the degree of obtrusiveness for the driver's perspective toward the windscreen need to be taken into account. Further video processing and analysis tasks require, indeed, a frontal viewpoint. The connection of the device to the smartphone and their positioning inside the cabin are shown in Figures 4.3(a) and 4.3(b), whereas an example of a captured infrared video frame is illustrated in 4.3(c). In particular, the FLIR One Pro LT thermal camera may work according to three different modalities that allow to acquire

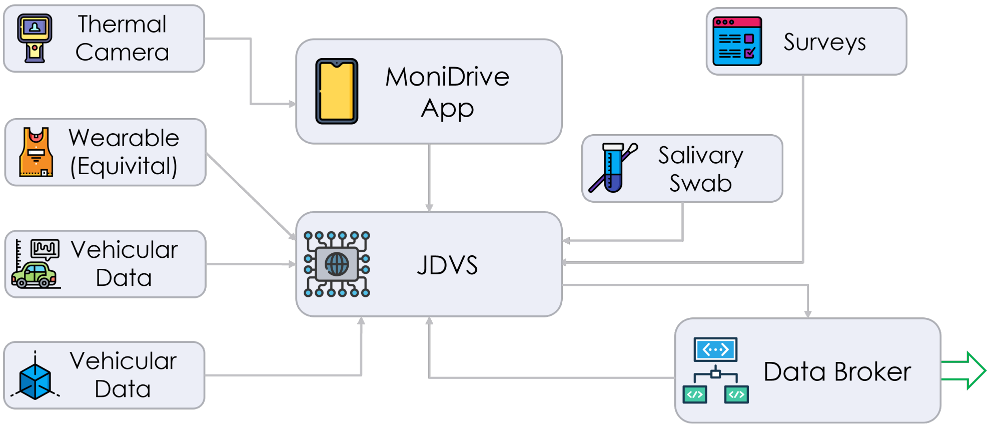


Figure 4.4: Data acquisition protocol.

RGB, thermal or blended data.

4.3 Data Acquisition

The various components of the system transmit the collected data to an on-board multi-interface gateway through off-the-shelf communication protocols, e.g., Wi-Fi, Bluetooth and serial bus, where data fusion and processing algorithms are also implemented. A schematic illustration of the data acquisition protocol is provided in Figure 4.4, where the connections between the components are highlighted. In particular, a mobile application (MoniDrive in Figure 4.4) running on the smartphone has been developed in order to acquire data from the FLIR One Pro LT thermal camera connected to the smartphone itself. The MoniDrive application collects the RGB and thermal images, plus additional frame and camera information, i.e., temperature scale, framed hottest and coldest points, etc., that are subsequently sent via a Message Queue Telemetry Transport (MQTT) transmission protocol to an external hub. This hub is called Joint Driver Vehicle Status (JDVS) in Figure 4.4 and also receives data from the Equivital EQ02 Life monitor sensing device (i.e., HR, HRV, RR) via a Transmission Control Protocol (TCP) socket and from the inertial sensors embedded in the vehicle. Additional input information is also collected. Salivary cortisol analysis and offline surveys are administrated to drivers to collect reference indications about their psycho-physiological status, in terms of stress and anxiety levels.

The JDVS module has a central role in the system architecture as it performs different functions:

- acting as a Wi-Fi Access Point and a MQTT broker,
- processing the information received from the mobile app, through dedicated image and video processing algorithms,
- processing the data received from the sensored belt,
- operating data fusion in order to compute and transmit an estimated value of arousal to the data broker.

The data broker is finally able to communicate with the external environment and may possibly forward the arousal information to other external modules.

4.3.1 Driving protocol

In order to collect data on the driver's psycho-physiological status, driving tests are performed according to a well-defined operating protocol. Both smooth and fast driving are included in the analysis to evaluate the driver's response to different external stimuli, associated to different amount of perceived stress. To this end, the driving sessions take place both in a controlled environment, i.e., a driving simulator, and in realistic scenarios, i.e., on urban and highway roads in situations of smooth and heavy traffic. The correlation between stress and road traffic and between stress and road type has been, indeed, demonstrated by scientific evidence. E.g., in [72] an experimental route that includes city and highway driving was tested to assess the driver stress level in different road conditions. The driving tests are administered, upon signing the informed

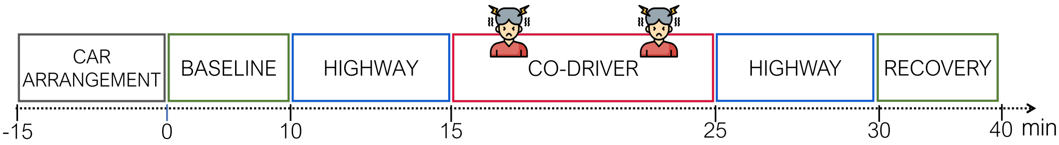


Figure 4.5: Driving protocol.

consent, to a total number of 40 healthy subjects: 20 women and 20 men aged 20 to 50. The participants must hold a driving license from at least 3 years and own a car, that is specifically arranged and employed for the test.

The driving protocol is divided into six time intervals, as shown in Figure 4.5. In particular, a driving test can start after an initial arrangement phase where the vehicle is equipped with the various components of the proposed monitoring system. In this phase, the Equivital EQ02 Life monitor sensor is worn by the driver and the FLIR One Pro LT thermal camera is positioned inside the cabin according to the indications provided in Section 4.2. The surveys and the salivary swab are also administrated. During the first test period (BASELINE in Figure 4.5), lasting 10 minutes, the vehicle engine is off and the subject is required to sit still inside the vehicle to acquire reference data to be compared with data collected afterwards. The baseline epoch is followed by three driving phases (HIGHWAY, CO-DRIVER, HIGHWAY in Figure 4.5) where the subject is required to drive on highways and urban roads. During the urban driving, the driver is guided by a co-driver, sitting in the back of the vehicle, who provides for road indications and stressful stimuli. Finally, at the end of the driving session, data are acquired for 10 more minutes to check the driver's response after a recovery phase (RECOVERY in Figure 4.5). The salivary swab and questionnaires about the perceived stress are administered to the participants also during the final recovery phase to investigate the impact of the driving task on the subject's status.

An example of a route followed during a driving test performed in the city of Parma, Italy, is shown in Figure 4.6.

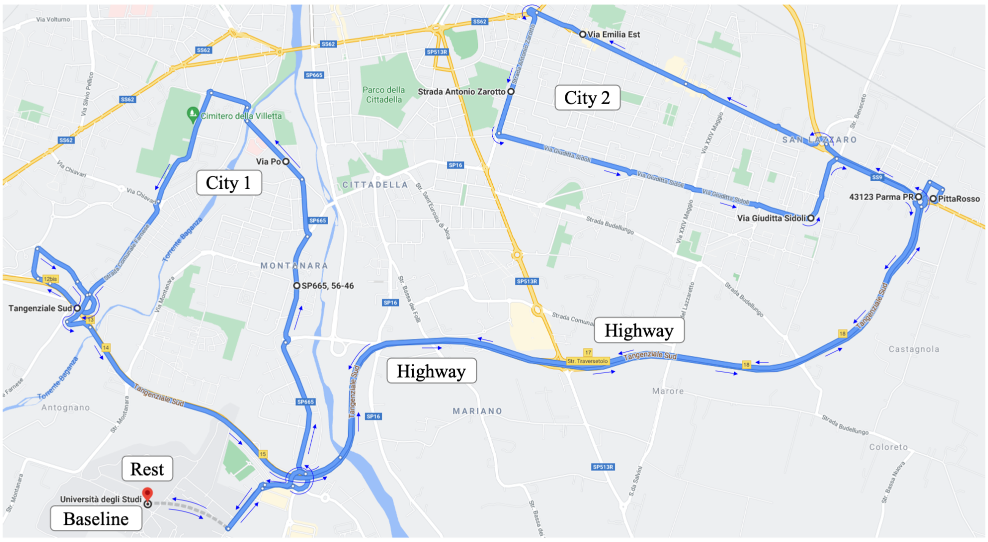


Figure 4.6: Urban and highway roads traveled in the city of Parma, Italy, during a driving test.

4.4 Data Processing for Temperature Extraction

In order to extract useful information from the collected heterogeneous data, processing techniques need to be implemented. In particular, a video processing algorithm was developed to extract the temperature from the acquired thermal videos. When a thermal video is recorded by the FLIR One Pro LT device, the temperature information is not stored due the supported file format. Thermal video sequences are, indeed, saved as mp4 files and the extracted frames are actually coded according to the RGB standard. To this purpose, the MoniDrive app, introduced in Section 4.3, has been implemented to retrieve necessary data properly handled by the proposed processing algorithm, whose schematic overview is shown in Figure 4.7.

In particular, the RGB and thermal frames are stored by the MoniDrive app along with the respective temperature scale and represent the inputs of the processing algorithm. The locations and temperature values of the hottest

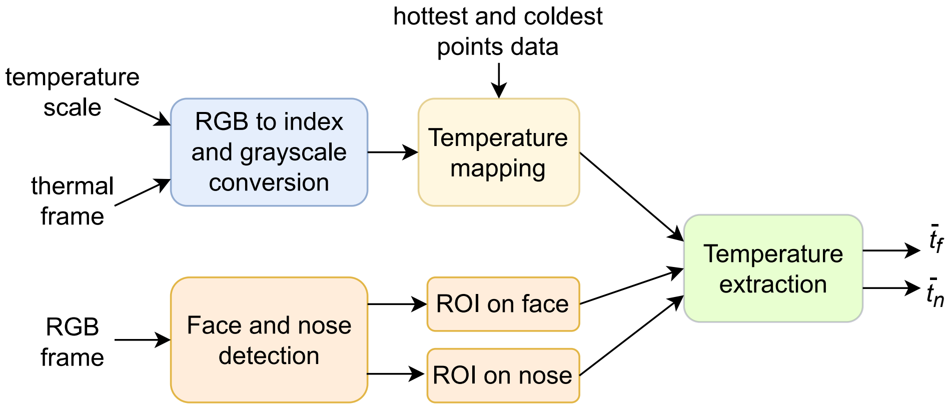


Figure 4.7: Video processing algorithm to extract the temperature from thermal frames.

and coldest points within the frame under analysis are also stored.

An RGB thermal frame is initially converted to an indexed image whose pixel intensity values are replaced by indices determined according the corresponding colormap, i.e., the temperature scale. The temperature scale contains, indeed, the values of the RGB components of each color in the image that are associated to integer indices. The indexed image is then converted to grayscale to simplify the following operations and is scaled according to a specific function to map the grayscale indices into actual values of temperature, expressed in degrees. The mapping function is defined as follows

$$I_t = t(l_1, l_2) + \big(t(h_1, h_2) - t(l_1, l_2)\big) I_i; \tag{4.1}$$

where I_t and I_i are the two-dimensional (2D) matrices representing the thermal image whose pixel values are actual temperature values and the grayscale indexed image, respectively, whereas $t(l_1, l_2)$ and $t(h_1, h_2)$ are the hottest and coldest temperature values associated to the pixels at position (l_1, l_2) and (h_1, h_2) within the indexed frame I_i.

On the other hand, the RGB frame is directly processed by the algorithm in order to detect face and nose regions through the Viola-Jones method [73], which is trained on standard RGB images. Two rectangular Regions of Interest (ROIs) are then extracted on the corresponding thermal image and the mean temperatures on the selected areas are computed according to the following spatial averaging operation

$$\bar{t} = \frac{1}{WH} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} t(w,h) \tag{4.2}$$

where W and H are the width and height of the considered region, respectively, and $t(w,h)$ denotes the temperature value associated to the pixel at position (w,h) within the ROI. In order to improve the efficiency of the method, the ROIs on the subject's nose which are incorrectly detected are discarded. Incorrect nose ROIs are declared if the upper-left corner of the corresponding area falls above the 80% or below the 25% of the height of the face region.

4.5 Experimental Results

An example of three processed frames of a video recorded during a driving test is visible in Figure 4.8, where the ROIs on the subject's face and nose are correctly detected.

The mean temperatures extracted from each frame of a considered video sequence are, instead, plotted in Figure 4.9 against time values that correspond to the video duration, expressed in seconds. The video under analysis has a total duration of around 20 minutes, but the first 7 minute interval is discarded as artefacts due to the subject's motion affect the obtained results. In Figure 4.9(a) the mean temperatures computed on the face and nose ROIs are shown along with the mean temperature extracted from the whole frame and are expressed in degrees. At time instants where the ROIs are incorrectly detected by the Viola-Jones algorithm, the temperature curves are not defined as the mean temperatures are not computed. For a further analysis, the mean temperature values extracted from the two ROIs are also normalized, at each

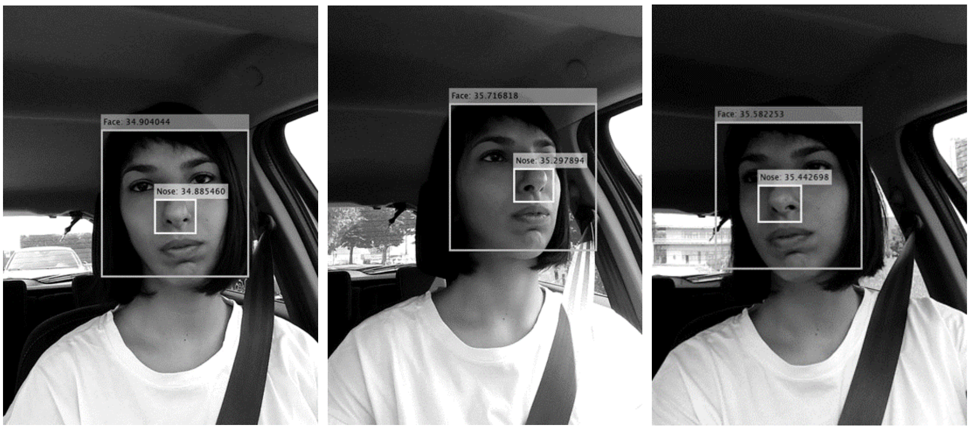


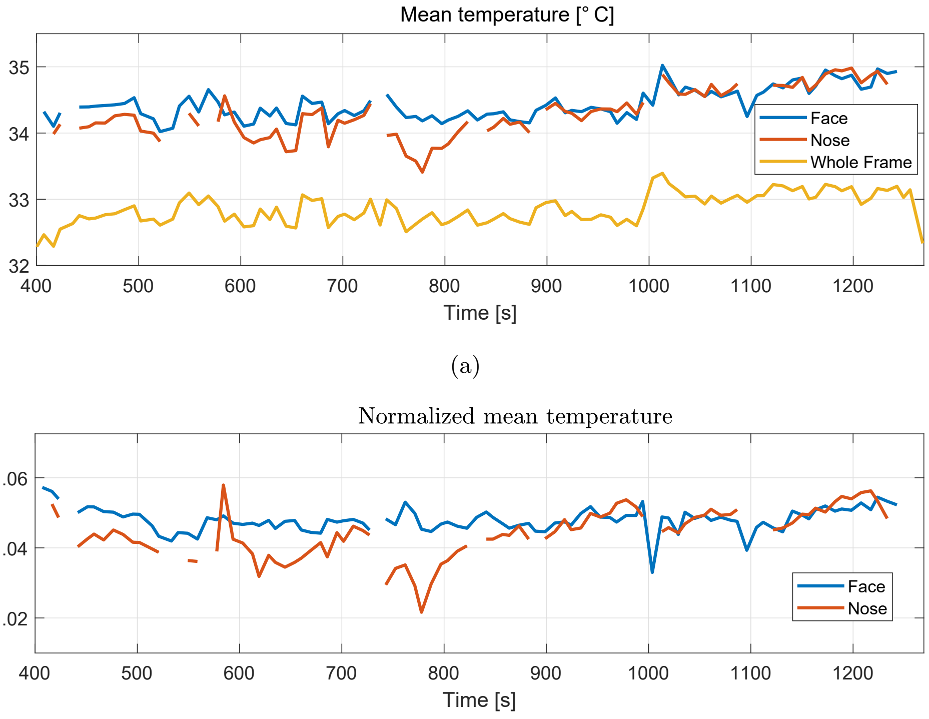
Figure 4.8: Processed frames of a video sequence acquired during a driving test.

time instant, with respect to the mean temperature of the whole frame. The normalized temperature curves are shown in Figure 4.9(b). The normalization operation is performed in order to remove undesired trends possibly due to temperature oscillations inside the vehicle and unrelated to the perceived stress.

Additional results need to be acquired in order to extract useful information about the stress level of the driver from the temperature data. The correlation with the physiological parameters extracted by the wearable sensor need also further investigation.

4.6 Conclusions

In this chapter, a non-invasive in-vehicle system for the monitoring of the psycho-physiological status of a driver is presented. The proposed infrastructure is based on an Internet of Things (IoT) network composed by different sensing devices. In particular, the joint use of a wearable sensor belt, worn by the driver, and a thermal camera, installed in the cabin of the vehicle to frame the driver's face, is investigated. The architecture of the system is pre-



(a)

(b)

Figure 4.9: Extracted mean temperatures.

sented along with a description of each component and their function. The protocol adopted to acquire data during driving sessions is also detailed and methods to process the collected information are described. In particular, a dedicated video processing algorithm to extract temperature values from the thermal images is developed. Regions of interest on the subject's face and nose are considered to analyse temperature variations possibly related to the perceived stress. Preliminary results are presented, whereas additional experimental recordings are needed to assess the robustness of the proposed system.

As a future activity, the extracted temperature information could be jointly analysed with the physiological data obtained by the wearable sensor in order to enhance the estimation of the arousal value.

Conclusions

In this thesis, video processing techniques to extract relevant information from video sequences are implemented, with a specific attention to the automotive and healthcare sectors. In particular, the application of video-based solutions to the tasks of motion estimation, breathing analysis and driver monitoring is discussed on the basis of comparisons with state-of-the-art methods and numerical evaluations, that demonstrate the feasibility and effectiveness of the presented techniques.

The first proposed application concerns the automotive sector and consists in the speed estimation of framed objects, that may be subject to other dynamic modifications. The presented method is divided into two main parts. Robust video pre-processing operations are first applied to reduce or neglect the effects of transformations due to the projection of real world scenes onto the camera plane and superimposed to the motion of the framed object. Then, speed estimation is performed by exploiting the Maximum Likelihood (ML) principle to derive a reliable estimator of the unknown speed. The effectiveness of the proposed method is demonstrated against the lower performance achieved by a reference method, i.e., the block matching algorithm, testing a number of real videos.

Periodicity is another characteristic related to motion which is investigated in this thesis. In particular, respiration, which consists of periodic movements of the chest and abdomen of a subject, is extracted from video sequences by proper processing algorithms. Two breathing monitoring systems are presented and discussed in terms of performance. Respiratory signals are first amplified and extracted by a sequence of spatial and temporal operations. Then, the ML principle is exploited to estimate the Respiratory Rate (RR) of the framed subjects. These applications demonstrate the versatility of this statistical tool, which can be applied to various problems in order to extract unknown parameters of interest.

In the last application scenario considered in this thesis, a contactless and non-invasive system to monitor the psycho-physiological status of a driver is presented. The system is composed by heterogeneous devices to extract relevant physiological information. In particular, an Infra Red (IR) camera is employed to acquire thermal images of a driver in order to analyse skin temperature variations possibly induced by stressing factors related to the driving task. The temperature information, extracted by a dedicated video processing algorithm, may be correlated to data obtained from other sensors in order to provide a robust estimation of the arousal level of the subject. In particular, a wearable device is used to measure physiological indices such as ElectroCardioGram (ECG) signals and the Heart Rate (HR).

In conclusion, the analysed performance of the algorithms developed and presented in this thesis demonstrates the effectiveness of the proposed innovative video-processing techniques, against other conventional solutions. It is also proved that these techniques may be adopted in a wide range of application scenarios, being versatile and adaptable to various tasks.

List of Publications

The list of publications based on this thesis work at time of preparation of this final version is here reported.

International Journals

(j1) V. Mattioli, D. Alinovi, G. Ferrari, F. Pisani and R. Raheli, “Motion Magnification Algorithms for Video-Based Breathing Monitoring”, under review, 2022

(j2) V. Mattioli, D. Alinovi, and R. Raheli, “Maximum likelihood speed estimation of moving objects in video signals”, *Signal Processing* (Elsevier), vol. 196, pp. 108528, July. 2022. doi:10.1016/j.sigpro.2022.108528.

International Conferences

(c1) V. Mattioli, D. Alinovi, and R. Raheli, “A Maximum Likelihood Approach to Speed Estimation of Foreground Objects in Video Signals”, in *Proc. 2020 28th European Signal Processing Conference* (EUSIPCO 2020), Amsterdam, Netherlands, Jan. 2021, pp. 715-719. doi:10.23919/Eusipco47968.2020.9287813.

(c2) V. Mattioli, D. Alinovi, and R. Raheli, “On Motion Analysis of Multiple Time-Variant Objects in Video Sequences’,’ in *2020 43rd International Conference on Telecommunications and*

Signal Processing (TSP), Milan, Italy, July 2020, pp. 445-448. `doi:10.1109/TSP49548.2020.9163577`.

Book Chapters

(b1) L. Davoli, V. Mattioli, S. Gambetta, L. Belli, L. Carnevali, M. Martalò, A. Sgoifo, R. Raheli, and G. Ferrari, "Non-invasive psycho-physiological driver monitoring through IoT-oriented systems", Chapter 2 in *The Internet of Medical Things: Enabling technologies and emerging applications*, edited by S.K. Pani, P. Patra, G. Ferrari, R. Kraleva, and X. Wang, 2021. Online ISBN 978-1-83953-274-0, Print ISBN: 978-1-83953-273-3. `doi:10.1049/PBHE034E_ch2`.

(b2) V. Mattioli, D. Alinovi, F. Pisani, G. Ferrari, and R. Raheli, "Video-based solutions for newborn monitoring", Chapter 14 in *The Internet of Medical Things: Enabling technologies and emerging applications*, edited by S.K. Pani, P. Patra, G. Ferrari, R. Kraleva, and X. Wang, 2021. Online ISBN 978-1-83953-274-0, Print ISBN: 978-1-83953-273-3. `doi:10.1049/PBHE034E_ch14`.

(b3) V. Mattioli, D. Alinovi, F. Pisani, G. Ferrari, and R. Raheli, "Respiration Monitoring by Video Signal Processing", Chapter 8 in *ICT for Health: Sensing, Data Analysis, Applications*, edited by M. Rossi, G. Cisotto, and R. Raheli, Roma, 2021.Online ISBN 978-88-94982-54-1, Print ISBN: 978-88-94982-53-4.

Oral Presentations

(o1) V. Mattioli, D. Alinovi and R. Raheli, "A Novel Method to Estimate the Speed of Moving Objects in Video Signals", Annual Meeting of GTTI, Lecce, Sept. 2021.

(o2) V. Mattioli, D. Alinovi and R. Raheli, "Maximum Likelihood Motion Estimation and Periodic Feature Extraction in Video Signals", GTTI Meeting on Multimedia Signal Processing, Virtual event, Feb. 2021.

The methods and results presented in Chapter 2 of this thesis are based on the works described in (j2), (c1) and (c2) and were partially presented during the events (o1) and (o2); Chapter 3 is based on the works presented in (b2), (b3) and on the submitted article (j1); Chapter 4 is based on the works presented in (b1).

Bibliography

[1] C. Solomon, T. Breckon, Fundamentals of Digital Image Processing, 1st Edition, Wiley-Blackwell, Croydon, UK, 2011.

[2] R. C. Gonzalez, R. E. Woods, Digital Image Processing, 2nd Edition, Prentice Hall, Upper Saddle River, NJ, USA, 2008.

[3] R. Szeliski, Computer Vision Algorithms and Applications, 1st Edition, Springer, London, UK, 2011.

[4] D. Alinovi, G. Ferrari, F. Pisani, R. Raheli, Respiratory rate monitoring by maximum likelihood video processing, in: Proc. 2016 IEEE Intern. Symp. Signal Processing and Information Technology (ISSPIT), Limassol, Cyprus, 2016, pp. 172–177. doi:10.1109/ISSPIT.2016.7886029.

[5] L. Cattani, D. Alinovi, G. Ferrari, R. Raheli, E. Pavlidis, C. Spagnoli, F. Pisani, Monitoring infants by automatic video processing: A unified approach to motion analysis, Computers in Biology and Medicine 80 (2017) 158–165. doi:10.1016/j.compbiomed.2016.11.010.

[6] A. M. Tekalp, Digital Video Processing, 2nd Edition, Prentice Hall, Upper Saddle River, NJ, USA, 2015.

[7] A. C. Bovik, Handbook of Image and Video Processing, 1st Edition, Academic Press series in communications, networking and multimedia, Academic Press, San Diego, CA, USA, 2000.

[8] D. J. Fleet and Y. Weiss, Handbook of Mathematical Models in Computer Vision, Springer, Boston, MA, USA, 2006, Ch. Optical Flow Estimation, pp. 239–257.

[9] B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proc. 7th Int. Joint Conf. Artif. Intell., Vancouver, BC, Canada, 1981, pp. 674–679.

[10] B. K. P. Horn and B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1) (1981) 185–203.

[11] R. Ke, S. Kim, Z. Li, Y. Wang, Motion-vector clustering for traffic speed detection from UAV video, in: Proc. 2015 IEEE First Intern. Smart Cities Con. (ISC2), Guadalajara, Mexico, 2015, pp. 1–5. doi:10.1109/ISC2.2015.7366230.

[12] X. Qimin, L. Xu, W. Mingming, L. Bin, S. Xianghui, A methodology of vehicle speed estimation based on optical flow, in: Proc. 2014 IEEE Intern. Conf. Service Operations and Logistics, and Informatics, Qingdao, China, 2014, pp. 33–37. doi:10.1109/SOLI.2014.6960689.

[13] X. Yu, P. Xue, L. Duan and Q. Tian, An algorithm to estimate mean vehicle speed from MPEG skycam video, Multimedia Tools Appl. 34 (1) (2007) 85–105. doi:10.1007/s11042-006-0073-8.

[14] F. Y. Hu, H. Sahli, X. F. Dong and J. Wang, A high efficient system for traffic mean speed estimation from MPEG video, in: Proc. IEEE Int. Conf. Artif. Intell. and Comput. Intell., Vol. 3, Shangai, China, 2009, pp. 444–448. doi:10.1109/AICI.2009.358.

[15] J. Jiang, C. Mi, M. Wu, Z. Zhang, Y. Feng, Study on a real-time vehicle speed measuring method at highway toll station, in: 2019 International Conference on Sensing and Instrumentation in IoT Era (ISSI), 2019, pp. 1–5. doi:10.1109/ISSI47111.2019.9043732.

[16] J. Sochor, R. Juránek, A. Herout, Traffic surveillance camera calibration by 3D model bounding box alignment for accurate vehicle speed measurement, Computer Vision and Image Understanding 161 (2017) 87–98. doi:10.1016/j.cviu.2017.05.015.

[17] P. Giannakeris, V. Kaltsa, K. Avgerinakis, A. Briassouli, S. Vrochidis, I. Kompatsiaris, Speed estimation and abnormality detection from surveillance cameras, in: Proceedings of the IEEE Conference on Com-

puter Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 93–936. doi:10.1109/CVPRW.2018.00020.

[18] R.Hartley, A. Zisserman, Multiple View Geometry, 2nd Edition, Cambridge University Press, Cambridge, UK, 2003.

[19] C. Stauffer, W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in: Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognit. (CVPR), Vol. 2, Cambridge, MA, USA, 1999, pp. 246–252. doi:10.1109/CVPR.1999.784637.

[20] A. Briassouli, N. Ahuja, Integration of frequency and space for multiple motion estimation and shape-independent object segmentation, IEEE Trans. Circuits and Systems for Video Technology 18 (2008) 657 – 669. doi:10.1109/TCSVT.2008.918799.

[21] J. Shi, Tomasi, Good features to track, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, WA, USA, 1994, pp. 593–600. doi:10.1109/CVPR.1994.323794.

[22] C. Tomasi, T. Kanade, Detection and Tracking of Point Features, Shape and motion from image streams, School of Computer Science, Carnegie Mellon Univ., 1991, accessed in October 2022.
URL https://books.google.it/books?id=20wpSQAACAAJ

[23] S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, 1st Edition, Prentice Hall, Upper Saddle River, NJ, USA, 1993.

[24] I. Costanzo, D. Sen, L. Rhein, U. Guler, Respiratory monitoring: Current state of the art and future roads, IEEE Reviews in Biomedical Engineering 15 (2022) 103–121. doi:10.1109/RBME.2020.3036330.

[25] J. J. Freundlich, J. C. Erickson, Electrical impedance pneumography for simple nonrestrictive continuous monitoring of respiratory rate, rhythm and tidal volume for surgical patients, Chest 65 (2) (1974) 181–184. doi:10.1378/chest.65.2.181.

[26] D. Yıldızdaş, H. Yapıcıoğlu, H. L. Yılmaz, Y. Sertdemir, Correlation of simultaneously obtained capillary, venous, and arterial blood gases of pa-

tients in a paediatric intensive care unit, Archives of Disease in Childhood 89 (2) (2004) 176–180. doi:10.1136/adc.2002.016261.

[27] W. H. Spriggs, Essentials of Polysomnography, 2nd Edition, Jones & Bartlett Learning, Burlington, MA, USA, 2015.

[28] C. Massaroni, A. Nicolò, M. Sacchetti, E. Schena, Contactless methods for measuring respiratory rate: A review, IEEE Sensors Journal 21 (11) (2021) 12821–12839. doi:10.1109/JSEN.2020.3023486.

[29] C. Massaroni, D. S. Lopes, D. L. Presti, E. Schena, S. Silvestri, Contactless monitoring of breathing patterns and respiratory rate at the pit of the neck: A single camera approach, Journal of Sensors 2018 (2018) 1–13. doi:10.1155/2018/4567213.

[30] S. Sanyal, K. K. Nundy, Algorithms for monitoring heart rate and respiratory rate from the video of a user's face, IEEE Journal of Translational Engineering in Health and Medicine 6 (2018) 1–11. doi:10.1109/JTEHM.2018.2818687.

[31] M. Mateu-Mateus, F. Guede-Fernández, N. Rodriguez-Ibáñez, M. García-González, J. Ramos-Castro, M. Fernández-Chimeno, A non-contact camera-based method for respiratory rhythm extraction, Biomedical Signal Processing and Control 66 (2021) 102443. doi:10.1016/j.bspc.2021.102443.

[32] R. Janssen, W. Wang, A. Moço, G. Haan, Video-based respiration monitoring with automatic region of interest detection, Physiological measurement 37 (2015) 100–114. doi:10.1088/0967-3334/37/1/100.

[33] S. Alam, S. P. N. Singh, U. Abeyratne, Considerations of handheld respiratory rate estimation via a stabilized video magnification approach, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea (South), 2017, pp. 4293–4296. doi:10.1109/EMBC.2017.8037805.

[34] A. Al-Naji, J. Chahl, Remote respiratory monitoring system based on developing motion magnification technique, Biomedical Signal Processing and Control 29 (2016) 1–10. doi:10.1016/j.bspc.2016.05.002.

[35] K. Mutlu, J. Esquivelzeta Rabell, P. Martin del Olmo, S. Haesler, IR thermography-based monitoring of respiration phase without image segmentation, Journal of Neuroscience Methods 301 (2018) 1–8. `doi:https://doi.org/10.1016/j.jneumeth.2018.02.017`.

[36] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Transactions on Graphics 31 (4) (Jul. 2012). `doi:10.1145/2185520.2185561`.

[37] N. Wadhwa, M. Rubinstein, F. Durand, W. T. Freeman, Riesz pyramids for fast phase-based video magnification, in: 2014 IEEE International Conference on Computational Photography (ICCP), Santa Clara, CA, USA, 2014, pp. 1–10. `doi:10.1109/ICCPHOT.2014.6831820`.

[38] D. Alinovi, L. Cattani, G. Ferrari, F. Pisani, R. Raheli, Spatio-temporal video processing for respiratory rate estimation, in: 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings, Turin, Italy, 2015, pp. 12–17. `doi:10.1109/MeMeA.2015.7145164`.

[39] D. Alinovi, Video processing for remote respiration monitoring, Doctoral thesis, Università degli Studi di Parma. Dipartimento di Ingegneria dell'Informazione, accessed on October 2022 (2017).
URL `https://hdl.handle.net/1889/3416`

[40] D. Alinovi, G. Ferrari, F. Pisani, R. Raheli, Respiratory rate monitoring by video processing using local motion magnification, in: 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 1780–1784. `doi:10.23919/EUSIPCO.2018.8553066`.

[41] P. Burt, E. Adelson, The Laplacian pyramid as a compact image code, IEEE Transactions on Communications 31 (4) (1983) 532–540. `doi:10.1109/TCOM.1983.1095851`.

[42] A. V. Oppenheim, J. R. Buck, R. W. Schafer, Discrete-time Signal Processing, 3rd Edition, Pearson - Prentice Hall, Croydon, UK, 2010.

[43] N. Wadhwa, H.-Y. Wu, A. Davis, M. Rubinstein, E. Shih, G. J. Mysore,

J. G. Chen, O. Buyukozturk, J. V. Guttag, W. T. Freeman, F. Durand, Eulerian video magnification and analysis, Commun. ACM 60 (1) (2016) 87–95. doi:10.1145/3015573.

[44] M. Felsberg, G. Sommer, The monogenic signal, IEEE Transactions on Signal Processing 49 (12) (2001) 3136–3144. doi:10.1109/78.969520.

[45] M. Unser, D. Sage, D. Van De Ville, Multiresolution monogenic signal analysis using the Riesz–Laplace wavelet transform, IEEE Transactions on Image Processing 18 (11) (2009) 2402–2418. doi:10.1109/TIP.2009.2027628.

[46] N. Wadhwa, M. Rubinstein, F. Durand, W. Freeman, Quaternionic representation of the Riesz pyramid for video magnification, https://people.csail.mit.edu/mrub/papers/RieszPyr_Quaternion_TechReport.pdf, accessed on October 2022 (Apr. 2014).

[47] N. Wadhwa, M. Rubinstein, F. Durand, W. T. Freeman, Phase-based video motion processing, ACM Transaction on Graphics 32 (4) (Jul. 2013). doi:10.1145/2461912.2461966.

[48] D. Alinovi, G. Ferrari, F. Pisani, R. Raheli, Respiratory rate monitoring by maximum likelihood video processing, in: 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Limassol, Cyprus, 2016, pp. 172–177. doi:10.1109/ISSPIT.2016.7886029.

[49] L. Cattani, D. Alinovi, G. Ferrari, R. Raheli, E. Pavlidis, C. Spagnoli, F. Pisani, Monitoring infants by automatic video processing: A unified approach to motion analysis, Computers in Biology and Medicine 80 (2017) 158–165. doi:10.1016/j.compbiomed.2016.11.010.

[50] N. Patwari, J. Wilson, S. Ananthanarayanan, S. K. Kasera, D. R. Westenskow, Monitoring breathing via signal strength in wireless networks, IEEE Transactions on Mobile Computing 13 (8) (2014) 1774–1786. doi:10.1109/TMC.2013.117.

[51] L. Cattani, D. Alinovi, G. Ferrari, R. Raheli, E. Pavlidis, C. Spagnoli, F. Pisani, A wire-free, non-invasive, low-cost video processing-based approach to neonatal apnoea detection, in: 2014 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings, Rome, Italy, 2014, pp. 67–73. doi:10.1109/BIOMS.2014.6951538.

[52] Equivital Life Monitor [online], https://www.equivital.com/products/eq02-lifemonitor, accessed in October 2022 (2022).

[53] W.-Y. Chung, T.-W. Chong, B.-G. Lee, Methods to detect and reduce driver stress: A review, International Journal of Automotive Technology 20 (5) (2019) 1051–1063. doi:10.1007/s12239-019-0099-3.

[54] J. D. Hill, L. N. Boyle, Driver stress as influenced by driving maneuvers and roadway conditions, Transportation Research Part F: Traffic Psychology and Behaviour 10 (3) (2007) 177–186. doi:10.1016/j.trf.2006.09.002.

[55] A. Ziebinski, R. Cupek, D. Grzechca, L. Chruszczyk, Review of Advanced Driver Assistance Systems (ADAS), AIP Conference Proceedings 1906 (1) (2017) 120002. doi:10.1063/1.5012394.

[56] L. Davoli, M. Martaló, A. Cilfone, L. Belli, G. Ferrari, R. Presta, R. Montanari, M. Mengoni, GiraldiLuca, E. G. Amparore, M. Botta, I. Drago, G. Carbonara, A. Castellano, J. Plomp, On driver behavior recognition for increased safety: A roadmap, Safety 6 (4) (2020) 1–33. doi:10.3390/safety6040055.

[57] M. Satoh, S. Shiraishi, Performance of antilock brakes with simplified control technique, in: SAE International Congress and Exposition, SAE International, United States, 1983. doi:10.4271/830484.

[58] M. Vollrath, S. Schleicher, C. Gelau, The influence of cruise control and adaptive cruise control on driving behaviour – A driving simulator study, Accident Analysis & Prevention 43 (3) (2011) 1134–1139. doi:10.1016/j.aap.2010.12.023.

[59] N. Sharma, T. Gedeon, Objective measures, sensors and computational

techniques for stress recognition and classification: A survey, Computer Methods and Programs in Biomedicine 108 (3) (2012) 1287–1301. doi:10.1016/j.cmpb.2012.07.003.

[60] R. K. Dishman, Y. Nakamura, M. E. Garcia, R. W. Thompson, A. L. Dunn, S. N. Blair, Heart rate variability, trait anxiety, and perceived stress among physically fit men and women, International Journal of Psychophysiology 37 (2) (2000) 121–133. doi:10.1016/S0167-8760(00)00085-4.

[61] U. R. Acharya, K. P. Joseph, N. Kannathal, C. Min Lim, J. S. Suri, Heart rate variability: a review, Medical and Biological Engineering and Computing 44 (12) (2006) 1031–1051. doi:10.1007/s11517-006-0119-0.

[62] R. Cassani, T. H. Falk, A. Horai, L. A. Gheorghe, Evaluating the measurement of driver heart and breathing rates from a sensor-equipped steering wheel using spectrotemporal signal processing, in: Proc. 2019 IEEE Intelligent Transportation Systems Conf. (ITSC), 2019, pp. 2843–2847. doi:10.1109/ITSC.2019.8916959.

[63] A. Lanatà, G. Valenza, A. Greco, C. Gentili, R. Bartolozzi, F. Bucchi, F. Frendo, E. P. Scilingo, How the autonomic nervous system and driving style change with incremental stressing conditions during simulated driving, IEEE Trans. on Intelligent Transportation Systems 16 (3) (2015) 1505–1517. doi:10.1109/TITS.2014.2365681.

[64] B. Eilebrecht, S. Wolter, J. Lem, H. Lindner, R. Vogt, M. Walter, S. Leonhardt, The relevance of HRV parameters for driver workload detection in real world driving, in: Proc. 2012 Computing in Cardiology, Krakow, Poland, 2012, pp. 409–412.

[65] A. Konieczka, E. Michalowicz, K. Piniarski, Infrared thermal camera-based system for tram drivers warning about hazardous situations, in: Proc. 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 2018, pp. 250–254. doi:10.23919/SPA.2018.8563417.

[66] D. Cardone, D. Perpetuini, C. Filippini, E. Spadolini, L. Mancini, A. M. Chiarelli, A. Merla, Driver stress state evaluation by means of thermal imaging: a supervised machine learning approach based on ECG signal, Applied Sciences 10 (16) (2020) 5673. doi:10.3390/app10165673.

[67] S. Ebrahimian, A. Nahvi, H. Bakhoda, A. Homayounfard, M. Tashakori, Evaluation of driver drowsiness using respiration analysis by thermal imaging on a driving simulator, Multimedia Tools and Applications 79 (25–26) (2020) 17793–17815. doi:10.1007/s11042-020-08696-x.

[68] V. Engert, A. Merla, J. A. Grant, D. Cardone, A. Tusche, T. Singer, Exploring the use of thermal infrared imaging in human stress research, PLOS ONE 9 (3) (2014) 1–11. doi:10.1371/journal.pone.0090782.

[69] C. Diaz-Piedra, E. Gomez-Milan, L. L. Di Stasi, Nasal skin temperature reveals changes in arousal levels due to time on task: An experimental thermal infrared imaging study, Applied Ergonomics 81 (2019) 102870. doi:10.1016/j.apergo.2019.06.001.

[70] S. Matsuno, T. Mizuno, H. Asano, K. Mito, N. Itakura, Estimating autonomic nerve activity using variance of thermal face images, Artificial Life and Robotics 23 (06 2018). doi:10.1007/s10015-018-0436-z.

[71] FLIR ONE Pro LT [online], https://www.flir.it/products/flir-one-pro-lt/, accessed in October 2022 (2022).

[72] O. V. Bitkina, J. Kim, J. Park, J. Park, H. K. Kim, Identifying traffic context using driving stress: A longitudinal preliminary case study, Sensors 19 (9) (2019). doi:10.3390/s19092152.

[73] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition. CVPR 2001, Vol. 1, Kauai, HI, USA, 2001, pp. I–512–I–518. doi:10.1109/CVPR.2001.990517.

Acknowledgments

This thesis concludes an important chapter of my life. The past three years have been particularly intense. I celebrated good times, but also got lost in hard ones and that is when I learned that people are the most precious treasure in life. Now it's time to express all my gratitude.

The first thank you is, and will always be, for my parents, for their unconditional love and support in every part of my life.

I wish to thank my mentor, Professor Raheli, for still carefully guiding me after many years of work together, being always available and ready to help. I feel blessed with all the opportunities he keeps providing me. I also need to thank professors Ferrari and Pisani for their support and collaboration.

I would like to thank my dear colleagues, who have always been nice to me since my first day. I thank Alessandro for being a good friend and for always brightening up the office with his congeniality. I thank Luca, Laura and Chiara, for always encouraging me and for their fundamental help and advice. Their kindness warms my heart. I thank Marco, for supporting me, too.

I also wish to thank my beloved friends, who gift me with their precious love and make my life so much happier and lighter. I feel like no words could accurately describe how much grateful and lucky I feel to have them.

The last thank you is the hardest one. With my greatest fondness, I wish to dedicate each word of this work to Davide, who always guided me with infinite patience, passion and kindness. I know he has found his own way to keep guiding me from wherever he is.