Doctoral Thesis

Non-invasive, automatic, and real-time characterization of drowsiness based on eye closure dynamics

Quentin Massoz

Supervisors:
Jacques G. Verly
Marc Van Droogenbroeck

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Engineering Sciences

Department of Electrical Engineering and Computer Science
Faculty of Applied Sciences
University of Liège
January 2019

The present thesis has been evaluated by the following members of the Jury (listed in alphabetical order):

Dr. Christer AHLSTRÖM		VTI[1], Sweden;
Dr. Clémentine FRANÇOIS		Phasya s.a., Belgium;
Prof. Pierre GEURTS	*President*	ULiège, Belgium;
Prof. Jean-Philippe THIRAN		EPFL[2], Switzerland;
Prof. Marc VAN DROOGENBROECK	*Supervisor*	ULiège, Belgium;
Prof. Jacques G. VERLY	*Supervisor*	ULiège, Belgium.

[1] VTI refers to the Swedish National Road and Transport Research Institute.
[2] EPFL refers to the "École Polytechnique Fédérale de Lausanne".

The research described in the present thesis was financially supported by the Belgian "Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture" (FRIA).

Abstract

This thesis is about drowsiness characterization systems that operate non-invasively, automatically, and in real-time from a video stream of face images, with a focus on the analysis of eye closure dynamics. In addition to providing a comprehensive discussion about the development of such systems, we present three novel systems: a baseline system, a multi-timescale system, and a parametric system. While our three systems each characterize drowsiness in their own manner, they share the following two aspects of design: (1) they use a 1-min sequence of eyelids distances as an intermediate representation, and (2) their models are trained with a ground truth of drowsiness based on impairments of psychomotor performance.

The baseline system characterizes drowsiness from a set of pre-defined ocular features, which is a typical approach used by most systems of other studies. As output, this system estimates a binary Level of Drowsiness (LoD) or the mean reaction time (RT). This system allows us to study the relationship between eye closure dynamics and performance impairments, as well as to rank the ocular features in terms of importance in the decision of the machine learning models. Our system tends to confirm that several standard ocular features, such as the number of long eye closures, the percentage of eye closure (PERCLOS), and the average eye closure duration, are well correlated with drowsiness.

The multi-timescale system characterizes drowsiness from four sets of data-driven ocular features, with each set extracted at a different timescale, i.e., window length. As output, this system estimates four binary LoDs with diverse trade-offs between accuracy and responsiveness. By combining these four LoDs, this system is able to (1) detect drowsiness onsets further in advance (but at the cost of lower accuracy), and to (2) detect drowsiness onsets with high accuracy (but at the cost of lower responsiveness). We show that the use of data-driven ocular features instead of pre-defined ones increases the performance of the system.

The parametric system characterizes drowsiness from a set of data-driven ocular features. As output, this system estimates the two parameters (mean and variance) of the instantaneous reciprocal normal ("recinormal") probability density function of drowsiness-induced RTs. The results show that this system estimates well the mean parameter, but estimates poorly the variance parameter. With the goal of understanding the inner-workings of this system, we conduct a visual, and partly analytic interpretation of the ocular features learned by this system. We show that these features, which are data-driven, are closely related to some pre-defined features typically found in the literature such as the PERCLOS and the number of long eye closures, as well as to some novel ones that we introduce such as the integration of a "droopy eye" signal and the number of "sudden recovery in alertness" events.

Résumé

Cette thèse traite de systèmes de caractérisation de la somnolence opérant de manière non-invasive, automatique, et temps-réel à partir d'un flux vidéo d'images du visage, et ce en se concentrant sur la dynamique de la fermeture des yeux. En plus de mener une discussion complète sur le développement de tels systèmes, nous présentons trois systèmes innovants : un système de référence, un système à multi-facteurs d'échelle temporel, et un système paramétrique. Bien que ces trois systèmes caractérisent chacun la somnolence de manières différentes, ils partagent deux points communs suivants dans leur conception : (1) ils utilisent, comme représentation intermédiaire, une séquence d'une minute composée des distances inter-paupières, et (2) leurs modèles sont appris avec une référence de somnolence basée sur la dégradation des performances psychomotrices.

Le système de référence caractérise la somnolence à partir d'un ensemble de paramètres oculaires pré-définis, ce qui est une approche typiquement utilisée par la plupart des systèmes de la littérature. En sortie, ce système soit estime soit prédit un niveau de somnolence binaire ou le temps de réaction moyen. Ce système nous permet d'étudier le lien entre la dynamique de la fermeture des yeux et la dégradation des performances, ainsi que d'identifier les paramètres oculaires les plus importants dans la décision des modèles. Nous confirmons bien que plusieurs paramètres oculaires standards, tels que le nombre de longues fermetures des yeux, le pourcentage de fermeture (PERCLOS), et la durée moyenne de fermeture, sont fortement corrélés avec la somnolence.

Le système à multi-facteurs d'échelle temporel caractérise la somnolence à partir de quatre ensembles de paramètres oculaires appris, où chaque ensemble est lié à un facteur d'échelle temporel différent. En sortie, ce système estime quatre niveaux de somnolence binaires avec divers compromis entre la précision et la réactivité. En combinant ces quatre niveaux de somnolence, ce système est capable (1) d'avertir à l'avance l'opérateur du véhicule d'épisodes de somnolence (au prix d'une précision réduite), et (2) de détecter avec précision les épisodes de somnolence (au prix d'une réactivité réduite). Nous montrons que l'utilisation de paramètres oculaires appris-des-données plutôt que pré-définis améliore la performance du système.

Le système paramétrique caractérise la somnolence à partir d'une ensemble de paramètres oculaires appris. En sortie, ce système estime les deux paramètres (moyenne et variance) de la fonction de densité de probabilité réciproque-normale instantanée des temps de réactions dégradés par un état de somnolence. Les résultats montrent que ce système estime avec succès le paramètre lié à la moyenne, mais pas celui lié à la variance. Afin de comprendre le fonctionnement interne de ce système, nous effectuons une interprétation visuelle—et en partie analytique—des paramètres oculaires appris. Nous montrons que ces paramètres appris sont étroitement liés à certains paramètres pré-définis utilisés dans la littérature tels que le PERCLOS et le nombre de longues fermetures des yeux, ainsi qu'à de nouveaux paramètres que nous introduisons tels que l'intégration d'un signal de "yeux tombants" et le nombre de "regains soudain de la vigilance".

Acknowledgements

I would like to first express my gratitude to my supervisors, Jacques Verly and Marc Van Droogenbroeck. I thank them both for their availability, their trust in my work, their valuable teachings, and their thorough review of this manuscript. I thank Jacques Verly in particular for giving me the opportunity to do a doctorate in the fields of computer vision and drowsiness characterization.

I would also like to thank my former and current colleagues for their support, their help, our great conversations, and the awesome times we spent together. These include, but are not limited to, Thomas H., Thomas L., Clémentine, Philippe, David, Delphine, Anaïs, Mohamed, Antoine, Thomas S., Jérôme, Eva, JB, et Maryse. Special thanks to Thomas H. and Clémentine for taking the roles of my mentors in computer vision and in drowsiness characterization, respectively. They helped me learn very quickly the tips and tricks of scientific research!

And finally, I am infinitively grateful to my friends (including those from the swim team) and my family for their well-needed support.

Contents

Chapter 1

Introduction

1.1 Context

Drowsiness is an unavoidable physiological state with significant repercussions on the mind and body. Whether drowsiness is caused by sleep degradation due to sleep apnea, by sleep restriction due to a baby regularly crying at night, or by sleep deprivation due to staying up late with friends—the negative effects are vast and include: impairments of psychomotor performance, poor decision making, mood alteration, attentional lapses, and other physical and mental health consequences [10, 57, 132, 142]. Although one can experience drowsiness at any time of the day, drowsiness becomes more likely the longer one remains awake, particularly during the night and early morning. For instance, at a waking duration of 17 hours, impairments of performance reaches levels equivalent to a blood alcohol concentration (BAC) of 0.05%, and 24 hours to 0.10% [37]. For comparison, in Belgium, the legal BAC limit that must not be exceeded to be authorized to drive is 0.05%.

Clearly, when performing a critical task, e.g., operating a vehicle or controlling air traffic, drowsiness can lead to accidents with considerable environmental, property, financial, and health-related costs. In the road transportation sector, drowsiness is a leading cause of fatal accidents. Indeed, a drowsy driver will experience difficulties in tracking lanes, in maintaining a constant speed, and in keeping a safe distance from other vehicles. On the highway, accidents thus happen in the matter of a few seconds [88], and definitely result in severe damages both to humans and property. In numbers, drowsiness is estimated to be responsible for about 10% of all road accidents [106, 131], and about 20–30% of all fatal road accidents [17, 39, 100, 131]. Each year in the United States, drowsiness therefore leads to more than 6000 deaths, and causes an estimated societal cost of 109 billion dollars [66, 101]. Notably because of these figures, the National Transportation Safety Board (NTSB), i.e., the U.S. government agency responsible for the investigation of civil transportation accidents, has made the reduction of drowsiness-related accidents an important item on its most-wanted list of transportation safety improvements since 2016 [16, 17].

However, official statistics are widely regarded as substantial underestimates of reality. Indeed, identifying drowsiness as a probable cause or a contributing factor to an accident is a challenging task to perform [49, 131]. Reasons are numerous: the lack of identifiable or conclusive evidence of such involvement during the police post-accident investigation; the unawareness or forgetfulness of the drivers about the role of drowsiness in the accident; the reluctance of the drivers to admit they had fallen asleep or were tired; and/or the death of the driver.

Even though the figures differ across studies, the majority of experts consider drowsy driving as an important traffic safety problem, and so does the public. In a 2002 National

Highway Traffic Safety Administration (NHTSA) survey [119], 95% of drivers evaluated drowsy driving by others to be a major threat to their safety, whereas 5% of them evaluated it as a minor threat. In the 2017 AAA's Traffic Safety Culture Index [5], 87.9% of drivers considered drowsy driving as a serious or somewhat serious threat, and 95.2% of them considered it as an unacceptable behavior. Yet, despise the public's correct assessment of this safety threat, 75.4% of Belgian drivers reported having experienced drowsiness at the wheel at least once in their lifetime [73], 37% of US drivers reported having nodded off or fallen asleep at the wheel at least once in their lifetime [119], and 30.8% of US drivers admitted having driven while having a hard time keeping their eyes open at least once in the past month [5].

In addition to the road transportation sector, drowsiness is responsible for serious accidents in other types of transportation sectors (aviation, marine, and rail), as well as in the industry sector. For examples, in the aviation sector, the NTSB has identified drowsiness as a contributing cause [17] to the crash of the Colgan Air Flight 3407 into a residence in Clarence Center, New York, in 2009 (50 fatalities); the crash of a sightseeing tour helicopter in Las Vegas, Nevada, in 2011 (5 fatalities); the UPS Flight 1354 accident in Birmingham, Alabama, in 2013 (2 fatalities). In the industry sector, human errors and poor judgments induced by drowsiness have been recognized as having contributed to the disaster at the Three Mile Island nuclear plant in Dauphin County, Pennsylvania, in 1979 [29]; the catastrophe at the Chernobyl nuclear plant in Pripyat, Ukraine, in 1986 [30]; and the launch disaster of the Space Shuttle Challenger in Florida, in 1986 [28].

Automatic and real-time drowsiness characterization systems have certainly the potential to prevent transportation accidents by issuing timely drowsiness warnings to the vehicle operator. Such systems are generally based on driving performance (e.g., wheel steering, braking, and line crossing) and/or operator physiology (e.g., brain signals, heart rate, and facial expressions). In addition to providing drowsiness warnings, such systems could also adjust the way the vehicle systems operate, and this as a function of the driver's estimated level of drowsiness. For instance, collision avoidance systems could adjust their settings and start braking further in advance; integrated navigation systems could give directions to the nearest rest area, and inform about the most-adequate countermeasures against drowsiness given the remaining trip duration; and (semi-)autonomous systems could autonomously bring the vehicle to the nearest rest area. Furthermore, drowsiness monitoring systems could provide valuable information to investigators for determining the contributing causes of an accident.

1.2 Goals and approach

The main goal of this thesis is to develop novel drowsiness characterization systems that operate automatically from a video stream of face images, with a focus on the analysis of eye closure dynamics. Systems based on eye closure dynamics have the significant advantages of being mostly independent of applications and vehicle types, less sensitive to external conditions (e.g., weather, and traffic), and non-intrusively implementable with remote sensors such as cameras. As such, they require no action from the vehicle operator other than reacting to the timely-issued drowsiness warnings. The dynamics of eye closures is recognized as a strong and reliable physiological indicator of drowsiness [4, 40, 41, 122, 136]. And, considering that blinks naturally occur once every few seconds, eye closure dynamics constitutes a regular stream of insights about the physiological impacts of drowsiness. This inherent attribute makes the eye closure dynamics an indicator of choice to base automatic, real-time drowsiness characterization systems upon.

Training these automatic, real-time systems requires data labeled with a measure of drowsiness, i.e., a ground truth of drowsiness. Unfortunately, the level of drowsiness is not a precisely and numerically defined quantity that can be measured directly. Therefore, the practical approach to quantifying drowsiness is by characterizing it (i.e., describe its distinctive nature) based on measurable, but imperfect, indicators of drowsiness. The choice of which indicator to use depends on whether it will be used (1) as an input to the system, or (2) to produce a ground truth to train and evaluate the system. When used as an input, the indicator has to be automatically measurable in operational settings, which is the case for driving performance, facial expressions, and eye closure dynamics. When used to produce a ground truth, the scientific community has yet to reach a clear consensus on which indicator is the best. Therefore, the choice of indicator is typically based on its ease of use, and on whether the study protocol enables its acquisition or not.

In this thesis, to produce a ground truth of drowsiness, we made the choice of using the reaction times (RTs) observed during the performance of a Psychomotor Vigilance Task (PVT). The impairment of RTs is recognized as a reliable and sensitive performance-based indicator of drowsiness [10, 13, 14, 42]. The RT has the significant advantages of being automatically-annotated, of being objective (i.e., free of the subjective interpretation of human annotators), of being relatively densely-annotated (every few seconds), and of embodying meaningful effects of drowsiness *w.r.t.* operating a vehicle. Note that there exist diverse approaches to producing a ground truth of drowsiness from a set of RTs. We consider and present some of them in this thesis. Also note that, although RTs are well-suited for producing a ground truth of drowsiness, they are not well suited to be used as inputs to an automatic system since their stimulus-based acquisition would hinder the performance of the main task in operational settings, e.g., operating a vehicle.

In addition to the development of systems, the secondary goals of this thesis are (1) to provide a comprehensive discussion about the development of drowsiness characterization systems, and (2) to propose novel algorithms for the extraction of the eyelids distance from a face image.

1.3 Contributions and outline

In **Chapter 2**, we present a comprehensive review on drowsiness and its characterization. In particular, we (1) define drowsiness, (2) review the determinants of its onset, (3) address the question whether drowsy individuals can reliably evaluate their own level of drowsiness, and (4) report on the efficient and inefficient countermeasures against drowsiness at the wheel. Moreover, we propose a classification of indicators of drowsiness, present their associated standard measures, and identify which indicators are most suited to be used as (1) an input or as (2) a ground truth for the development of drowsiness characterization systems. Finally, we provide a technical review of the scientific literature on automatic, real-time drowsiness characterization systems based on eye closure dynamics.

In **Chapter 3**, we describe the sleep-deprivation dataset we collected for the purpose of developing the novel systems presented in this thesis. There are many possibilities for designing the protocol of such dataset. Therefore, we motivate our design choices but still point out our limitations and potential improvements. Furthermore, we discuss, as best as possible, the ecological validity of such laboratory dataset, i.e., the extent to which the conclusions and findings drawn from a laboratory dataset can be generalized to real-life, operational settings. Although very little work discuss this intricate topic, we highlight some key considerations to keep in mind when developing drowsiness characterization systems intended to be adapted for real-life, operational settings. Finally, we analyze the diverse indicators of drowsiness that we recorded alongside images of the face, and highlight some of their properties.

In **Chapter 4**, we present a baseline drowsiness characterization system. We designed this system to extract a set of pre-defined ocular features prior to characterizing drowsiness, which is a typical approach used by most systems of other studies. In addition to providing a performance baseline, this system enabled us to study the relationship between eye closure dynamics and impairments of psychomotor performance, as well as to rank the ocular features in terms of importance in the decision of the machine-learning classification/regression models. This chapter is based on the following published conference paper [91]:

- Q. Massoz, T. Langohr, C. François, and J. Verly. The ULg multimodality drowsiness database (called DROZY) and examples of use. In IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–7, Lake Placid, NY, USA, March 2016.

In **Chapter 5**, we present a novel multi-timescale drowsiness characterization system that aims at dealing with the trade-off between accuracy and responsiveness. Indeed, we identified that most systems typically make use of eye closure dynamics by averaging ocular features over a time window of fixed length. However, this strategy suffers from an inherent trade-off between accuracy (best achieved with long time windows, i.e., at long timescales) and responsiveness (best achieved with short time windows, i.e., at short timescales). With the goal of satisfying both accuracy and responsiveness, our multi-timescale system characterizes drowsiness from a sequence of eyelids distances by using four binary classifiers operating at four distinct timescales (5s, 15s, 30s, and 60s). We jointly trained these data-driven classifiers using a carefully-crafted multi-timescale ground truth of drowsiness based on impairments of psychomotor performance. We discuss how to combine and use these four drowsiness classifiers in operational settings. This chapter is based on the following published journal article [93]:

- Q. Massoz, J. Verly, and M. Van Droogenbroeck. Multi-timescale drowsiness characterization based on a video of a driver's face. Sensors, 18(9):1–17, August 2018.

In **Chapter 6**, we present a novel parametric drowsiness characterization system that aims at estimating the instantaneous probability density function (pdf) of drowsiness-induced reaction times (RTs). Whereas our previous systems were trained with a continuous or discrete ground-truth quantity mapped from a set of RTs, our parametric system is trained directly with such a set of RTs as ground truth. Considering that the RT follows relatively well a reciprocal normal distribution, our parametric system can therefore learn to estimate the pdf parameters by maximizing their likelihood given the set of observed RTs. In the same process, our system also learns to extract the most informative features related to eye closure dynamics. With the goal of understanding the inner-workings of our system based on convolutional neural networks, we conduct a visual interpretation, and a first analytic analysis, of these data-driven ocular features. This chapter is based on the following published conference paper [92]:

- Q. Massoz and J. Verly. Vision-based system for monitoring vehicle operator responsiveness from face images. In International Conference on Managing Fatigue, pages 1–3, San Diego, CA, USA, March 2017.

In **Chapter 7**, we conclude this thesis, identify the limitations and potential improvements, and discuss the path forward for mitigating drowsy driving.

The list of publications is given at the end of this thesis.

Chapter 2

Background on drowsiness and its characterization

This chapter presents a background on drowsiness, as well as a review on its characterization in operational and non-operational settings. Section 2.1 defines drowsiness. Section 2.2 reviews the determinants and influencing factors of drowsiness. Section 2.3 discusses whether drowsy individuals are reliably self-aware of being drowsy. Note that this latter topic is fundamentally important for motivating operational drowsiness characterization systems. Section 2.4 reviews the reliable and non-reliable countermeasures against drowsiness at the wheel. Section 2.5 reviews the diverse indicators of drowsiness, along with their standard measures typically found in the literature. Section 2.6 reviews related work on drowsiness characterization systems, in particular with respect to the design choice of ground truth of drowsiness and of system architecture. Section 2.7 concludes this chapter.

2.1 Definitions of drowsiness

Drowsiness is defined as the state of being drowsy, that is, having a difficulty of staying awake, a strong inclination toward falling asleep. Here follow the main characteristics of drowsiness:

- it is an intermediate state between fully awake and asleep;
- it is experienced at a continuous level that varies in time;
- it is characterized by physiological changes [4, 7, 10, 15], and by impairments of both cognitive performance [10, 11, 62] and psychomotor performance [44, 54, 55].

It is also often interpreted as a basic physiologic need state, like hunger and thirst, existing for the survival of the individual organism [117]. As such, drowsiness could be seen as a way for our own body to indicate that we need to sleep, similarly to hunger indicating a need to eat, and thirst a need to drink. Complying with this need, i.e., by sleeping, reverses the state of drowsiness—at least for healthy individuals.

Note that the term "drowsiness" is regularly used interchangeably with the terms "sleepiness", "somnolence", and "fatigue". However, the use of the term "fatigue" is not recommended as it can also describe the subjective feeling of tiredness, induced by prolonged periods of physical exercise or cognitive activity. Fatigue and drowsiness are distinct states; the first is alleviated by taking a break, whereas the second is alleviated by sleeping. Yet, fatigue and drowsiness are not unrelated: both manifest themselves as the inability to complete a task at normal performance. It is therefore easy to misdiagnose one for the other.

In this thesis, we consider only the terms “drowsiness”, “sleepiness”, and “somnolence” as synonyms.

In the field of sleep medicine, drowsiness is not defined as a time-varying state, but it is instead defined as a symptom with a high “propensity to fall asleep” [10, 75]. In this field, drowsiness represents an ease of falling asleep as measured objectively by the sleep latency, i.e., the time it takes to fall asleep, when the patient is instructed either to “try to sleep” [25] or to “try to stay awake” [96]. To sleep clinicians, the sleep latency is a valuable metric, as a low-enough value indicates that the patient suffers from a sleep disorder such as excessive somnolence or narcolepsy [96].

In this thesis, we will not adopt this propensity definition as we are interested in characterizing the state of drowsiness automatically, in real-time, and in situations where the individual is busy performing a task such as driving. In such active situations, the individual is torn between his/her need to sleep and the necessity to perform his/her task, and thus adopts perceptible countermeasures to fight the urge to sleep. Therefore, we define drowsiness as a physiological state, the continuous level of which varies in time.

2.2 Determinants of drowsiness

The continuous and time-varying level of drowsiness is determined by diverse factors, which are mostly related to sleep. In this section, we take the time to introduce these determinants of drowsiness, and alongside a few figures, so as to inform about how one could preventively reduce the risk of being drowsy in critical situations.

2.2.1 Sleep-wake cycle

To a great degree, drowsiness is driven by the sleep-wake cycle, the timing and structure of which are considered well described by the two-process model [3, 10]. The first process, called the sleep-wake homeostasis, corresponds to a sleep debt/pressure that builds up during wakefulness (mainly due to daytime cerebral/cognitive activity) and declines during sleep in a non-linear fashion [10]. It explains why the level of drowsiness increases during the day, and increases day by day when the sleep duration is lower than needed. The second process, called the circadian rhythm, corresponds to a body clock oscillating with a 24-h period that regulates the favorable periods of sleep and wake throughout the day. It explains why the level of drowsiness peaks in the early morning (at 2–6 am) and in the mid-afternoon (at 2–4 pm), and also why one is relatively more alert outside of these peak hours [3, 69, 117]. In such a manner, the effective sleep-wake cycle, and therefore drowsiness, is driven by the dynamic balance between these two regulatory processes.

2.2.2 Sleep quantity

As modeled by the homeostatic process, the quantity of sleep directly impacts the amount of daytime drowsiness. The quantity of sleep can be shortened either by acute sleep deprivation, i.e., with a wake duration greater than the usual duration of 16–18 hours, or by sleep restriction, i.e., with a reduced sleep duration for prolonged periods. Based on sleep research, the recommended quantity of sleep per night is of 7–8 hours for adults [142], and of about 9–10 hours for adolescents and youths [3, 117]. Extending bedtime beyond this recommended baseline has been shown to be beneficial as it increases alertness (i.e., decreases drowsiness) throughout the day [117]. Recent surveys reported that 35–40% of the adult U.S. population sleep less that 7–8 hours, and about 15% less than 6 hours [10]. Moreover, another survey reported that 50% of French drivers sleep intentionally less the night prior to a departure, and 10% do not even sleep at all [3]. These numbers are quite alarming because sleep deprivation and sleep restriction have been shown to induce—in addition to drowsiness and impairments of performance—a mood of irritability and reduced empathy, higher sensitivity and reactivity to stress, and several health-related consequences such as an increased risk of obesity, diabetes, hypertension, cardiovascular diseases, and all-around mortality [10, 42, 62, 111, 142].

For the purpose of comparison, Dawson and Reid [37] measured the impairments of performance of subjects in two distinct conditions: acutely sleep deprived and alcohol intoxicated. They showed that performance impairments after 17 hours of sustained wakefulness (measured at 1 am) were comparable to those with a blood alcohol concentration (BAC) of 0.05%, whereas 24 hours of being awake (measured at 8 am) was comparable to a BAC of 0.10%. In other words, a drowsy driver with minimal alcohol levels may be as dangerous as an alert driver who is legally intoxicated. In another study, Van Dongen *et al.* [43] showed that sleep restriction to 4 hours per night during two weeks led to comparable impairments of performance than acute sleep deprivation during 3 days. However, note that sleep loss appears to affect each individual differently [15]. Indeed, some individuals are very resilient to sleep loss and sustain minimal performance impairments from it, some are moderately affected by it, and others are particularly vulnerable to it. These differences have been shown to be trait-like, which may reflects underlying genetic involvements [15, 57, 120].

2.2.3 Sleep quality and sleep disorders

Daytime drowsiness also strongly relates to the quality and continuity of sleep [3, 117], which can be severely degraded and fragmented by various sleep disorders. Many sleep disorders, such as sleep apnea and narcolepsy, lead to the symptom called "excessive daytime sleepiness". About half of the patients with "excessive daytime sleepiness" report having road and/or work accidents, with some being life threatening and some costing them their job [117]. Common sleep disorders include insomnia, sleep apnea, and narcolepsy; information about them follows. Insomnia is characterized by a difficulty of falling asleep and/or difficulty of maintaining sleep. Insomnia is often considered as a symptom, but can be considered as a sleep disorder when no other cause is found. It is present in approximately 20–30% of the adult general population [3, 118]. Sleep apnea is characterized by frequent, brief micro-awakenings with duration of up to 15 seconds during sleep, leading to a non-restorative sleep. It is present in about 3–7% of adult men and 2–5% of adult women [111], and is also commonly present in the elderly population and overweight population [3]. Narcolepsy is characterized by "excessive daytime sleepiness" even after having an adequate night of sleep. A narcoleptic patient is likely to experience unavoidable episodes of drowsiness and fall asleep several times a day, indiscriminately of the place, situation, or time of day. It is rather a rare condition as it is estimated to be present in 0.047% of the European general population [105] and in less than 0.01% of the general population [3].

2.2.4 Other factors

On top of the quantity and quality of sleep, drowsiness can also be influenced by one or more external factors including sleep inertia; jet lag; alcohol, drugs, and medication intakes [3, 117]; bad environmental conditions such as high density traffic, poor visibility, and darkness [1, 3]; the type, length, and monotonicity of the task that is being performed [3, 7, 108]; and irregular and prolonged working schedules [2, 3, 7]. Some of these external factors are more likely to affect some sub-populations than others. For example, young drivers, in addition to having higher requirements of sleep, tend to overestimate their driving skills and more often engage in risky behaviors such as driving under the influence of alcohol and drugs [3]. Likewise, professional drivers are particularly affected as their work involves long and monotonous drives every day, as well as regular modifications in their sleep-wake schedules.

2.3 Self-awareness of drowsiness

Are drowsy individuals self-aware of being in a state of drowsiness while performing a task? If so, is this self-assessment reliable and accurate? Having concrete answers to these important questions has many implications, such as the motivation for drowsiness characterization systems (like the ones developed in the present thesis) and the validity of subjective self-ratings of drowsiness. Unfortunately, studies often draw contradicting conclusions on this matter. On the one hand, many studies conclude that people are well aware of their drowsiness [4, 11, 54, 55, 68, 78, 88, 95, 123], as evidenced by the correlation between their subjective ratings and objective indicators of drowsiness such as performance impairments or physiological changes. On the other hand, many studies conclude that people are unable to correctly rate their drowsiness [10, 43, 77, 97, 107], as evidenced by the lack of correlation between subjective and objective ratings. It is likely that these disputed results are caused by differences in the experimental protocol, the subjective scale they used, and/or the way in which the subjects had their sleep duration reduced, i.e., the type of sleep loss. The type of sleep loss has been proven of importance. Indeed, Van Dongen *et al.* [43] studied the effects of the type of sleep loss on the subjective ratings and performance impairments. They found that (1) sleep-deprived subjects (acutely over three days) reported self-ratings that were coherent with their increasing performance impairments, but that (2) sleep-restricted subjects (cumulatively over two weeks) reported significantly lower self-ratings even though they experienced equivalent performance impairments to sleep-deprived subjects. These findings suggest that, once sleep restriction becomes chronic, individuals lose the ability to reliably assess their drowsiness, as if they had forgotten how being alert feels like and consider “drowsy” as the new “alert”. This loss-of-perspective theory is further supported by the fact that patients suffering from chronic sleep disorders may consider themselves as alert even when they fall asleep throughout the day [117].

Furthermore, a fair amount of considerations are generally raised against the validity and reliability of self-ratings of drowsiness. Indeed, many researchers consider self-ratings of drowsiness to be susceptible to manipulation [122]. It implies that, by knowing the fact that the level of drowsiness typically increases with sleep deprivation, subjects report increasingly greater self-ratings as the study they participate in progresses, thereby introducing fake correlations between self-ratings and objective ratings in studies. This manipulative behavior has yet—to our knowledge—to be scientifically proven, which may prove to be challenging if subjects adopt such behavior unconsciously. Besides, it is likely that by periodically asking subjects to self-rate their drowsiness, their perception of drowsiness is being heightened [68, 116]. It implies that drivers may need to ask themselves whether they are drowsy to become aware of any drowsiness, otherwise they may not be so aware of it. This may be true, but then educating drivers about the risks of drowsy driving may encourage them to ask themselves this question more often, which could help prevent accidents.

That being said, the majority of research data supports the fact that healthy, educated individuals are able to reliably report on their drowsiness when induced by acute sleep deprivation. Nevertheless, even with a good perception of drowsiness, not everyone has a good perception of the likelihood of falling asleep while being drowsy [116]. In practice, drowsy drivers often underestimate how much time it would take them to fall asleep [69], and then even deny having fallen asleep when this happened for a short duration [18, 69].

2.4 Countermeasures against drowsiness at the wheel

The best advice to a driver falling asleep at the wheel is to stop driving as soon as possible. However, lots of drivers keep driving anyway [8]. The given reasons for why they do so are that they want to get to their destination, that they are close to home, and/or that they are in a hurry [8]. Even though drivers correctly perceive stopping to take a nap and swapping drivers as the most effective countermeasures [8], that knowledge of the safest strategies does not translate into their actual use. Anund *et al.* [6] identified the most self-administered countermeasures to be: stopping to take a walk (reported by 54% of the 1885 respondents), turning on the radio (52%), opening a window (47%), drinking coffee (45%), and engaging in social interactions (35%). However, only 18% of drivers reported counteracting drowsiness by stopping to take a nap. Further analysis indicated that the following sub-populations were more likely to adopt naps as countermeasures: individuals that have already been involved in sleep-related crashes, individuals that have already experienced driving with severe drowsiness, professional drivers, males, and drivers aged 46–64 years.

Several laboratory studies have been conducted to evaluate the effectiveness of the most common countermeasures. In this endeavor, naps (up to 15 minutes long) and intakes of caffeine (150 mg) were found to be effective countermeasures [70]. The effects of caffeine were found to be more consistent than those of naps as not every subjects managed to fall asleep on the spot. Experts even recommend to combine both by drinking a caffeinated beverage immediately before a 15-min nap [3], because the effects of caffeine take about 20 minutes to kick in. On the opposite, taking a break, turning on the radio, or opening the window were found to be ineffective countermeasures [70, 115]. The radio and the opened window have—at best—only temporary effects that would give the driver just enough time to find a suitable resting area [115].

Lastly, it is of great importance to mention that not all countermeasures are self-administered; well-accommodated road infrastructures and technologies can be considered as countermeasures against drowsiness. Indeed, the presence of rumble strips (also known as sleeper lines) and resting areas have significant, positive impacts on the reduction of the number of accidents on the highway [3]. Furthermore, technologies such as collision avoidance systems, lane departure warning systems, and drowsiness characterization systems have the clear potential to reduce accidents and save lives on the road.

2.5 Indicators of drowsiness, and their standard measures

Drowsiness is a complex physiological state that one can experience at diverse continuous levels. Indeed, a drowsy individual can be slightly drowsy, moderately drowsy, critically drowsy, or at any levels in-between. However, the level of drowsiness is not a precisely and numerically defined quantity that can be directly measured. Therefore, the practical approach to quantifying drowsiness is by characterizing it (i.e., describe its distinctive nature) based on measurable, but imperfect, indicators of drowsiness. We distinguish four categories of indicators: (1) the indicators based on physiology, (2) those based on impairments of performance, (3) those based on spontaneous facial expressions, (4) those based on subjective ratings. These indicators are measurable, and associated with multiple standard measures proven to be sensitive to drowsiness.

In this section, we present the most common indicators of drowsiness and their standard measures. It is important to note that, to date, there is no clear consensus on which indicator is best to use to characterize drowsiness [75]. However, the most used indicators of drowsiness are the brain activity, eye movements and eye closures, psychomotor performance, driving performance, and subjective assessment. The structure of the subsections below follows that of Table 2.1, which contains an overview of the indicators of drowsiness and their related standard measures.

Category	**Indicator of drowsiness**	**Standard measures**
Based on physiology	Brain activity (EEG)	θ and α (relative) powers
	Eye movements and closures (EOG)	Mean blink duration and interval; mean closure and opening speed
	Experts scoring (EEG & EOG)	KDS; OSS
	Heart rate (ECG)	LF/HF power ratio of HRV
	Skin conductance	Mean frequency of skin response; skin conductance level
Based on impairments of performance	Psychomotor performance	Number of lapses; mean RT; mean RS (to a stimulus-based task)
	Cognitive performance	Mean RT (to a cognitive task)
	Driving performance	SDLP; TLC; steering wheel variability
Based on spontaneous facial expressions	Eye closures	PERCLOS; mean blink duration and interval; mean closure and reopening speed
	Pupil diameter instabilities	Mean and variability of pupil diameter; power of diameter variations at LF
	Yawns	Occurrence frequency
	Eyebrows rises	Correlation with eye openness
	Head pose	Variability of head roll and pitch
	Experts scoring	Scale of [104]; scale of [144]
Based on subjective ratings	Self-assessment via a questionnaire	KSS; SSS; VAS; TSS

Table 2.1 – Overview of drowsiness indicators and their standard measures.

2.5.1 Indicators based on physiology

The state of drowsiness is first and foremost a physiological state. In reality, drowsiness is manifested by glands releasing various hormones, mainly melatonin [48], as a way to signal to multiple organs to regulate their physiology and functioning behavior. These physiological and behavior changes—albeit complex—can be measured via electrodes conveniently positioned in contact with the skin. Electroencephalography (EEG) enables the electrophysiological monitoring of the brain activity, the electrooculography (EOG) enables the monitoring of the eye movements, the electrocardiogram (ECG or EKG) enables the monitoring of the heart activity, and skin conductance electrodes enable the monitoring of the electrodermal activity. Note that, while they are pragmatic and practical to measure physiological changes in well-controlled laboratory settings, these electrodes-based methods are not well suited for operational use, i.e., in real-world situations, mostly due to the occurrence of many artifacts and noises induced by vibrations, movements, and loose electrodes.

Brain activity and eye movements

The most substantial changes in physiology are found in the brain activity and eye movements, as measured by the EEG and EOG, respectively. In a general context, different frequency bands in the brain waves correlate with different physiological states. As such, delta (δ) waves (0.5–3 Hz) are generally associated (for healthy individuals) to deep sleep, theta (θ) waves (4–7 Hz) to drowsiness and idling, alpha (α) waves (8–12 Hz) to relaxation and closed eyes, and beta (β) waves (13–25 Hz) to active thinking and alertness. In a context where an individual is performing a task such as driving, drowsiness has been proved to be characterized by increased levels of energy in the θ and α bands [4], as well as long blinks and slow eye movements (SEM) [4, 122]. The standard measures of brain activity are the θ and α powers, P_θ and P_α, which are generally (1) extracted via spectral analysis, (2) expressed relative to either the total EEG power [110] or some baseline power measured in a state of alertness [4], and (3) finally combined in various ways (e.g., the sum $P_\theta + P_\alpha$ or the ratio $\frac{P_\theta+P_\alpha}{P_\beta}$). The standard measures of eye movements include the mean and standard deviation of blink durations, blink intervals, eye closure speeds, and eye opening speeds [122]. Furthermore, trained experts can also score drowsiness by visually counting the symptoms in both the EEG (i.e., θ and α activity bursts) and the EOG (i.e., SEMs). To this end, several objective scales have been developed, such as the Karolinska Drowsiness Scale (KDS) [7, 55] and the Objective Sleepiness Scale (OSS) [99, 109]. In practice, these expert-produced scores are each associated with a 20-s epoch, and range from 0% to 100% by steps of 10% for KDS and from 0 to 4 for OSS. Table 2.2 contains the criteria of OSS.

OSS	**Cumulative duration of θ and/or α activities**	**Blinks and eye movements**
0	Negligible	Normal
1	Less than 5s	Normal
2	Less than 5s or Less than 10s	Slow Normal
3	Less than 10s or More than 10s	Slow Normal
4	More than 10s	Slow

Table 2.2 – The Objective Sleepiness Scale (OSS), adapted from [109].

Heart activity

Changes in physiology also manifest themselves in the heart activity, as measured by the ECG. Indeed, as drowsiness increases, the heart rate (HR) decreases and the heart rate variability (HRV) increases [137]. HRV is generally analyzed in the frequency domain via spectral analysis, and decomposed into a low frequency (LF) band (0.04–0.15 Hz) and a high frequency (HF) band (0.15–0.4 Hz). The standard measures of heart activity include the ratio of the LF power to the HF power which decreases as drowsiness increases [137]. However, note that the HR is influenced by various other factors such as age, health condition, stress, anxiety, and, most importantly, body movements, which increase the difficulty in linking its variation to drowsiness.

Electrodermal activity

Drowsiness also leads to changes in the electrodermal activity (also known as skin conductance or galvanic skin response) which relates to the electrical resistance measured via electrodes on the surface of the skin. This skin resistance fluctuates with sweating, the level of which is controlled by the sympathetic nervous system, which also autonomously regulates emotional states such as drowsiness [95]. However, skin conductance is influenced by the environment (temperature, humidity, *etc.*) and other arousing factors (stress, anxiety, emotions, demanding tasks, *etc.*). Therefore, skin conductance is, in practice, an indicator of unspecified arousal rather than purely drowsiness. However, it becomes a good indicator of drowsiness under neutral, controlled conditions (e.g., opened eyes and resting conditions). The standard measures of electrodermal activity include the skin conductance level and the mean frequency of the (non-specific) skin conductance response [95].

Gold standard

Overall, changes in physiology as captured via the EEG and EOG are recognized as well-validated indicators of drowsiness. As a matter of fact, their combination is often—but arguably—called the “gold standard” to estimate drowsiness. Concerning the other physiological indicators such as heart rate, and skin conductance, researchers have certainly linked them to drowsiness in controlled laboratory settings, but have yet to demonstrate their ability to automatically detect drowsiness in more complex situations such as driving in real-world settings [84].

2.5.2 Indicators based on impairments of performance

As by-products of the above physiological changes, the performances of a drowsy individual on diverse—and sometimes critical—tasks are impaired. When tasked to respond quickly to a sudden event, a drowsy individual will inevitably demonstrate a slower reaction time [13, 24]. Moreover, when tasked to solve a cognitive challenge, a drowsy individual will likely demonstrate reductions in innovation; flexibility of thinking; the abilities to avoid distractions and to communicate effectively; and the assessments of the risk, the task feasibility, and their own strengths and weaknesses [11].

Psychomotor and cognitive performance

In the context of driving, the performance requirements may range from simple to highly complex depending on the situation. In the most extreme of cases, the driver has to react quickly to sudden and unexpected situations by first deciding on the best strategy to adopt, and then executing it in a timely manner. It is thus not surprising that the standard measures of impairments of performance are related to the reaction time (RT), the impairment of which is known to be a reliable and very sensitive indicator of drowsiness [10, 13, 14, 40, 42, 88].

For the psychomotor performance, the RT is defined as the time that it takes to react to a visual or auditory stimulus. A well-known stimulus-based task is the Psychomotor Vigilance Task (PVT), which has the advantages of being simple and requiring minimal mental processing. The standard measures of psychomotor performance include the number of lapses (where a lapse is conventionally defined as a RT greater than 500ms [44]), the mean RT, the mean reaction speed (RS, corresponding to the reciprocal of the RT), and the mean of the 10% fastest (or slowest) RTs (or RSs) [13].

For the cognitive performance, the RT is defined as the time that it takes to solve and answer a cognitive task. We did not find any standard measure of cognitive performance, but we found the work of Baranski *et al.* [11] to be providing an interesting measure of cognitive performance. Their measure is the mean RT, as measured by the time it takes to calculate the sum of eight numbers consecutively presented every 1.25s. In this way, they were able to adjust the task difficulty by selecting sets of different numbers to add (e.g., a level-one difficulty requires to add numbers in the range 1–2, a level-three in the range 4–8, and a level-six in the range 12–16). It is important to note that this cognitive task is not excessively complex. Because, when the task becomes too complex, subjects will most probably apply compensatory efforts and perform normally [62], and by doing so render the measure of cognitive performance significantly less sensitive to sleep loss.

Driving performance

In practice, the RT is preferably measured in controlled laboratory settings. However, measuring the RT is not always feasible or adequate, especially in operational settings such as on the road. Under these circumstances, the RT task for measuring the RT is secondary, and would hinder the performance of the main task, i.e., driving. An alternative approach is to take advantage of the fact that the driver is already performing a task, i.e., driving, and measure their performance at this task, which is known to degrade with increasing drowsiness [50, 84, 145]. Standard measures of driving performance include standard deviation of lateral position (SDLP), steering wheel variability, and time to line crossing (TLC) [56, 136]. Indeed, as drowsiness progresses, the driver looses his/her ability (1) to track lanes, (2) to apply breaks and accelerator adequately, and (3) to apply regular and micro wheel corrections to adjust the vehicle trajectory [84, 145]. As a result, the driving of a drowsy individual tends to be characterized by delayed, sudden brakings, and large wheel corrections. While the above standard measures reflect well the erratic behavior of drowsy driving, they are also heavily influenced by external conditions such as the weather (snowy, rainy, or sunny), road type (curvy or straight), road condition (presence of potholes), and traffic (dense or not).

2.5.3 Indicators based on spontaneous facial expressions

We define spontaneous facial expressions as perceptible patterns of facial muscle contractions (facial expressions) that are not consciously controlled by the individual (hence spontaneous). Like impairments of performance, spontaneous facial expressions are closely related to physiology, and could even be considered as such. Yet, we make the distinction between (1) changes in physiology and (2) changes in facial expressions as changes in facial expressions are non-invasively and visually observable from outside the body, whereas changes in physiology are semi-invasively observable via electrodes in contact with the skin. In the context of drowsiness, the spontaneous facial expressions of interest mostly include fast and slow eye closures; pupil diameter instabilities; yawns; eyebrows rises; and—if we extend our definition to neck muscles—head nods and rolls.

Eye closure

Long and slow spontaneous eye closures are generally considered the indicators of choice to identify drowsiness [4, 40, 41, 84, 122, 136, 144]. A reason for this is that blinks naturally occur once every 2–10 seconds. Meaning that eye closures constitute a regular stream of insights about the physiological impacts of drowsiness, which is a well-suited attribute for basing drowsiness characterization systems upon. The most standard measure of spontaneous eye closure is the percentage of closure (PERCLOS) [40, 41, 145]. The PERCLOS is usually defined as the proportion of time (over a given time window) that the pupils are at least 70% (or 80%) covered by the eyelids. As the level of drowsiness increases, the eye closures become slower and longer, and the upper eyelid droops, all of which contribute to an increase of the PERCLOS. Other reliable, standard measures include mean blink duration [7, 122], mean blink frequency or interval [88, 122], and eye closing and reopening speed [122].

Pupil diameter

The pupil diameter instability has also been linked to drowsiness. Indeed, several studies found that the pupil diameter fluctuates at a low frequency with a high amplitude whenever subjects reported being drowsy [90, 103, 146]. Furthermore, this fluctuation is, most of the time, preceded by a gradual miosis (i.e., a severe constriction of the pupil) [103], during which the subject has yet to be aware of his/her own drowsiness. In other words, gradual miosis is a strong premonitor of drowsiness, and the large, low frequency fluctuation of the pupil diameter is a reliable indicator of drowsiness. Nevertheless, the pupil diameter is highly influenced by external lighting conditions, which makes it inappropriate for outdoor, operational uses. The standard measures of pupil diameter include the average pupil diameter, the variability of pupil diameter, and the power of pupil diameter variation of low frequencies (e.g., below 0.8 Hz [146]).

Yawning

Yawning is as much associated to drowsiness, as it is to boredom and to nervousness. One also yawns when someone nearby yawned recently, hypothetically out of social empathy. To date, yawning is considered to be “the least understood, common human behavior” [60]. The scientific literature offers multiple theories about the purpose of yawns, the most probable being: (1) to increase alertness, (2) to reduce the brain temperature so as to increase its effectiveness, and (3) to show social empathy. Purposes (1–2) are clearly compatible with the need of leaving states of boredom, nervousness, or drowsiness; whereas purpose (3) could explain why one yawns when somebody nearby yawned recently. It is suspected that yawns serve more than one of these physiological functions. In the context of drowsiness, the study of Vural *et al.* [140] suggests that yawns occur less often during the 60s period before a crash. An explanation could be that individuals on the brink of falling asleep lack the energy to even yawn. In other words, yawns appear to be indicators of drowsiness, but negative indicators of high levels of drowsiness. The standard measures of yawning usually consist of the frequency of yawn occurrence.

Eyebrows rising

In the same study, Vural *et al.* [140] showed that the raising of the eyebrows could be indicative of drowsiness. More specifically, they showed that drowsy individuals displayed an increased correlation between eyebrows raising and eye opening. They argue that drowsy individuals fight the urge to sleep by applying supplementary efforts to keep their eyes open, and this by raising the eyebrows. The standard measure would be the correlation between eyebrow positions with eye openness.

Microsleeps

At some point, a severely drowsy individual will uncontrollably experience microsleeps, i.e., brief episodes of sleep. The duration of each microsleep is typically of 0.5–1 second [88]. During this brief period, the individual is unable to assess and respond to his/her surroundings, which may be enough to cause an accident in itself. During a microsleep, the head may drop by either nodding downwards or rolling sideways. Therefore, the variability of head roll [139, 140] and variability of head yaw [94] are considered indicators of microsleeps and, by extension, of drowsiness.

Qualitative evaluation

As can be seen, there exists a wide variety of indicators of drowsiness related to quantitative measures of spontaneous facial expressions. However, having measures that are qualitative (e.g., answering "are the eye closures unusually long?") rather than quantitative (e.g., PERCLOS) may be fundamentally valuable to develop drowsiness characterization systems. To do so, several authors have developed their own scale so that trained experts could produce a score of drowsiness by visually inspecting—in a qualitative way—the spontaneous facial expressions. Wierwille and Ellsworth [144] proposed a linear, gradual scale with verbal descriptions ranging from "not drowsy", "slightly drowsy", "moderately drowsy", "very drowsy", to "extremely drowsy". The trained experts based their scoring decisions on the eye closure dynamics of the subjects. Nopsuwanchai *et al.* [104] proposed a bidimensional scale (Table 2.3). The first dimension of this scale represents the degree of drowsiness (ranging from 1 to 4, corresponding to "high alertness", "slightly low alertness", "very low alertness", and "extremely low alertness", respectively), whereas the second dimension is whether or not the driver applied efforts to counter drowsiness (binary). To do so, the trained experts based their scoring decisions on the facial tone, eye closure dynamics, and countermeasure efforts. Overall, the main concern of such scales is the inter-experts and intra-experts reliability. Wierwille and Ellsworth [144] showed that sufficiently trained experts were found to be consistent within and among themselves.

LoD	**Definition**	**Facial expressions**	**Countermeasure effort**
1	High alertness	Normal facial tone (mouth firmly closed, clear-eye appearance), normal or fast blinks	No effort
2 α	Slightly low alertness	Facial tone is likely to decrease (loosing mouth, upper eyelid slightly falls down), slightly longer blinks period	No effort
2 β	Slightly low alertness with struggling	Facial tone is likely to decrease (loosing mouth, upper eyelid slightly falls down), burst blinks	Moving mouth, rubbing face, moving restlessly in the seat
3 α	Very low alertness	Facial tone decreases (mouth opened, upper eyelid falls down, not properly focusing the eyes), slow blinks	Lack of activity
3 β	Very low alertness with struggling	Facial tone decreases (mouth opened, upper eyelid falls down, not properly focusing the eyes), slow blinks, facial part movements	Head movements, conscious blinking, conscious deep breathing
4	Extremely low arousal	Facial tone absolutely decreased, prolonged eye closure	Lack of activity

Table 2.3 – Bidimensional (1–4 and $\{\alpha, \beta\}$) scale of drowsiness, based on changes in spontaneous facial expressions, adapted from [104].

2.5.4 Indicators based on subjective ratings

The validity and reliability of self-ratings of drowsiness have already been thoroughly discussed in Section 2.3. However, we have yet to introduce the different scales that have been developed to record self-ratings in a standardized manner. The most popular subjective scale is the Karolinska Sleepiness Scale (KSS) [4, 54, 78]. The KSS consists of nine verbally-anchored levels, out of which the respondent has to pick the one that most accurately reflects his/her subjective state of drowsiness. The nine KSS levels and their verbal description, which ranges from "extremely alert" to "very sleepy—fighting sleep", can be found in Table 2.4. Similarly, the Stanford Sleepiness Scale (SSS) [67] is a 7-valued scale (Table 2.5), but it is less popular due to its use of unusual or vague words such as "vital", "foggy", and "woozy". Another subjective scale is the Visual Analogue Scale (VAS) [98]. In drowsiness studies, the VAS usually takes the form of a 100 mm straight line on paper, the ends of which correspond to the extreme limits of alertness and drowsiness, i.e., "very alert" and "very sleepy" [4]. The VAS enables researchers to obtain subjective measures of drowsiness across a continuum of values as the respondent marks their subjective state on the 100 mm line. Differently, the Tiredness Symptom Scale (TSS) [124] asks the subject to indicate how many drowsiness-related symptoms—out of a list of 14—occurred during the performance of a task, and which ones. The proposed symptoms are as follow [95]: (1) heavy head, (2) sore eyes, (3) watering eyes, (4) heavy eyelids, (5) heavy legs, (6) general weakness, (7) feeling cold, (8) sensitivity to noise, (9) yawning, (10) loss of interest, (11) poor concentration, (12) irritability, (13) little desire to speak with others, (14) urge to move around. Upon completion, the TSS score is computed as the total number of confirmed symptoms. A neat advantage is that the TSS enables researchers to obtain separate subjective scores of drowsiness based on either cognitive symptoms (items 8, 10–14) or physical symptoms (items 1–7, 9).

KSS	Level description
1	Extremely alert
2	Very alert
3	Alert
4	Rather alert
5	Neither alert nor sleepy
6	Some signs of sleepiness
7	Sleepy, but no effort to remain awake
8	Sleepy, some effort to stay awake
9	Very sleepy, great effort to stay awake, fighting sleep

Table 2.4 – The Karolinska Sleepiness Scale (KSS), after [4].

2.6 Operational, real-time drowsiness characterization systems

As seen in the previous section, there exists a wide diversity of indicators of drowsiness. In the context of developing an operational drowsiness characterization system, the choice of which indicator to use depends on whether it will be used (1) as an input to the system, or (2) as a ground truth of drowsiness to train the system and evaluate its performance.

When used as an input, the indicator has to be pragmatically and automatically measurable in operational settings, which is the case for driving performance, facial expres-

SSS	Level description
1	Feeling active, vital, alert, or wide awake
2	Functioning at high levels, but not at a peak; able to concentrate
3	Awake, but relaxed; responsive but not full alert
4	Somewhat foggy, let down
5	Foggy; loosing interest in remaining awake; slowed down
6	Sleepy, woozy, fighting sleep; prefer to lie down
7	No longer fighting sleep, sleep onset soon; have dream-like thoughts

Table 2.5 – The Stanford Sleepiness Scale (SSS), adapted from https://web.stanford.edu/∼dement/sss.html.

sions, and eye closure dynamics. However, other indicators are generally inadequate for operational use as they would either hinder the performance of the primary task (e.g., a secondary psychomotor/cognitive task), or be afflicted with a low signal-to-noise ratio (e.g., EEG, EOG, ECG, and pupil diameter).

When used as a ground truth, the scientific community has yet to reach a clear consensus on which indicator is the best [75]. Therefore, the choice of indicator is typically based on its ease of use, and on whether the study protocol enables its acquisition or not.

The below subsections review the state of the art on operational drowsiness characterization systems, the development of which is the main purpose of the present thesis. We focus on two important, yet essentially independent, design aspects: how to produce a ground truth of drowsiness, and what kind of system architecture to use. Furthermore, for conciseness, we limit the scope of this review to drowsiness characterization systems that are based on ocular features. We summarize this state-of-the-art review in Table 2.6, which can be found at the end of this chapter.

2.6.1 Design of ground truth of drowsiness

The scientific literature displays a large variety of approaches to producing a ground truth of drowsiness. In general, a ground truth of drowsiness is produced by mapping an indicator of drowsiness (or, to be more accurate, a measure of this indicator) to a quantized Level of Drowsiness (LoD) taking 2 (a binary LoD) or n (n-valued LoD) distinct integer values. In this way, the task of characterizing drowsiness can be formulated as a classification problem, and be tackled using well-known machine learning models trained in a supervised manner. Note that the ground-truth LoD could be continuous, the task of characterizing drowsiness would then be formulated as a regression problem.

The most straightforward approach to produce a ground truth of drowsiness is by means of a subjective scale, such as those presented in Section 2.5.4. Following this approach, Wang and Xu [141] and Ebrahim *et al.* [45] asked their subjects to self-rate their own LoD in terms of the KSS. To reduce the number of levels, Wang and Xu [141] associated KSS values of 1–6 to “alertness”, a KSS value of 7 to “moderate-level drowsiness”, and KSS values of 8–9 to “high-level drowsiness”. Similarly, Ebrahim *et al.* [45] considered the subjects reporting a KSS value in the range 1–6 as “alert”, and 7–9 as “drowsy”.

Another approach is to take notes of when breakdowns of driving performance occur. Following this approach, Vural *et al.* [139, 140] and Liang *et al.* [87] labeled the drivers as drowsy whenever a line crossing occurred, either in a driving simulator or on straight segments of real roads, respectively. More precisely, Vural *et al.* [139, 140] considered as “drowsy” the minute prior to each line crossing, whereas Liang *et al.* [87] considered as "drowsy" any minute during which a minimum of one line crossing occurred. In all other cases, the subject was considered as "alert". This performance-based approach has the significant advantage of producing an LoD that is both objective, and meaningful for practical use. The annotation of such a ground truth is automatic in a simulator, and fast and simple on real roads.

A family of approaches involves having trained experts score the LoD by visually searching for specific indicators of drowsiness, either in brain signals or in the face video. Along these lines, François *et al.* [51] had one trained expert score the 11-valued LoD according to the KDS, i.e., by looking, in the EEG and EOG signals, for neurophysiological indicators of drowsiness such as alpha rhythm, theta activity, and slow eye movement. Similarly, Liang *et al.* [87] had one trained expert label each minute either as "drowsy" (if there was at least one burst of theta activities lasting longer than 3 seconds), or as "alert" (if there was none). García *et al.* [53] had three psychologist experts subjectively score the binary LoD by a majority vote from the video of the driver's face. Also from the face video, Matsuo and Khiat *et al.* [94] and Nopsuwanchai *et al.* [104] each had three trained experts score, by majority vote, a 6-valued LoD defined on a bidimensional scale crafted by Nopsuwanchai *et al.* [104] (see Section 2.5.3).

One last approach is to establish a list of facial expressions indicative of drowsiness, and then have subjects act them out according to a pre-defined script. In this fashion, Weng *et al.* [143] constructed the Drowsy Driver Detection (DDD) dataset, upon which Shih and Hsu [125] and Huynh *et al.* [72] developed and tested their systems. To enact "drowsiness", the subjects were asked to act out the following facial expressions: (1) yawning, (2) high PERCLOS (i.e., long and frequent eye closures), (3) high PERCLOS then frequent nodding, and (4) a combination of yawning, high PERCLOS, and frequent nodding. To enact "alertness", the subjects were asked to act out the following facial expressions: (1) normal driving (i.e., low PERCLOS), (2) surprised face and bursting out laughing, and (3) combination of talking, laughing, and looking sideways. This approach has the advantage of having the possibility to produce an LoD in both controlled conditions and real-world conditions, but has the significant disadvantage of being based on pre-defined and non-spontaneous facial expressions. Training on non-spontaneous facial expressions is problematic for operational use, as a system trained with such ground truth may not be able to detect subtle, less exaggerated facial expressions that would in fact be indicative of actual drowsiness.

2.6.2 Design of system

Drowsiness characterization systems generally adopt a cascade structure, consisting in (1) extracting an intermediate representation, e.g., a feature vector or a sequence of feature vectors, from the input sequence; and then (2) characterizing drowsiness, as defined by the selected type of ground truth of drowsiness. As stated previously, we focus the present state-of-the-art review on systems that are based on ocular features, i.e., systems that use, at the very least, ocular features as their intermediate representation.

Extraction of the intermediate representation

For most drowsiness characterization systems, the input consists of a sequence of face images. Face images can be acquired remotely, which is convenient and non-intrusive, and they can be processed to extract a wide range of indicators of drowsiness (see Section 2.5.3). On the downside, extracting the semantic information from an image remains a task that is computationally expensive and far from being straightforward.

The systems of Wang and Xu [141] and García *et al.* [53] extract a vector of features related to eye closure dynamics (i.e., ocular features) and driving performance (i.e., driving features). Driving features (i.e., average speed, SDLP, variability of steering wheel, *etc.*) are automatically recorded via either the driving simulator software [141], or multiple in-car sensors in real-roads conditions [53]. The ocular features of Wang and Xu [141] (i.e., PERCLOS, average blink duration and frequency, and average pupil diameter) are automatically measured with the proprietary Smarteye Pro software, whereas those of García *et al.* [53] (i.e., PERCLOS) are extracted (1) by using adaptive image filters, and then (2) by fitting a Gaussian function on the vertical profile of the image variance.

In addition to the PERCLOS, Matsuo and Khiat [94] incorporates the variability of the head center position (i.e., head pose features) and the frequency of subsidiary behaviors into the intermediate representation. Subsidiary behaviors are defined as "behavioral events that are unrelated to and unnecessary for the main task of driving" [94], which include yawning, stifling a yawn, exhaling and breathing deeply, touching one's face (i.e., hand motion), flexing one's neck or one's shoulders, adjusting one's pose, and closing one's eyes for a long duration. Their extraction algorithm involves tracking facial landmarks [147] to produce the ocular and head pose features, and using an array of specialized, data-driven detectors to count the frequency of subsidiary behaviors.

With investigative goals in mind, Vural *et al.* [139, 140] do not restrict themselves to eye closure dynamics, but instead evaluate the relevance of more than 20 facial action units (AUs) codified by the Facial Action Coding System (FACS) [46]. These facial actions include inner and outer brow raising (AU 1 and 2); blinking and eye closing (AU 45); lid tightening (AU 7); jaw dropping (AU 26); nose wrinkling and nostril compressing (AU 9 and 39); cheek raising (AU 6); lip puck, funneling, and pressing (AU 18, 22, and 24); and upper lip raising (AU 10). In both work, the system automatically extracts the facial AUs using algorithms based on Gabor filters, feature selection, and machine learning classifiers; the implementations of which are bundled into the Computer Expression Recognition Toolbox (CERT) [12]. In their work of 2009 [140], they concatenated the mean intensity of each of the 31 AUs into a feature vector as the intermediate representation. They observed that drowsy individuals tend to display a correlation between eyebrows raising and eye opening, as they "raised their eyebrows in an attempt to keep their eyes open". Furthermore, they also found that yawning occurred less often in the 60-s period prior to an accident (i.e., line crossing), suggesting that yawns are a negative indicator of drowsiness in situations of most extreme drowsiness. However, these appealing findings still need to be confirmed via a study at a much larger scale than on 4 subjects. They showed that the most discriminative features are related to eye blink and eye closure, followed by outer brow raise, frown, chin raise, and nose wrinkle. One year later, in 2010, Vural *et al.* [139] incorporated aspects of temporal dynamics by applying 306 temporal Gabor filters on each of 20 AUs, and concatenated their magnitudes, real components, and imaginary components into the intermediate vector. They showed that the most discriminative features were related to eye closing, lip puck, head rolling, nose wrinkling, and lid tightening.

In a further intent to incorporate temporal dynamics, Weng *et al.* [143] devise their intermediate representation to be a sequence of feature vectors, rather than a feature vector. To do so, their extraction algorithm proceeds in two steps. First, they extract low-level spatiotemporal features (related to the eye, mouth, and head pose) based on face landmarks aligned on the face image via the supervised descent method of Xiong and Torre [148]. Then, they generate the sequence of high-level features with three separate deep belief networks (DBNs), each applied at a frame level. These high-level features consist of the probabilities of stillness, nodding, and looking aside for the head pose features;

the probabilities of stillness, laughing/talking, and yawning for the mouth features; and the probabilities of normal and sleepy eyes for the eye features. Having such sequence of feature vectors, rather than a feature vector, has the advantage of enabling the drowsiness model to discover the most discriminative temporal patterns.

Likewise, the system of Nopsuwanchai *et al.* [104] extracts a sequence of normalized histograms of five blink categories, i.e., a sequence of ocular features. The five blink categories differ from one another according to their structural characteristics, such as their amplitudes, their closing durations, and their opening durations. The intermediate representation is obtained in three steps. First, the sequence of normalized eyelids distance is produced by the alignment of eyelid landmarks via an Active Shape Model (ASM) [34]. Second, the blinks are categorized using five hidden Markov models (HMMs), one for each category. Third, the normalized histograms of blink categories are computed within a sliding window.

Shih and Hsu [125] use existing algorithms to extract the intermediate representation. As such, they pre-processed each face image with a pre-trained Convolutional Neural Network (CNN), i.e., VGG-16 [127], so as to produce, as an intermediate representation, a sequence of VGG-16 features. Note that the features of such an intermediate sequence are not easily interpretable.

To further push the concept of discovering patterns in the temporal data, the next step is to map directly the input to the output, i.e., to remove the intermediate representation. In this way, the learning algorithm is not limited to discovering patterns in a sequence of pre-defined features. Instead, the learning algorithm has access to the sequence of raw data, which may enable the discovery of a better intermediate representation than the ones that one would design manually. However, this potential increase in discovery capability comes at the cost of more complexity in the training of the system, since the input dimensionality and the required amount of data grow larger. Adopting such approach, the system of Huynh *et al.* [72] consists in a three-dimensional CNN (3D-CNN), which directly maps the sequence of raw face images into a binary LoD.

In some cases, instead of face images, the vertical EOG is used as input by drowsiness characterization systems. The EOG provides a simpler way of obtaining eye closure dynamics than a sequence of face images. However, interpretation of the EOG may be ambiguous at times since the movements of the eyeball and of the eyelids can manifest in a very similar fashion in the EOG signal. Situations can be found where a downward glance of a few seconds may be misinterpreted as an eye closure of the same duration, even visually by experts. Besides, the EOG requires the wearing of electrodes, which would make EOG-based technologies unlikely to be accepted by the operators.

Ebrahim *et al.* [45] extract, from each blink individually, 8 amplitude-based features and 10 duration-based features by applying thresholds on the derivative of the vertical EOG signal. Amplitude-based features of a blink include its amplitude; its energy; its average and maximum opening and closing speed; and ratios between some of these quantities. Duration-based features of a blink include its frequency; its duration; its closing, closed, and opening durations; PERCLOS (here defined as the ratio between blink duration and closed duration). The vectorized intermediate representation is then produced by averaging these blink features over one minute.

In other cases, the input of drowsiness characterization systems is provided via an instrumented pair of eyeglasses. Even though wearing a pair of eyeglasses may be inconvenient for some people, it is well adapted for applications where the wearer has to be mobile, i.e., able to move freely in his/her environment. Furthermore, this solution has the advantages of producing measures or images of the eye that are stable (the head-mounted sensors perfectly follow the head) and clean (the glass can filter the sun away with a proper coating).

For instance, the system of Liang *et al.* [87] produces, as an intermediate representation, a vector of ocular features (i.e., amplitude-velocity ratio and PERCLOS) and driving features (i.e., SDLP and standard deviation of steering wheel position). While the driving features were measured via automotive sensors, the ocular features were measured via eye reflectometry by the Optalert system [76]. In practice, the Optalert system consists of a pair of eyeglasses equipped with infrared light emitters and transducers. At a frequency of 500 Hz, the emitters send pulses of low-power infrared light over the eye and eyelids. The emitted light is reflected on the complex, multi-layered surface composed of the cornea, iris, sclera, conjunctiva, and skin of the eyelids, each of which has their own reflectance, distance, and orientation with respect to the transducers. The characteristics of this complex surface mostly vary with the eyeball and eyelids movements, making the intensity of the reflected light mostly a function of the eye configuration. As a result, the transducers convert the reflected light into an electrical signal that is mostly sensitive to eyeball and eyelid movements, which allows the characterization of eye closure dynamics.

In a like manner, François *et al.* [51] use the Phasya system to extract a vector of ocular parameters, such as the PERCLOS, mean blink duration, and the percentage of microsleeps (with a microsleep being defined as an eye closure longer than 500ms). In practice, the Phasya system consists of a pair of eyeglasses equipped with an infrared LED, a hot mirror, and a high-speed infrared camera, so as to produce eye images at a framerate of 120 FPS. On the upside, eye images provide more information about the state of the eye than eye reflectometry, such as the pupil diameter and gaze orientation. The latter may be of great help in disambiguating whether the eye is closed or looking down, as both cases manifest themselves with a smaller distance between the eyelids. On the downside, eye images require greater computation power to be processed, which is a challenge on a pair of eyeglasses.

Characterization of drowsiness

The task of characterizing a physiological state such as drowsiness from a set of imperfect indicators is complex and challenging. Therefore, drowsiness characterization systems generally use machine learning models trained in a supervised manner. Of course, the choice of which model to use depends on whether the intermediate representation is a feature vector or a sequence of feature vectors. For processing a feature vector, popular models include Artificial Neural Network (ANN) [45, 53, 94, 141], Support Vector Machine (SVM) [45], and logistic regression [87, 139, 140]. For processing a sequence of feature vectors, popular models include Hidden Markov Model (HMM) [104, 143], and Long Short-Term Memory (LSTM) network [125]. For processing a sequence of raw face images, Huynh *et al.* [72] use a 3D-CNN.

2.6.3 Comparison of performance

The comparison of performance of the reviewed systems requires some caution. Indeed, the various studies differ in many ways, including: the number of subjects, the type of task performed by the subjects, the acquisition settings, the evaluation procedure, the reported performance metric(s), and the ground truth of drowsiness. Furthermore, the unavailability of datasets and of trained models makes fair comparisons infeasible in practice.

Nevertheless, we provide the classification performances of the reviewed systems in Table 2.7, which can be found at the end of this chapter. We observe that some studies (e.g., Garcìa *et al.* [53]) report significantly higher performance than others (e.g., Ebrahim *et al.* [45]). Considering that all the reviewed systems are based on ocular features, these differences in performance are most probably due to the differences in ground truth. Indeed, as we have seen, there exists a wide range of approaches to annotating the ground-truth LoD. In particular, some of these annotated ground-truth LoDs are intrinsically more correlated with the eye closure dynamics than others. For instance, Garcìa *et al.* [53] annotate the ground-truth LoD via three experts visually looking for behavioral signs of drowsiness in the face video, which is more correlated with the eye closure dynamics than ground-truth LoDs annotated from physiology-based indicators or subjective indicators. It is therefore expected that using a ground truth that is strongly correlated with the eye closure dynamics will lead to higher performance than using a ground truth that is not as much.

2.7 Conclusion

Drowsiness is the intermediate physiological state between fully awake and asleep. The continuous and time-varying level of drowsiness is determined by diverse factors including time of day; sleep quantity, quality, and schedule; sleep disorders; task type and length; age; and—most probably—genetics, as supported by the difference in vulnerability to drowsiness across individuals. Generally, educated and healthy individuals are able to reliably assess their own level of drowsiness when it is induced by acute sleep deprivation. However, when the sleep loss turns chronic, due to either sustained sleep restriction or sleep disorders, which are both very commonly experienced in the general population, this self-assessment of drowsiness has been shown to become less reliable. When experiencing drowsiness at the wheel, the best countermeasure is to cease driving and to take a short, 15-min nap. The intake of caffeine (150 mg) is also reliable countermeasure against drowsiness, and can even be combined with a nap. On the opposite, taking a break, turning on the radio, and opening the windows are not reliable countermeasures and only provide—at best—temporary arousal.

The level of drowsiness is not a precisely and numerically defined quantity that can be directly measured. Therefore, the practical approach to estimate the onset and level of drowsiness is by means of measurable, yet imperfect, indicators of drowsiness. There exists a wide range of indicators, including brain activity, spontaneous eye closure dynamics, impairments of performance, and subjective ratings. The most standard measures for brain activity are the α and θ activity powers; the ones for eye closure dynamics are the PERCLOS, mean blink duration and interval, and mean closure and opening speeds; the ones for psychomotor performance are the number of lapses, mean RT, and mean RS; the ones for driving performance are the SDLP and steering wheel variability; the one for subjective ratings is KSS.

In the context of developing automatic, real-time drowsiness characterization systems, the choice of which indicator to use depends on whether it will be used (1) as an input to the system, or (2) as a ground truth of drowsiness to train the system and evaluate its performance. Only a subset of the indicators are well suited to be used as an input (e.g., driving performance, spontaneous facial expressions, and eye closure dynamics) as they require to be pragmatically and automatically measurable in operational settings. The brain activity, subjective ratings, and psychomotor performance are well suited to used as a ground truth, but there is still no clear consensus on which one is best to use. Generally, the ground truth of drowsiness consists of a quantized Level of Drowsiness (LoD) taking 2 or n distinct integer values annotated using various approaches.

Automatic, real-time drowsiness characterization systems usually adopt a cascade structure, which consists in (1) extracting an intermediate representation, e.g., a feature vector or sequence of feature vectors; and then in (2) characterizing drowsiness, using machine learning models trained in a supervised manner. Compared to systems with a single module, systems with a cascade structure have the key properties of having greater interpretability, modularity, and data efficiency. Interpretability facilitates the explanation of the system's decisions, which is of great importance since wrong decisions—although intrinsically unavoidable—should be explainable to humans for (1) the legal and public acceptance of the technology, and for (2) its future improvements, in particular for safety-related applications where human lives are at stake. Modularity enables (online and offline) adaptations to how the intermediate representation is extracted so as to perform better in real-life, operational settings, while being able to keep the characterization of drowsiness as is, i.e., as developed in laboratory settings. Data efficiency enables the system to obtain better performance with an equivalent, limited amount of data.

In the next chapters, we describe three automatic, real-time drowsiness characterization systems. We made the following design choices for our systems:

1. the input must be a sequence of face images;
2. the intermediate representation must be based on the eye closure dynamics;
3. the ground truth of drowsiness must be based on responsiveness (i.e., psychomotor performance) indicators of drowsiness, acquired while performing a PVT.

We made choice (1) so that our systems are well-suited for operational use, choice (2) because eye closure dynamics is the most reliable indicator of drowsiness based on spontaneous facial expression [4, 40, 41, 84, 122, 136, 144], and choice (3) because impairments of performance is a reliable indicator of drowsiness [10, 13, 14, 40, 42, 88] that has the advantages of being objectively, automatically, and densely annotated.

Ground truth of drowsiness is...	Paper	Input sequence	Intermediate representation	Algorithm to produce the intermediate representation	Model to characterize drowsiness
...self-rated by subjects	Wang and Xu [141]	Face images	Ocular and driving features	Smarteye Pro system and driving simulator software	ANN
	Ebrahim *et al.* [45]	EOG	Ocular features	Based on thresholds on the derivative of the EOG signal	SVM/ANN
...marked positive at the occurrence of line crossings	Vural *et al.* [140]	Face images	Facial expressions features	Computer Expression Recognition Toolbox (CERT) [12]	Logistic regression
	Vural *et al.* [139]	Face images	Facial expressions features		Logistic regression
	Liang *et al.* [87]	Eye infrared reflectometry	Ocular and driving features	Optalert system [76] and instrumented vehicle	Logistic regression
...scored by trained experts from brain signals	François *et al.* [51]	Eye images	Ocular features	Phasya system	Phasya system
...scored by trained experts from face video	García *et al.* [53]	Face images	Ocular and driving features	Based on adaptive image filters and instrumented vehicle	ANN
	Matsuo and Khiat [94]	Face images	Ocular, head pose, and subsidiary behavior features	Based on facial landmarks [147] and specialized, data-driven detectors	ANN
	Nopsuwanchai *et al.* [104]	Face images	Sequence of ocular features	Based on facial landmarks (ASM) [34] and HMMs	HMMs
...acted out according to a pre-defined script (non spontaneous)	Weng *et al.* [143]	Face images	Seq. of ocular, mouth, and head pose features	Based on facial landmarks (SDM) [148] and deep belief networks	HMMs
	Shih and Hsu [125]	Face images	VGG-16 features	Pre-trained CNN (VGG-16 [127])	LSTM
	Huynh *et al.* [72]	Face images	Not applicable		3D-CNN
...automatically annotated based on the RTs	This thesis, Chapter 4	Face images	Ocular features	Based on facial landmarks (CLM) [121] and adaptive blink segmentation	SVM
	This thesis, Chapter 5	Face images	Seq. of raw eyelids distances	Based on facial landmarks [81] and CNN	Multi-timescale temporal CNN
	This thesis, Chapter 6	Face images	Seq. of raw eyelids distances	CNN	Temporal CNN

Table 2.6 – Design choices of the drowsiness characterization systems of other studies. For the purpose of comparison, we add the lines related to the systems developed in the present thesis.

Ground truth of drowsiness is...	Paper	Acquisition setting	Nb. of subjects	Data size	Performance metric(s)	Results
...self-rated by subjects	Wang and Xu [141]	Driving simulator	16	16h	Average recall (3 classes)	56.04%
	Ebrahim *et al.* [45]	Real road and driving simulator	43	67h	TNR; TPR (w/ SVM)	76.6%; 64%
					TNR; TPR (w/ ANN)	75.9%; 65.2%
...marked positive at the occurrence of line crossings	Vural *et al.* [140]	Driving simulator	4	12h	Area Under the ROC Curve (AUC)	0.98
	Vural *et al.* [139]	Driving simulator	11	< 44h	AUC	0.96
...scored by trained experts from brain signals	Liang *et al.* [87]	Real road	16	36h	TNR; TPR; AUC (subject-specific, not -generic)	98%; 67%; 0.92
	François *et al.* [51]	PVT	24	12h	TNR; TPR	80%; 72%
...scored by trained experts from face video	García *et al.* [53]	Real road	10	30h	TNR; TPR	95.8%; 85.5%
	Matsuo and Khiat [94]	Driving simulator	20	20h	Average recall (4 classes)	50.43%
	Nopsuwanchai *et al.* [104]	Driving simulator	13	11h	Visual (4 classes)	-
...acted out according to a pre-defined script (non spontaneous)	Weng *et al.* [143]	Driving simulator	36	9h30	F1 score; Accuracy	85.39%; 84.82%
	Shih and Hsu [125]	Driving simulator	36	9h30	F1 score; Accuracy	82.82%; 82.61%
	Huynh *et al.* [72]	Driving simulator	36	9h30	F1 score; Accuracy	87.97%; 87.46%
...automatically annotated based on the RTs	This thesis, Chapter 4	PVT	14	6h	TNR; TPR; Accuracy	87.73%; 75.46%; 86.19%
	This thesis, Chapter 5	PVT	29	13h40	TNR; TPR; Accuracy	94.80%; 74.19%; 94.22%
	This thesis, Chapter 6	PVT	29	13h40	TNR; TPR; Accuracy; AUC	95.2%; 73.8%; 94.5%; 0.74

Table 2.7 – Classification performances of the systems of other studies. For systems using binary classification, the negative class corresponds to the "alert" label, and the positive class to the "drowsy" label. TNR stands for true negative rate, and TPR for true positive rate. For the purpose of comparison, we add the lines related to the systems developed in the present thesis.

Chapter 3

Sleep-deprivation dataset

This chapter describes the sleep-deprivation dataset that we collected for the purpose of developing automatic, real-time drowsiness characterization systems based on face images. Section 3.1 motivates our sleep-deprivation dataset, and discusses the design choices of the study protocol. Section 3.2 details the acquisition of data. Section 3.3 informs about the availability of our dataset to the research community. Section 3.4 discusses the ecological validity of our dataset, i.e., the extent to which the conclusions and findings drawn from our laboratory-acquired dataset can be generalized to real-life, operational settings. Section 3.5 highlights some limitations of our dataset. Section 3.6 analyzes the statistical distribution of drowsiness data in our dataset, i.e., the three recorded ground truths based on physiology, performance impairments, and subjective ratings. Section 3.7 concludes this chapter.

3.1 Motivation and design choices

The main goal of this thesis is the development of drowsiness characterization systems based on face images. To achieve this goal, one of the most crucial components is a well-designed dataset, which must—at the very least—contain the following ingredients:

1. the input that will be fed to the system in operational settings, i.e., the video of the face;

2. the output that the system should learn to produce, i.e., the ground truth of drowsiness derived from some indicator(s) of drowsiness.

Ideally, the dataset is acquired in conditions as close as possible to real-life, operational settings, meaning in a real car, on real roads, and with real, spontaneous drowsiness. However, in practice, several considerations make this difficult to achieve. First, as seen in the previous chapter, several indicators of drowsiness are not easily measurable in real conditions, such as brain signals and psychomotor performance. Second, inserting drowsy drivers into the road traffic is dangerous, unethical, and would require strong security precautions (e.g., an alert copilot ready to take the commands of the vehicle at any time). Acquiring the dataset in laboratory, controlled conditions however eliminates these inconveniences, and opens more possibilities to the development of drowsiness characterization systems. In practice, multiple design choices still have to be made in order to produce a sleep-deprivation dataset.

3.1.1 Type of ground truth of drowsiness

Some design choices relate to the ground truth of drowsiness: which indicator(s) of drowsiness we should acquire, and which task the participants should perform. In our case, we decided to gather as many indicators as possible, so as to gain as much freedom as possible in our research and development. We decided to record three indicators of drowsiness: an objective physiology-based indicator, an objective performance-based indicator, and a subjective indicator. The tasks of (1) performing a Psychomotor Vigilance Task (PVT) and (2) driving in a driving simulator can both be performed in adequate conditions for measuring such indicators. However, our choice leaned towards the PVT because its associated responsiveness performance metric is more standard, and more practical to work with than driving performance metrics. Moreover, the PVT's short duration of 10 minutes substantially simplifies the elaboration of the study protocol. The physiology-based indicator consists of the polysomnography (PSG) signals, i.e., the electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), and electrocardiography (ECG) signals, acquired during the PVT via electrodes in contact with the skin. The performance-based indicator consists of the reaction times (RTs) to visual stimuli displayed during the PVT. The subjective indicator is obtained by asking the subjects to self-assess their own drowsiness via the Karolinska Sleepiness Scale (KSS) prior to each PVT. It is important to note that the physiology-based and performance-based indicators were recorded in a perfectly time-synchronized manner with the face images during the PVT.

3.1.2 Inducement of drowsiness

Some design choices relate to how much drowsiness to induce, and in which manner. In our case, we aim to induce levels of spontaneous drowsiness in a range as wide as possible. In this way, the dataset contains realistic data with states ranging from full alertness, moderate drowsiness, to extreme drowsiness. To this end, we chose to induce drowsiness via acute sleep deprivation rather than accumulated sleep restriction. Multiple reasons justify this decision. First, acute sleep deprivation enables high levels of drowsiness to be reached quickly. Second, accumulated sleep restriction would require to be sustained for long periods, which would be difficult to enforce and far more burdensome for the subjects. Third, it has been shown that subjects self-rate their drowsiness more reliably when acutely sleep deprived than when chronically sleep restricted [43]. On top of the sleep deprivation, several factors may help in further reaching high levels of drowsiness, such as performing a monotonous task in the dark, in a quiet and isolated room, and at specific times of the day. The latter is particularly important given that the circadian and homeostatic processes regulate together the most adequate times for sleep, which correspond to the most vulnerable times for drowsiness. Therefore, we carefully selected the right times of day during which the PVT should be performed at so as to favor data of drowsiness.

3.1.3 Choice of participants

Some design choices concern the inclusion and exclusion criteria for the subjects participating in the study. In our case, we wanted a dataset composed of healthy subjects, meaning subjects without any sleep disorders, drugs addictions, or alcohol dependencies. The reasons are diverse. First, we wanted to dissociate the drowsiness induced by acute sleep deprivation from the one induced by these factors. In this way, we could study the drowsiness dynamics of healthy individuals before studying the one of patients. Studies incorporating patients suffering from sleep disorders would be of major importance since sleep disorders affect a significant part of the general population. Second, we wanted to obtain baseline data of alertness, so as to (1) have a wide range of levels of drowsiness and (2) be able to normalize inputs and ground truths considering that many indicators of drowsiness have significant inter-subject variability (see previous chapter). Patients with sleep disorders would not display such baseline data of alertness as they typically experience basic levels of drowsiness higher than those of healthy individuals. Finally, we had no exclusion criterion based on the age of the participants. However, it turned out that only young individuals (mostly students from our university) were interested in participating in our study.

3.1.4 Type of face images

Some design choices concern what type(s) of face images to record as inputs, i.e., color images, near-infrared intensity images, and/or range images, and with which camera. In our case, we needed types of face images that could be acquired in a lighted room as well as in total darkness, and in accordance with our study protocol and future operational use. Furthermore, we wanted—at the beginning—to explore the benefits of using 3D data in the context of facial expressions analysis. Taking all these requirements into account, we chose to acquire near-infrared intensity and range images directly using a 3D range camera, the best at the time in terms of resolution and image quality being the Microsoft Kinect v2 sensor.

3.2 Data acquisition

3.2.1 Study protocol

Thirty-five young, healthy subjects (14 males, 21 females), aged 23.3 ± 3.6 years (mean ± standard deviation), participated in our study that extended from November 2014 to June 2015. Our study protocol—approved by the Ethics Committee of the University of Liège—led each subject to perform three 10-minutes PVTs over two consecutive days, under conditions of increasing sleep deprivation conditions induced by acute, prolonged waking.

The protocol required subjects without any alcohol dependencies, drug addictions, or sleep disorders. Each was asked to maintain a normal sleep pattern for the week prior to taking the first PVT, and to have a full night sleep (of 7–8 hours at least) just before this PVT. Each was also asked to maintain a sleep diary during that week, to allow us to verify that the sleep requirements were met. Once a subject took the first PVT, he/she was not allowed to sleep until after the third and last PVT, thereby inducing a total sleep deprivation of 28–30 hours. We organized several sessions of three PVTs, each with a few subjects (typically 2–3 subjects) successively taking each PVT.

The details of the tests follow. **Day 1.** At 8:30 (in 24 hour time), the scheduled subjects arrived at the laboratory, and were equipped with the PSG electrodes. Between 10:00 and 11:00, they (successively) carried out the first PVT, called PVT1. Afterwards, they were equipped with wrist actigraphs to verify that they would not sleep, and were allowed to leave the laboratory. From 12:00 on, they were not allowed to consume any coffee, tea, energy drinks, or other stimulants. At 20:30, they returned to the laboratory, and were equipped with the PSG electrodes. They stayed overnight in the laboratory, and until the end of the tests. During the night, they were allowed to use multimedia devices, to play card and board games, to interact with the laboratory staff, and to consume the soft drinks and biscuits that we provided. **Day 2.** Between 3:30 and 4:00, they carried out the second PVT, called PVT2. At 8:30, we provided breakfast. Between 12:00 and 12:30, they

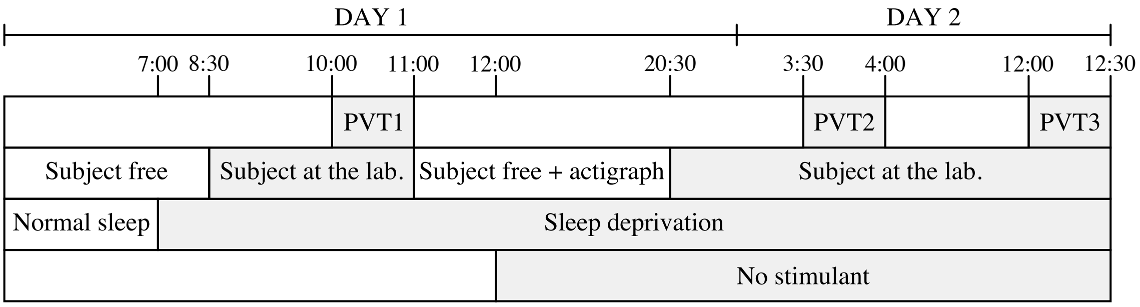


Figure 3.1 – Pictorial summary of the data acquisition schedule.

carried out the third PVT, called PVT3. This concluded the tests. We strongly advised the participants not to drive home by themselves, and we offered alternative transportation solutions when necessary. Figure 3.1 depicts the data acquisition schedule for each subject.

The PVTs were all performed in a quiet, isolated laboratory environment without any temporal cues (e.g., watch or smartphone). The room lights were turned off for PVT2 and PVT3. For a 15-minute period before each PVT, we instructed the subjects to part with their phones, computers, and any other screen devices. At the beginning of each PVT, we asked the participant to self-estimate their own level of drowsiness in terms of the KSS. During each PVT, we also recorded the PSG signals, the RTs (in milliseconds), and the face images of the subject, all in a perfectly time-synchronized manner. All data were collected anonymously.

3.2.2 Psychomotor Vigilance Task (PVT)

The PVT has become one of the most widely used tools to measure performance impairments induced by drowsiness. Multiple studies have shown its validity, reliability, and extreme sensitivity to sleep deprivation [13, 44], and by extension to drowsiness. The PVT gives the RTs to visual or auditory stimuli that occur at random inter-stimulus interval. Compared to other tests, the PVT has the advantage of being almost independent of aptitude (low inter-subject variability) and learning (high intra-subject reproducibility) [44]. In our study, we implemented our own version of the 10-minute PVT, adapted from the one proposed by Basner and Dinges [13]. The subjects were instructed to monitor a red rectangular box over a black background on a computer screen, and to press a physical, response button as soon as they noticed the appearance within the box of a yellow stimulus counter (expressed in milliseconds). When the button was pressed, the counter stopped and the achieved RT remained displayed for 1 second. RTs below 100 milliseconds were discarded as false starts (errors of commission). After 30 seconds without any response, the counter timed out and displayed a yellow "overrun" message inside the box for a few seconds. The inter-stimulus interval, defined as the time interval between the last response and the appearance of the next stimulus, was varied randomly between 2 and 10 seconds. Furthermore, each time the achieved RT was stopped being displayed, the red box position was randomly varied among five positions on the computer screen, i.e., at its center and at its four corners. In this way, the face images contain more variability in head pose and eye gaze direction, in a similar manner to what would be found in real-life settings.

3.2.3 Polysomnography (PSG) signals

The PSG signals, i.e., the EEG, EOG, ECG, and EMG, are regarded as the "Gold Standard" to study sleep and, in particular, to score sleep stages [112]. As seen in the previous chapter, such PSG signals are also useful to characterize drowsiness since the activity in the alpha (8–12 Hz) and theta (4–8 Hz) bands in the EEG signal, and the slow eye movements in the EOG signal are strong, reliable indicators of drowsiness when performing a task [4, 54]. In our study, we recorded the following channels via electrodes and the portable, laboratory Embla Titanium system, all sampled at a frequency of 512 Hz:

- EEG: Fz, Pz, Cz, C3, and C4 channels, all referenced to the A1 channel, via electrodes positioned on the scalp following the international 10–20 system [65];
- EOG: two channels for the vertical EOG, via electrodes positioned above and below the right eye; and two channels for the horizontal EOG, via electrodes positioned at the right and the left of the eyes;
- EMC: two channels, via electrodes positioned below the chin;
- ECG: two channels, via electrodes positioned on the chest;
- PGND: one ground channel, used for common mode rejection, via an electrode positioned on the scalp.

3.2.4 Face images

The face images were acquired with the Microsoft Kinect v2 sensor, which provides—for each video frame—a color image, and a pair of aligned, near-infrared intensity and range images. Since drowsiness characterization systems must generally operate in all lighting conditions, including in total darkness, we only retained the intensity and range images, which are both related to active near-infrared illumination. The near-infrared intensity and range images are of size 512×424 pixels, have 16-bit values, and are recorded at 30 FPS. Note that a framerate of 30 FPS corresponds to a temporal resolution of $\sim$ 33ms, which is sufficient as (1) the duration of a blink is on average greater than 100ms and (2) drowsiness is characterized by long blinks. The camera was positioned just below the computer screen used for the PVTs, at a distance of about 0.7 m from the subject. Figure 3.2 shows pairs of example of near-infrared intensity and range images.

3.2.5 Loss of data

Due to some technical issues, only 88 PVTs (out of 105) from 32 subjects (12 males, 20 females) turned out to be usable. In particular, the PVT1 data was lost for subjects 9, 11, 31, and 32, and never occurred for subjects 7 and 24; the PVT2 data was lost for subjects 9, 12, 14, 31, and 32; and the PVT3 data was lost for subjects 9, 14, 15, 16, 31, and 32.

3.3 Availability to the research community

We made available to the research community the data for a subset of 14 subjects via the “ULg Multimodality Drowsiness Database” [91], also called “DROZY”. The DROZY dataset contains all the modalities acquired during the study, i.e., the intensity and range images from the Kinect v2 sensor, the KSS scores, the PSG signals, and the PVT data, as well as 2D (in pixels) and 3D (in millimeters) annotations of 68 face landmarks (1) for 720 hand-selected frames via manual annotations and (2) for all frames via automatic annotations using subject-specific constrained local models [121]. Instructions for obtaining the DROZY dataset can be found on the associated website (http://www.drozy.ulg.ac.be).

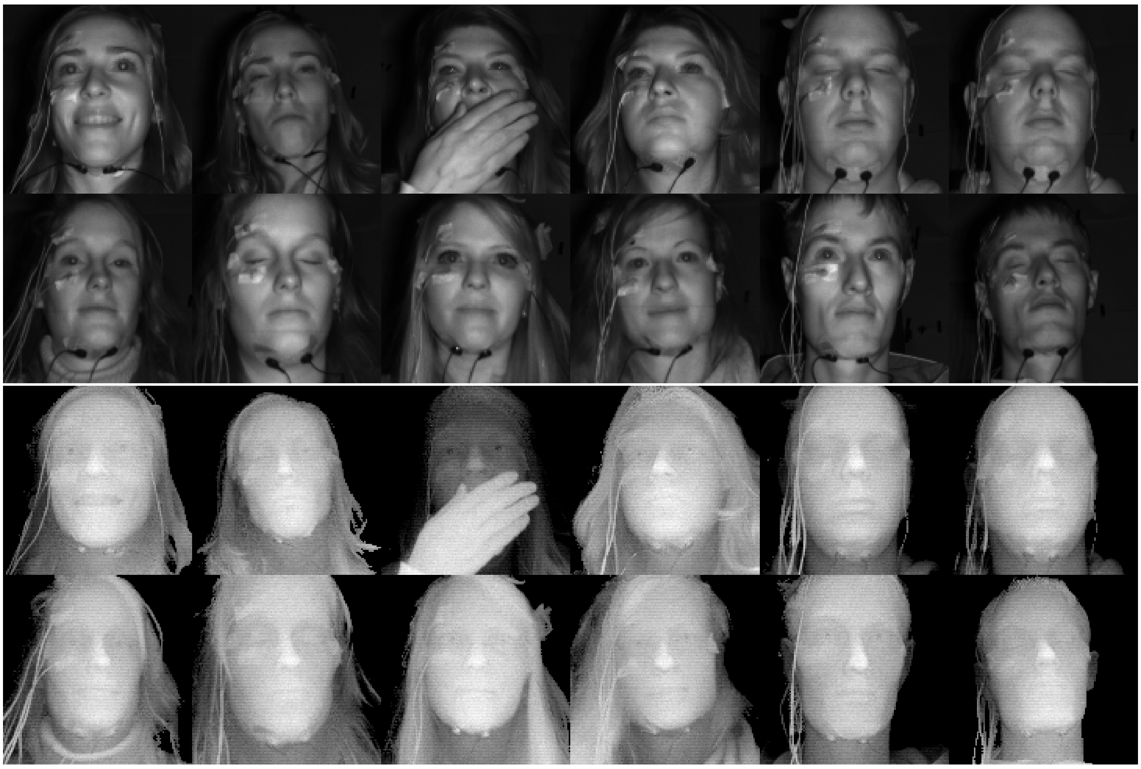


Figure 3.2 – Examples of cropped, near-infrared, intensity (top) and range (bottom) images of the face obtained with the Kinect v2 sensor. For visualization purposes, the range images were monochromatically colorized with colors ranging from white (for the closest point) to black (for points at least 18cm farther than the closest point).

In addition, we made available to the research community a subset of the data for all 32 subjects alongside the Massoz *et al.* [93] article. This data contains only two components: (1) the sequences of eyelids distances produced by the presented system, and (2) the PVT data. Links to this data can be found on the associated website (http://www.telecom.ulg.ac.be/mts-drowsiness).

3.4 Ecological validity

As stated previously, while the dataset should ideally be acquired in real-life, operational settings, it is instead acquired—for practical reasons—in controlled, laboratory settings. Therefore, an important topic of discussion is about the ecological validity of such laboratory dataset, i.e., the extent to which the conclusions and findings drawn from such laboratory dataset can be generalized to real-life, operational settings. Evaluating this ecological validity is far from being straightforward, and is really tackled by very few publications [61, 108]. We would recommend that the scientific community conduct further research so as to be able to evaluate this validity more thoroughly. Nevertheless, here is a list of key points, in the form of questions and first answers, so as to feed this discussion.

- *Are the RTs measured in a PVT indicative of real-life performance?* Yes, the RTs recorded in laboratory conditions are considered as valid and meaningful (though not absolute) measures of real-life performance [44, 107, 108]. Indeed, the PVT requires a sustained attention and quick responses to sudden events, just like driving and many other tasks in the real world.

- *Do performance impairments induced by (acute) sleep deprivation generalize to ones induced by other type of sleep loss, such as (accumulated) sleep restriction?* Yes, Van Dongen *et al.* [43] showed that the number of lapses are near-linearly related to the "cumulative duration of wakefulness in excess", regardless of the type of sleep loss (i.e., sleep deprived or sleep restricted). This is an important point since the general population probably experiences drowsiness induced by sleep restriction more regularly than by sleep deprivation, which is the type of sleep loss depicted in this dataset. However, the generalization to drowsiness induced by sleep disorders, most of which degrade sleep quality rather than sleep quantity, remains an open question.

- *What about the other indicators, i.e., the subjective and physiology-based ones, do they also generalize to real-life settings?* To some extent, yes. Hallvig *et al.* [61] compared such indicators (in terms of KSS and KDS, respectively) during both simulated driving and real driving, both during the day and during the night. Results show that higher KSS and KDS scores were reached in the simulator, which suggests low absolute validity. For instance, at the same time of day, KSS scores were about two units higher in the simulator. A possible explanation would be the soporific aspect of the simulator given the lack of danger, and the lower level of stimulation (e.g., traffic, lights, speed limits). However, it could also be that simulated driving manifests the latent, underlying, level of drowsiness better than real driving, as real driving is likely to mask latent drowsiness [117]. Reassuringly, results also indicate a good relative validity at night, meaning that both indicators of drowsiness, i.e., KSS and KDS, showed a similar response pattern at night both during simulated and real driving. We found no study comparing these indicators acquired during both a PVT (which is a monotonous but stimulating task) and real driving.

- *Given that the present thesis focuses on the analysis of eye closure dynamics, would eye closure dynamics be different if it was measured in operational settings?* To some extent, yes. Compared to performing a PVT in front of a computer, driving a car requires that the driver look at diverse elements (such as the road, the dashboard, and the rear-view mirrors), and in doing so modifies the eyelids configuration (e.g., looking down below the optical axis of a camera naturally brings the eyelids closer). This is also one of the reasons why we randomly moved the stimulus box of the PVT to various positions across the computer screen. Moreover, external conditions such as bright sunlight may cause the driver to squint, i.e., to close slightly their eyes in an attempt to see more clearly, which also modifies the eyelids configuration. Yet, the eye closure dynamics that are most indicative of drowsiness (i.e., slower and longer eye closures) are still found to be valid in real-life settings [45, 87], although the average eye closure duration during real driving appears to be shorter than during simulated driving [61]. In the meantime, the development of specialized detectors may help in disambiguating whether the driver closes his/her eyes because of drowsiness, bright sunlights, or downward glances, and thus improve the measurement of the eye closure dynamics of interest.

- *Are the recorded faces images comparable to the ones recorded in real-life settings?* No, there are some major differences in the properties of the face images, such as the variability in head poses and in illumination conditions, that motivate the need of further algorithm developments to process face images acquired "in the wild", i.e., in real-life, operational settings. Indeed, the recorded face images consist mostly of frontal and near frontal faces, acquired in a laboratory room that is either well lit or in total darkness, which does not cover the full range of head poses (e.g., head turned while checking a blind spot) and illumination conditions (e.g., face partly illuminated by the sunlight) occurring during naturalistic driving. Furthermore, as seen in Figure 3.2, the PSG electrodes are visible as artifacts on the recorded face images. However, experimental results indicate that these artifacts do not interfere with the processing of face images, considering that even off-the-shelf algorithms handle them well, as if these artifacts were absent.

Overall, this discussion highlighted some key considerations to keep in mind when developing drowsiness characterization systems from a dataset acquired in laboratory settings, and intended to be adapted for real-life, operational settings. The main conclusions concerning the ecological validity of laboratory datasets are as follow.

1. The indicators of drowsiness based on psychomotor performance, physiology, and subjective ratings appear to have high—though not absolute—ecological validity in laboratory settings. This suggests the need for some kind of post-development calibration of the characterized drowsiness to fit operational settings.

2. The eye closure dynamics appears to have high ecological validity in laboratory settings, but the processing of face images does not. This suggests that a good system architecture would be to isolate the processing of face images from the characterization of drowsiness based on eye closure dynamics, similarly to the cascade structure presented in the previous chapter. In this way, the processing of face images could be modified to perform better in operational settings while the other parts of the system could be kept as they are.

3.5 Limitations and potential improvements

Our sleep-deprivation dataset has some limitations that mostly concern the limited representativity in subjects, and the small amount of data.

Indeed, the subjects of our study were relatively young, i.e., with ages in the range of 19–34 years. Given the fact that the vulnerability to drowsiness varies with age [7, 69], there is a clear interest in incorporating older subjects in such study. Certainly, the more the general population is represented in the dataset in terms of facial expressions, of ethnicities, and of drowsiness dynamics, the better the developed systems will generalize to operational settings. Following the same idea, there is also a clear interest in incorporating patients suffering from sleep disorders (i.e., sleep patients), as they represent a significant part of the general population. However, our study protocol involves an inconvenient, acute sleep deprivation of 28–30 hours, which is definitely not appealing to older individuals and sleep patients.

Furthermore, the size of our dataset can be considered relatively small in terms of (1) the number of subjects (35), as well as of (2) the amount of data per subject (3 PVTs with durations of 10 minutes). Whereas the number of subjects could have been improved by prolonging the study, the amount of data per subject could have been improved by increasing the number of PVTs performed over the two consecutive days. However, the subjects were actually participating in three studies conducted at the same time: our study, another PVT study, and a stereoscopic-3D driving simulator study; the schedules of which were intertwined and spread over the same two days. As a result, increasing the number of PVTs in our study has been somewhat challenging.

3.6 Analysis of drowsiness data

Before blindly jumping into the development of drowsiness characterization systems, it is important to take the time to inspect and analyze the drowsiness data contained in our sleep-deprivation dataset. To this end, we plot and analyze the statistical distributions of the following standard measures of drowsiness:

- the Karolinska Drowsiness Scale (KDS) score, visually annotated, from the PSG signals, by one trained expert;
- the Karolinska Sleepiness Scale (KSS) score, directly annotated by the subject before each PVT;
- two performance measures, i.e., the mean RT and the percentage of lapses, computed from the raw RTs over 1-min time windows.

3.6.1 Statistical distribution of KDS scores

To annotate the KDS scores, the trained expert visually looks for signs of drowsiness (i.e., alpha rhythms, theta activities, and slow eye movements) within every successive, non-overlapping, and butting 2-s segments of the EEG and EOG signals. The KDS score is then computed, for each 20-s epoch, as the number of 2-s segments (within the epoch) containing at least one sign of drowsiness. The KDS score ranges thus from 0 to 10. Note that an epoch with a KDS score of 5 or above is generally associated with drowsiness [55]. This results in 30 KDS scores per 10-min PVT. Because of the difficulty and time-consuming aspects of this annotation, the KDS score was annotated only for subjects 1–4, resulting in a total of 12 annotated PVTs.

Table 3.1 contains the annotated KDS scores, and Figure 3.3 shows the box plot of the KDS score as a function of the PVT index. As expected, we observe, in the table and in the figure, that the physiological drowsiness score increases with the duration of sustained waking, which increases from PVT1 to PVT3. Furthermore, even with the small number of annotated subjects, we observe a significant variability in the vulnerability of subjects to drowsiness. For instance, subject #4 appears quite vulnerable given his/her relatively high minimum KDS score (i.e., mean KDS of 1.8 during PVT1), and his/her significant relative increase in drowsiness when sleep deprived (i.e., mean KDS of 2.4 and 6.1 for PVT2 and PVT3, respectively). On the opposite, subjects #2 and #3 appear less vulnerable given their low basal KDS scores, and their moderate increases in drowsiness when sleep deprived.

		Epoch index																														Mean
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**KDS**
S1	**P1**	0	0	0	1	1	0	1	0	1	0	0	2	0	1	0	0	0	0	2	0	1	0	2	0	0	1	1	0	1	0	0.5
	P2	1	0	0	3	3	4	3	3	**6**	2	3	3	3	4	1	4	2	3	4	3	3	3	3	**5**	3	4	4	3	**5**	3	3
	P3	1	4	3	**5**	4	**5**	3	2	4	4	3	3	4	1	2	4	3	3	2	4	4	2	3	2	2	**6**	**5**	**5**	3	3	3.3
S2	**P1**	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	2	0	1	0	0	0	0	0	0	0	0.3
	P2	0	0	0	1	1	2	0	1	1	1	2	0	0	4	1	1	1	3	1	2	1	3	2	0	0	0	0	1	1	2	1
	P3	1	2	1	2	0	3	1	2	2	3	3	2	2	2	2	1	1	2	3	1	2	1	0	1	2	2	2	1	1	1	1.6
S3	**P1**	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0	0	0.4
	P2	0	0	1	0	1	0	1	0	0	0	1	2	0	1	0	0	1	0	0	1	2	1	3	0	0	0	1	0	0	2	0.6
	P3	2	0	1	1	2	1	1	3	1	1	1	1	2	1	0	2	2	1	1	1	2	0	3	0	1	0	1	2	1	0	1.2
S4	**P1**	2	2	3	4	4	1	1	0	1	3	0	2	2	1	1	3	1	0	1	4	3	3	0	1	2	1	2	1	2	3	1.8
	P2	-	-	-	-	-	-	-	-	-	-	-	-	3	**5**	**5**	**5**	**5**	**8**	**7**	**10**	**8**	2	2	1	2	2	**5**	**7**	**6**	1	2.4
	P3	4	4	4	**6**	**6**	**5**	**7**	3	**7**	**7**	**8**	**10**	**5**	**6**	**7**	**6**	**8**	**6**	4	**7**	**5**	**9**	**5**	**5**	**6**	**8**	**6**	4	**8**	**8**	6.1

Table 3.1 – KDS score as a function of the subject index (Sx), the PVT index (Px), and the 20-s epoch index (1–30). KDS scores greater than, or equal to, 5 (generally associated with drowsiness) are highlighted in red. The value "-" indicates that scoring was impossible at the corresponding epoch because of high amplitude noise. We observe (1) the inter-subject difference in vulnerability to drowsiness, and (2) the effects of sustained waking duration on the physiology-based drowsiness score.

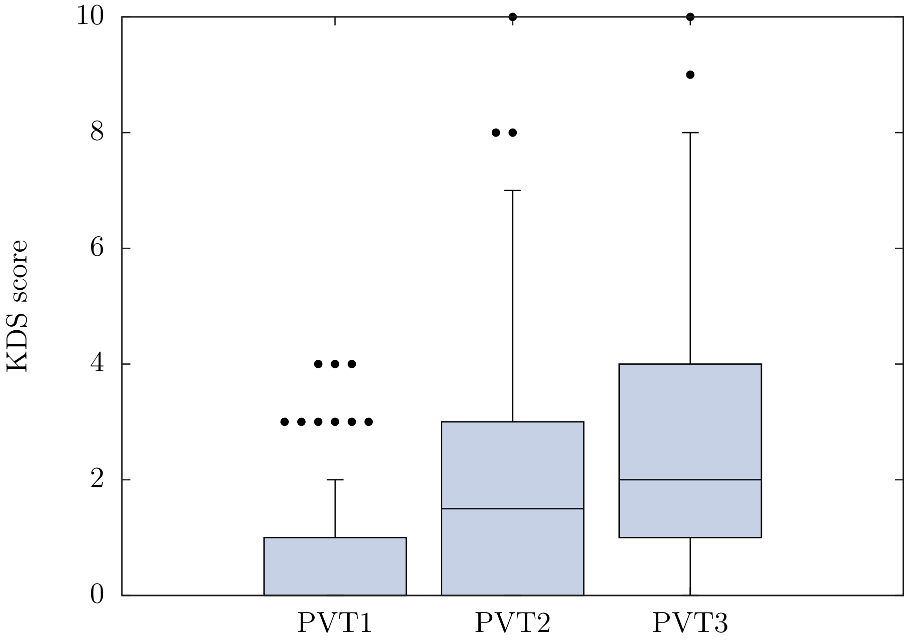


Figure 3.3 – Box plot of the KDS score as a function of the PVT index, over subjects 1–4. Note that the median KDS score for PVT1 is 0. We observe that objective drowsiness increases with increasingly sustained waking, i.e., from PVT1 to PVT3.

What is and how to read a box plot? The box plot is a standardized way of displaying the distribution of the data. The bottom and top of the box correspond to the first and third quartiles of the data, respectively. The bottom (resp. top) whisker corresponds to the lowest (resp. highest) datum still within 1.5 interquartile range (IQR) of the first (resp. third) quartile. The line inside the box represents the median of the data, whereas the dots represent outliers.

3.6.2 Statistical distribution of KSS scores

The analyses and observations based on the KSS scores are similar to those based on the KDS scores. Table 3.2 contains the self-annotated KSS scores (before each PVT), and Figure 3.4 displays the box plot of the KSS score as a function of the PVT index, i.e., 1, 2, and 3. We observe a difference in vulnerability to drowsiness across subjects, as well as an increase in the subjective drowsiness score with the PVT index, i.e., the duration of sustained waking.

Subject index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
PVT1	3	3	2	4	3	2	-	2	2	4	2	3	4	2	2	6	2	2
PVT2	6	7	3	8	7	3	4	6	7	3	6	6	7	5	3	3	5	5
PVT3	7	6	4	9	8	7	9	8	9	8	8	7	7	6	4	7	5	8
Subject index	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	
PVT1	3	3	5	2	1	-	3	2	2	2	2	3	4	2	1	1	2	
PVT2	6	5	7	6	4	6	5	4	6	5	5	3	6	7	4	5	6	
PVT3	3	4	8	7	6	6	4	4	7	8	7	5	8	3	4	6	5	

Table 3.2 – KSS score as a function of the subject and PVT indices. KSS scores greater than, or equal to, 7 (associated with drowsiness) are highlighted in red. The value "-" indicates that the corresponding subject did not perform the first PVT, and thus did not self-evaluate his/her KSS score. We observe (1) the inter-subject difference in vulnerability to drowsiness, and (2) the effects of sustained waking duration on the subjective drowsiness score.

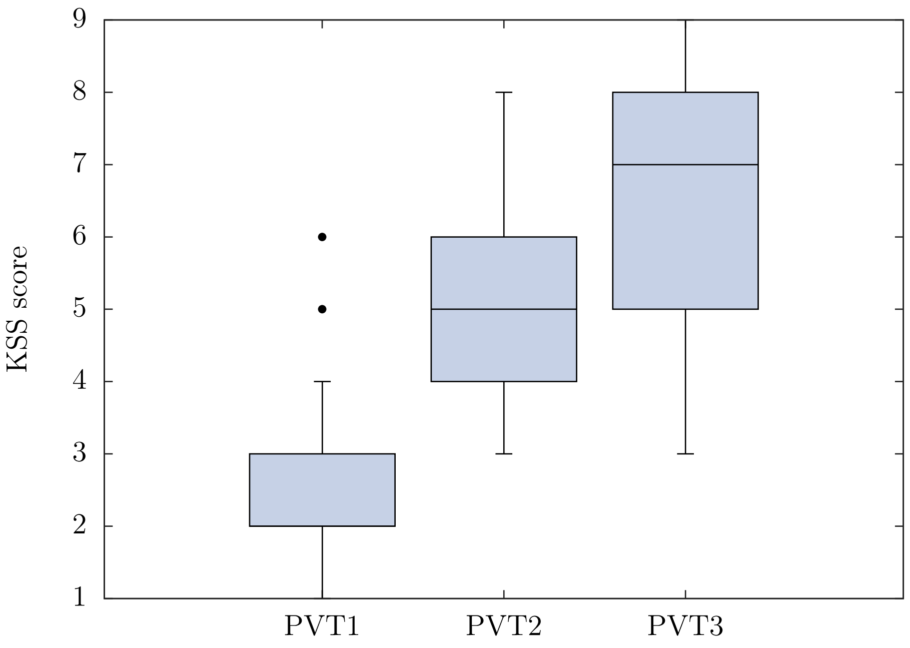


Figure 3.4 – Box plot of the KSS score as a function of the PVT index, over all the subjects. Note that the median KSS score for PVT1 is 2. We observe that subjective drowsiness increases with increasingly sustained waking, i.e., from PVT1 to PVT3.

3.6.3 Comparison between KDS and KSS scores

It is interesting to compare, for each PVT, (1) the KSS scores and (2) the mean KDS scores. To this end, Table 3.3 compiles the KSS and KDS scores for subjects #1–4, i.e., the subjects with both KSS and KDS annotations. We observe that subject #2 reported high KSS scores at the beginning of PVT2 and PVT3, even though the corresponding mean KDS scores were within the range of the alertness ones, i.e., low. However, we observe that subjects #1, #2, and #4 self-estimate correctly their physiological state of drowsiness, as evidenced by the Spearman's rank correlation coefficient of 0.94 between KSS and KDS for these three subjects.

Subject index		**1**	**2**	**3**	**4**
KSS	**PVT1**	3	3	2	4
	PVT2	6	**7**	3	**8**
	PVT3	**7**	6	4	**9**
Mean KDS	**PVT1**	0.5	0.3	0.4	1.8
	PVT2	3	1	0.6	2.4
	PVT3	3.3	1.6	1.2	6.1

Table 3.3 – Comparison between the self-annotated KSS scores, and the expert-annotated mean KDS scores, for the subjects 1–4. We observe that subject #2 reported high KSS scores at the beginning of PVT2 and PVT3, even though the corresponding mean KDS scores were within the range of alertness ones. However, subjects #1, #2 and #4 appear to self-estimate correctly their physiological drowsiness.

This disparity for subject #2 may hint at one, or both, of the following possibilities: (1) not everyone can reliably self-estimate their physiological state of drowsiness (implying that KSS is sometimes unreliable), and/or (2) the scoring criteria of the KDS are not suited for everyone (implying that KDS is sometimes unreliable). The latter is probable given that the expert only found theta activities and slow eye movements, i.e., no alpha rhythms, in the PSG signals of subject #2. This lack of alpha rhythms may be due to the "constant definition" of the frequency bands (i.e., 4–8 Hz for theta activity and 8–12 Hz for alpha rhythms) adopted by KDS. Indeed, the alpha frequency has been shown to vary to a large extent as a function of age, brain volume, neurological diseases, memory performance, and task difficulty [85]. Therefore, in practice, the alpha frequency band should rather be defined around an individualized anchor frequency called the individual alpha frequency (IAF) [85], which denotes the dominant EEG frequency (i.e., the frequency at peak EEG power). With this "individualized definition", a low IAF (i.e., lower alpha and theta frequency bands) would explain the lack of alpha rhythms: the alpha rhythms (in the "individualized definition") would be misdiagnosed as theta activities (in the "constant definition" of KDS), and the theta activities (in the "individualized definition") would be unnoticed as their frequencies would be below the frequency band of the "constant definition". As a result, the KDS scores would be underestimated, which is probably the case for subject #2.

3.6.4 Statistical distribution of measures of performance impairment

In practice, measures of performance impairments can be automatically and objectively produced from the RTs using sliding time windows, which is convenient. However, the mapping of performance measures to a score of drowsiness is not straightforward. To increase our understanding of performance measures and their link to drowsiness, let us analyze two standard measures computed over time windows of 60s with a step of 30s: the mean RT and the percentage of lapses.

Figure 3.5 shows the box plot of the mean RT and the percentage of lapses as a function of the PVT index. Overall, we observe that performance is increasingly impaired with increasing sustained waking, i.e., from PVT1 to PVT3. Furthermore, we observe that 75% of the mean RTs are below 382ms during PVT1; and that 50% of the mean RTs are above 382ms during PVT2, and above 404ms during PVT3. We also observe that (nearly) all mean RTs above 500ms (which is the threshold above which a RT is conventionally considered as a lapse) were recorded during PVT2 or PVT3, i.e., in sleep-deprived conditions where drowsiness is greatly favored.

Figure 3.6 shows the mean RT and the mean percentage of lapses as a function of time elapsed during each PVT. Again, we observe that performance is increasingly impaired with increasing sustained waking. However, we also observe the impact of the time-on-task on performance impairment: the performance measures remain almost constant during PVT1, slightly increase during PVT2, and significantly increase during PVT3. Indeed, via least-squares linear regression, we obtain the following line equations respectively for the mean RT (in ms), and mean percentage of lapses (in %): $0.7x+355$ms, and $-0.08x+5$% during PVT1; $3.6x+394$ms, and $0.5x+12$% during PVT2; and $11x+391$ms, and $1.5x+11$% during PVT3; where x is the time across the 10-min PVT expressed in minutes. Therefore, the more sleep-deprived an individual is, the more the time-on-task has an impact on his/her performance.

Figure 3.7 shows the box plot of the mean RT as a function of the subject index in two conditions: non sleep-deprived (PVT1) and sleep-deprived (PVT2 and PVT3). In non sleep-deprived conditions, we observe a significant inter-subject variability in the mean RT, as quantified by the median (of the mean RT) ranging from 264ms to 441ms. This greatly suggests that the PVT is affected by aptitude, which is in contradiction to what is stated in the literature [44]. However, the non sleep-deprived intra-subject variability is small, i.e., the distribution of their mean RTs is narrow, as quantified by the fact that 50% of the data (between the first and third quartiles, i.e., the box height) is contained within 27 ± 11ms (mean ± standard deviation). In sleep-deprived condition, the median (of the mean RT) ranges from 302ms to 615ms, which demonstrates the inter-subject difference in vulnerability to drowsiness, i.e., the impairments of performance are severe for some subjects, yet rather modest for others. The intra-subject variability is also greater when sleep-deprived, with 50% of the data contained within 73 ± 72ms.

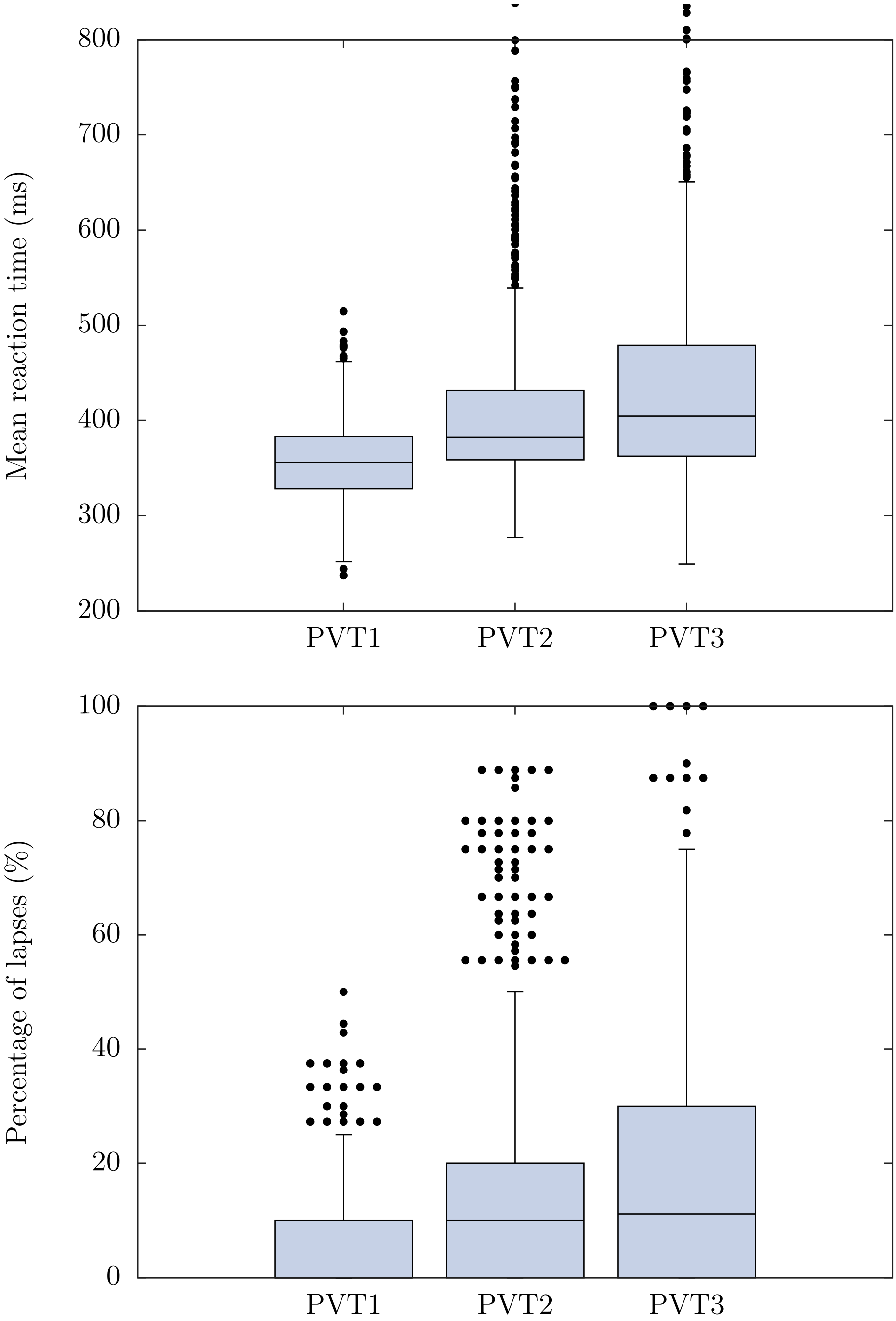
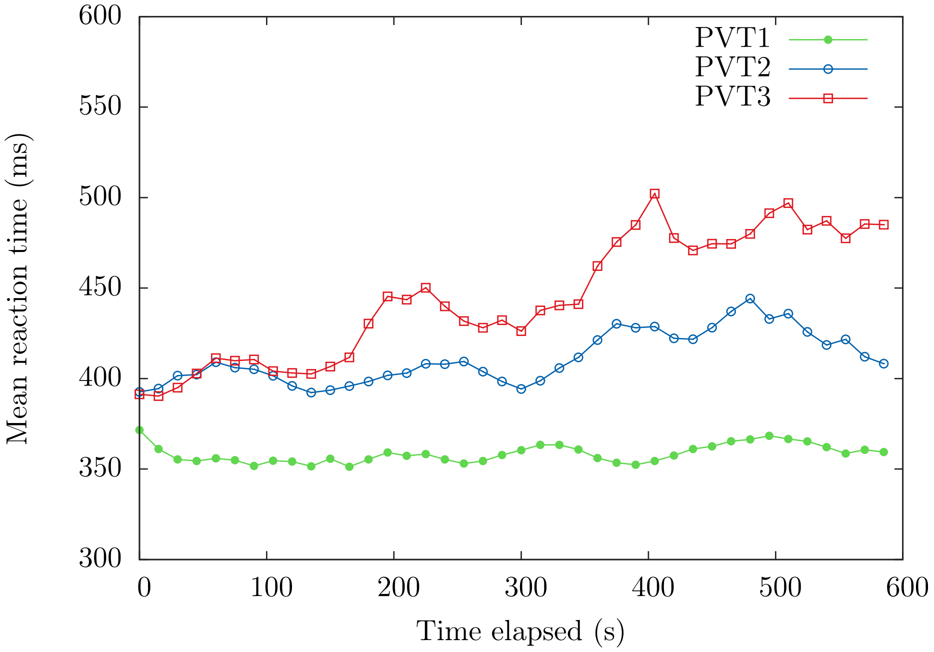


Figure 3.5 – Box plot of the performance measures (the mean RT at the top, and the percentage of lapses at the bottom) as a function of the PVT index, over all the subjects. The performance measures (i.e., mean RTs and percentages of lapses) are generated via a sliding time window with a step of 30s and a duration of 60s. Note that (1) the highest mean RT for PVT2 and PVT3 are 964ms and 1578ms, respectively, and (2) the median percentage of lapses for PVT1 is 0. Overall, we observe the increasing impairments of performance with the increasing sleep deprivation conditions, i.e., from PVT1 to PVT3.



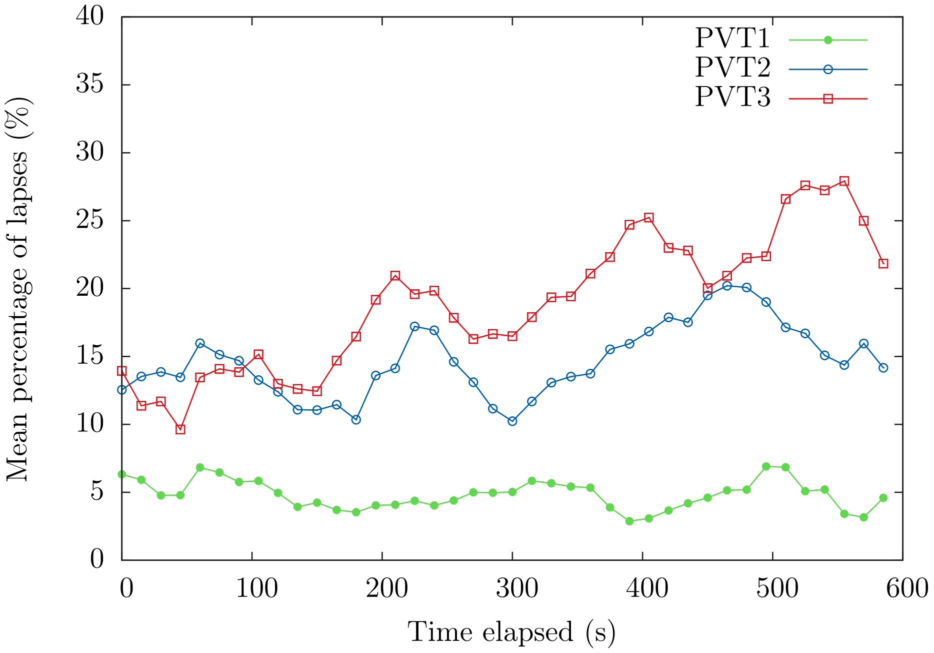


Figure 3.6 – Plot of the mean performance measures (the mean RT at the top, and the mean percentage of lapses at the bottom) as a function of time elapsed during each PVT. The performances measures are each generated differently, and defined as follows. At time t, the "mean RT" corresponds to the average of all the RTs that occurred within the $[t, t+60\text{s}]$ time window, across all subjects. At time t, the "mean percentage of lapses" corresponds to the average, over all subjects, of the percentage of lapses computed within the $[t, t+60\text{s}]$ time window. We observe the impairment effects of time on the performance metrics in terms of both the time across the day (from PVT1 to PVT3), and the time elapsed during the PVT (from 0 to 10 minutes).

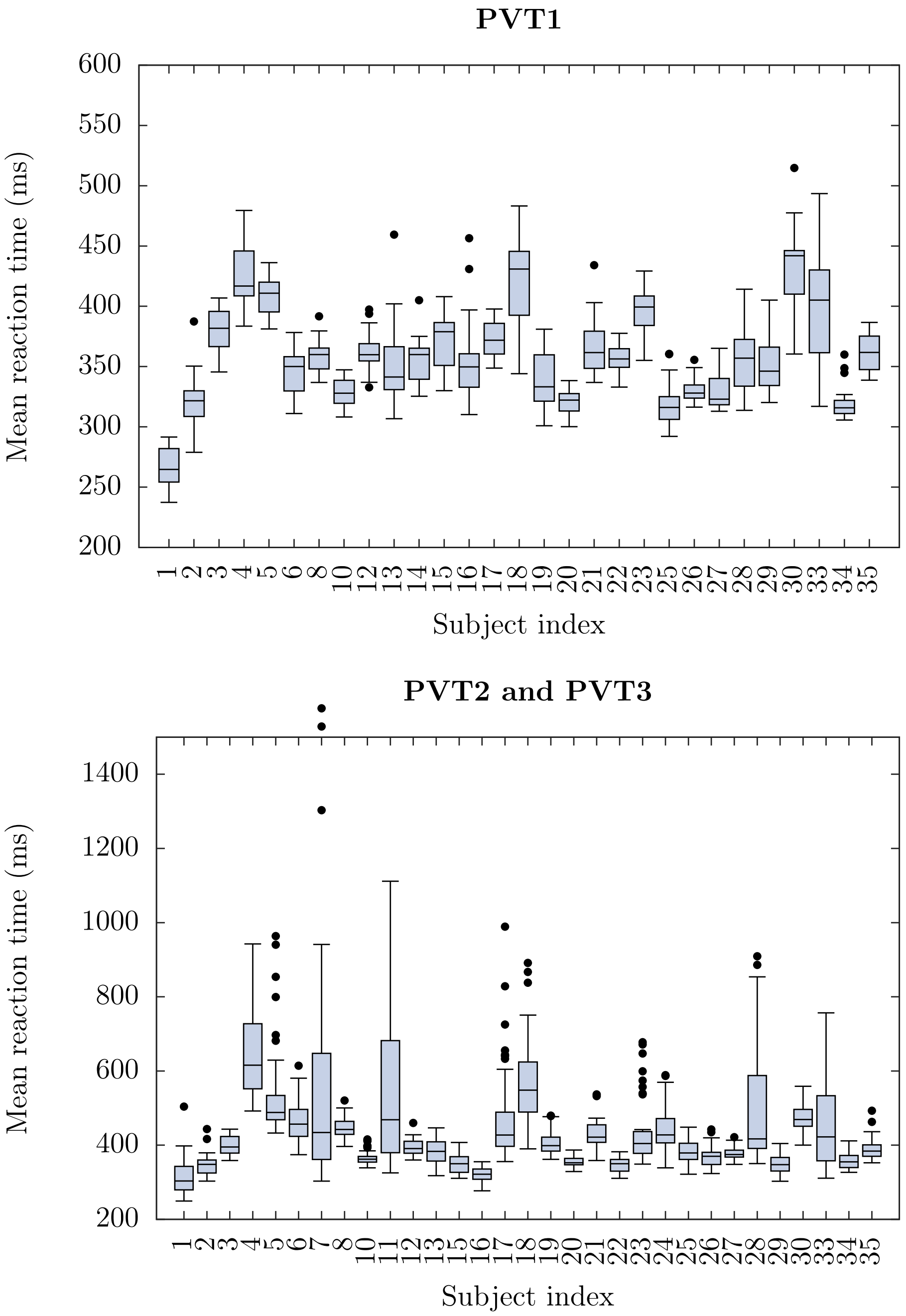


Figure 3.7 – Box plot of the mean RT as a function of the subject index during either PVT1 (top), or during PVT2 and PVT3 (bottom). Carefully note the difference in y-axis scale between both figures. The mean RTs are generated via a sliding time window with a step of 30s and a duration of 60s. We observe a significant inter-subject variability in the median (of the mean RT) during PVT1 (non sleep-deprived conditions). The intra-subject variability in the mean RT is small during PVT1, and increased during PVT2 and PVT3. We also observe the inter-subject difference in vulnerability to drowsiness.

3.7 Conclusion

In conclusion, we designed and acquired[1] a sleep-deprivation dataset that:

1. involves 32 young, healthy subjects, who each performed three 10-min PVTs;
2. involves a wide range of drowsiness levels, induced by acute sleep deprivation of up to 30 hours, i.e., by prolonged, sustained waking over 2 days;
3. contains (time-synchronized) near-infrared intensity and range images of the face at a frame rate of 30 frames per second;
4. contains the (time-synchronized) RTs to random, visual stimuli, i.e., performance-based indicator of drowsiness;
5. contains the (time-synchronized) PSG signals, i.e., physiology-based indicators of drowsiness;
6. contains the KSS, i.e., a subjective indicator of drowsiness, at the beginning of each PVT;
7. is available to the scientific community for both a subset of subjects [91], and for a subset of the data [93];
8. appears to have strong (relative) ecological validity.

By inspecting and analyzing standard measures of the three indicators of drowsiness we recorded, we highlighted the following properties about the indicators of drowsiness:

1. the measures of all three indicators increase when the duration of sustained waking increases, i.e., from PVT1 to PVT3;
2. there exists an inter-subject difference in vulnerability to drowsiness;
3. the annotation criteria of KDS might not be suited to all subjects, potentially because of the constant definition of the frequency bands of the brain waves;
4. the two measures of performance impairment increase with the time-on-task, and the impact of time-of-task grows larger the more the subject is sleep-deprived, i.e., from PVT1 to PVT3;
5. there exists an inter-subject difference in aptitude/skill when performing a PVT.

In the next chapters, we detail the drowsiness characterization systems that we designed in this thesis. Each of our drowsiness characterization systems use this sleep-deprivation dataset in a different way. However, the definition of their ground truth of drowsiness is always in terms of the RTs, i.e., a performance-based indicator of drowsiness. The reasons for not using other indicators for training our systems are as follows. For the physiology-based indicator, the analysis of the PSG signals to produce KDS scores is challenging, time-consuming, and affected by the subjective interpretation of the human annotator(s). For the subjective indicator, we have collected only one KSS score per PVT, which is too sparse of an annotation to develop real-time drowsiness characterization systems. In

[1] We thank Clémentine François, David Grogna, Thomas Langohr, and Philippe Latour for their help in setting up and supervising the collection of data.

contrast, the RTs are objective (i.e., free of any human's interpretation), automatically-annotated, and relatively densely-annotated. Although mapping the RTs to a level of drowsiness is not straightforward, the impairments of RTs still carry valuable information about the physiological impacts of drowsiness. However, note that this holds true as far as the impairments of RTs are induced by drowsiness, which is mainly the case given the controlled conditions of our study protocol that involves acute sleep deprivation.

Chapter 4

Baseline drowsiness characterization system

This chapter presents a baseline drowsiness characterization system that is representative of most systems of other studies, and that enables us to study eye closure dynamics with respect to the level of drowsiness. Section 4.1 introduces and motivates our baseline system. Section 4.2 describes our system. Section 4.3 details the training of our system. Section 4.4 reports experimental results, and evaluates the performance. Section 4.5 concludes this chapter. This chapter is based on the following published conference paper [91]: Q. Massoz, T. Langohr, C. François, and J. Verly. The ULg multimodality drowsiness database (called DROZY) and examples of use. In IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–7, Lake Placid, NY, USA, March 2016.

4.1 Introduction

As seen in Chapter 2, eye closure dynamics and performance impairments are both recognized as strong indicators of drowsiness. Whereas eye closure dynamics may be acquired in a wide range of settings via computer vision algorithms, performance impairments can be automatically recorded during Psychomotor Vigilance Tasks (PVTs). One may thus develop an automatic, real-time drowsiness characterization system that produces an estimate of drowsiness-induced performance impairments from eye closure dynamics. However, the scientific literature has so far given little attention to the relationship between eye closure dynamics and performance impairments.

With the goals of (1) representing a panel of systems of other studies that is as wide as possible and (2) studying the relationship between eye closure dynamics and performance impairments, we present a baseline drowsiness characterization system that automatically extracts a vector of standard, pre-defined ocular features from a face video. Based on this vector of ocular features, we consider four ways of characterizing drowsiness that are formulated as four problems: (1) an estimative regression problem; (2) a predictive regression problem; (3) an estimative binary classification problem; and (4) a predictive binary classification problem. The regression problems aim at outputting the mean reaction time (RT), whereas the classification problems aim at outputting a binary Level of Drowsiness (LoD) based on impairments of performance. The estimative problems aim at estimating the output of the current minute, whereas the predictive problems aim at predicting the output of the next minute. We evaluate the performance of the baseline system at these four problems, we compare the performance between estimative and predictive problems, we analyze the importance of each ocular features in the system decision, and we analyze the correlation of standard ocular features with several standard measures of drowsiness.

4.2 Baseline system

Our baseline drowsiness characterization system is composed of four modules operating in cascade: the "face landmarks" module, the "eyelids distance" module, the "ocular features" module, and the "drowsiness" module. Figure 4.1 depicts the baseline system and its four modules.

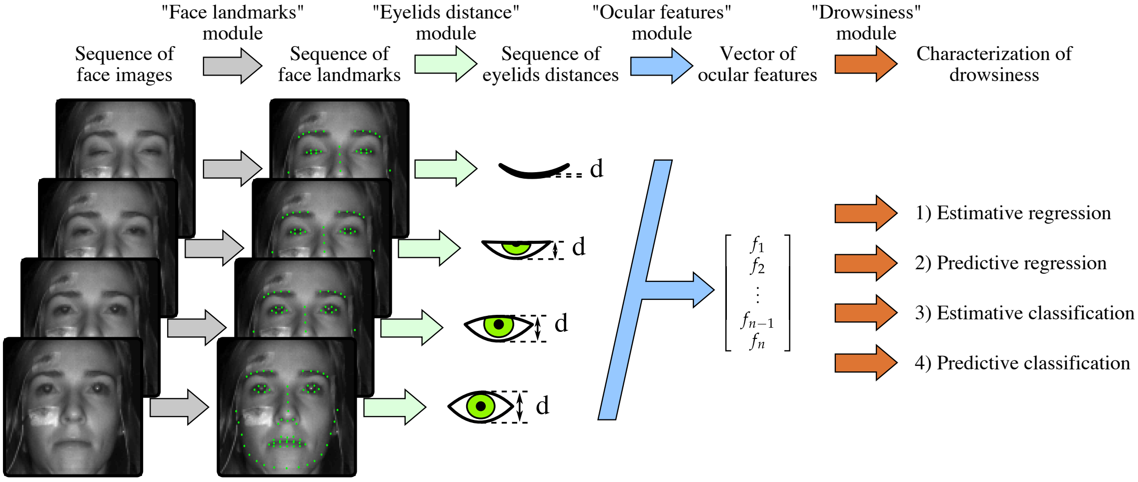

"Face landmarks" module
"Eyelids distance" module
"Ocular features" module
"Drowsiness" module
Sequence of face images
Sequence of face landmarks
Sequence of eyelids distances
Vector of ocular features
Characterization of drowsiness
d
d
d
d
f_1
f_2
⋮
f_{n-1}
f_n
1) Estimative regression
2) Predictive regression
3) Estimative classification
4) Predictive classification

Figure 4.1 – Overview of the baseline drowsiness characterization system operating on any given 1-min sequence of face images. First, from each face image, the "face landmarks" module tracks the position of 68 face landmarks via subject-specific constrained local models (CLMs). Second, from the eyelids landmarks of each frame, the "eyelids distance" module extracts the average eyelids distance. Third, from the most-recent 1-min sequence of eyelids distances, the "ocular features" module extracts a vector of ocular features via an algorithm that temporally segments each blink. Fourth, from the vector of ocular features, the "drowsiness" module characterizes drowsiness in four different ways. The "drowsiness" module consists of a Support Vector Machine (SVM) model for classification problems, and a Support Vector Regression (SVR) model for regression problems.

4.2.1 "Face landmarks" module

The "face landmarks" module tracks the 3D position of 68 face landmarks from the sequence of video frames, where a video frame consists of two co-registered images: a near-infrared (NIR) image, denoted by $\mathcal{I}$, and a depth map, denoted by $\mathcal{D}$. We filter $\mathcal{D}$ first with a median filter with a kernel size of 3×3, then with a bilateral filter with a kernel size of 5×5 and standard deviations of 200 (in the coordinate space and in the intensity space). Then, we iteratively align a 68-landmarks deformable shape model using constrained local models (CLM) and the regularized landmark mean-shift (RLMS) fitting procedure. We provide a technical background on these methods in Appendix A, and we list the several adjustments we made to the classic formulation of CLM and RLMS to obtain (1) better robustness to occlusions and (2) better alignment performance for landmarks with multi-modal appearance (such as the ones located on the eyelids and the mouth). We initialized

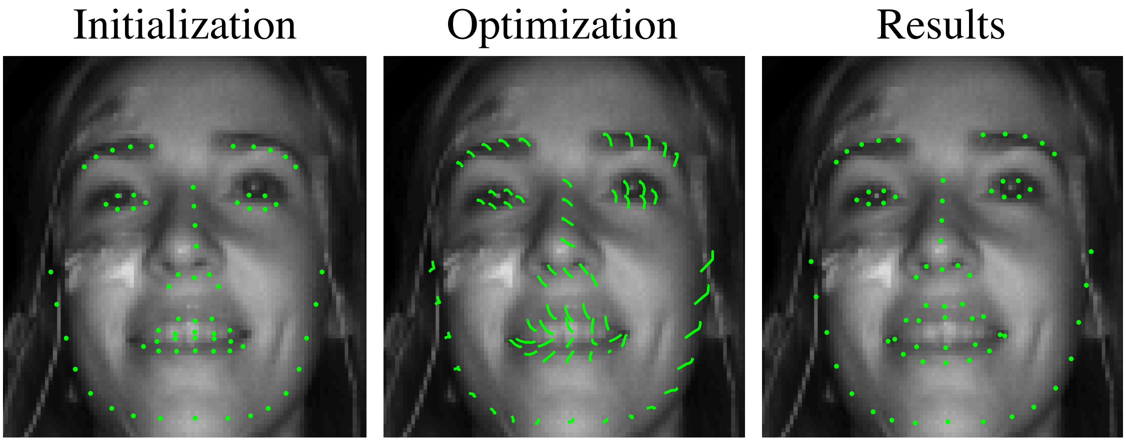


Figure 4.2 – Illustration of the "face landmarks" module. We initialize the 68 face landmarks from either the Viola and Jones algorithm or the previous frame (left), iteratively minimize the misalignment error using the RLMS fitting strategy (middle), and obtain the location of 68 landmarks aligned on the target image (right).

(or re-initialized, in case of tracking failures) the tracking with the OpenCV [20] implementation of the Viola and Jones algorithm [138] by centering a neutral face shape around the center of the detected box. Figure 4.2 illustrates the initialization, optimization, and results of the "face landmarks" module.

4.2.2 "Eyelids distance" module

The "eyelids distance" module extracts the average 3D eyelids distance (a real positive number, expressed in mm) from the 68 face landmarks, denoted by $\mathbf{X}_i \in \mathbb{R}^3$ with $i \in [1, 68]$. The eyelids of each eye are described by 6 face landmarks: 2 at the corners, 2 on the upper eyelid, and 2 on the lower one. The average eyelids distance is defined as the average of the four 3D inter-eyelid Euclidean distances, i.e., the distances, for each eye, between the two face landmarks positioned on the upper eyelid, and the two on the lower eyelid. Figure 4.3 illustrates the four pairs of eyelids landmarks used to compute the average eyelids distance.

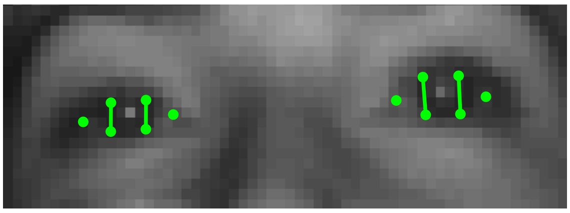


Figure 4.3 – Illustration of the "eyelids distance" module. The eyelids distance is computed from the 3D face landmarks (green dots) as the average of four 3D Euclidean distances (green lines).

4.2.3 "Ocular features" module

The "ocular features" module extracts a vector of ocular features from the most-recent 1-min sequence of eyelids distances, this in three steps. First, we adaptively-normalize the

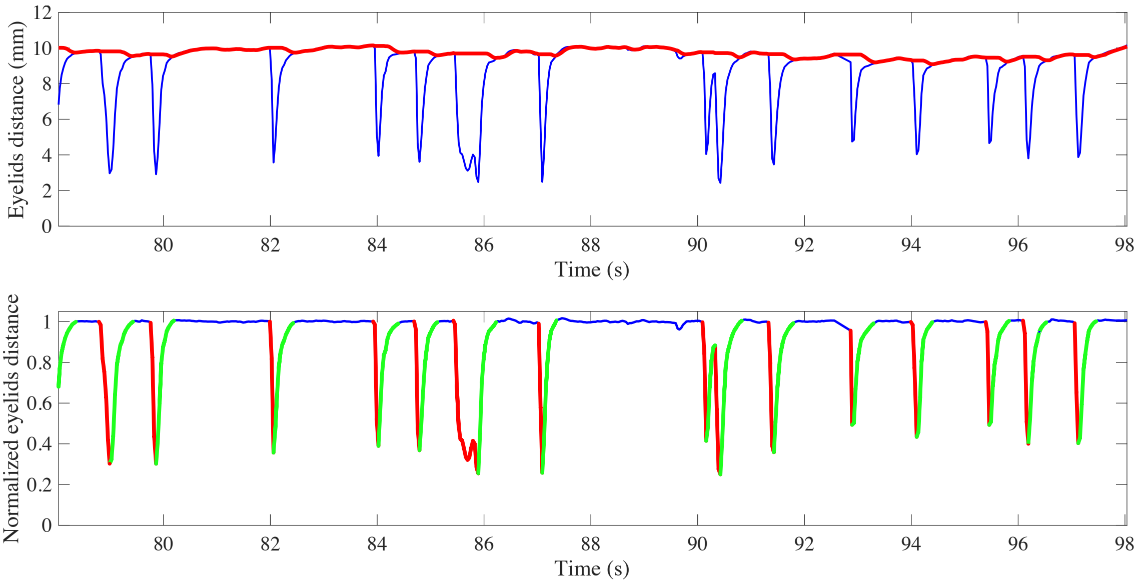


Figure 4.4 – Illustration of the "ocular features" module. The top graph illustrates a sequence of eyelids distances (in blue) and the corresponding baseline sequence (in red). The bottom graph illustrates the corresponding sequence of adaptively-normalized eyelids distances (in blue), where the eye-closing segments are colored in red and the eye-opening segments in green. Note that there is no eye-closed segment in this example.

sequence of eyelids distances. Second, we identify the time segments corresponding to the three phases (the closing phase, the closed phase, and the opening phase) of each blink. Third, we extract the ocular features. Figure 4.4 illustrates the adaptative normalization step and the segmentation step of the "ocular features" module.

Adaptive normalization of eyelids distances

We denote the sequence of eyelids distances by $d[n]$, where $n \in \mathbb{N}$ is the discrete time index. Since the maximum opening of the eye changes with time, e.g., with the gaze direction and the head pose, it proves useful to divide $d[n]$ by the maximum opening at time n, denoted by $b[n]$, which results in the sequence of adaptively-normalized eyelids distances $s[n] = d[n]/b[n]$. The baseline sequence $b[n]$ is computed recursively according to

$$b[n] = (1 - \alpha[n])\, b[n-1] + \alpha[n] d[n], \tag{4.1}$$

where $\alpha\,[n]$ is a smoothing factor defined as

$$\begin{aligned} \alpha[n] = \alpha_0 \exp &\left(-\alpha_d \left(d[n] - d[n-1]\right)^2\right) \\ &* \exp\left(-\alpha_a \left[d[n] - b[n-1]\right]_+\right) \\ &* \exp\left(-\alpha_b \left[b[n-1] - d[n]\right]_+\right) \\ &* H\left(d[n] - \alpha_m d_{median}[n]\right), \end{aligned} \tag{4.2}$$

where $*$ denotes multiplication, $[x]_+$ is x if $x \geq 0$, and 0 otherwise; $H\,(x)$ is the Heaviside step function defined as 1 if $x \geq 0$, and 0 otherwise; and $d_{median}[n]$ is the median value of

$\{d\,[i] : \forall i \in [1, n]\}$. The values of α_0, α_d, α_a, and α_m are empirically set to 0.4, 15, 0.5, 2, and 0.7, respectively.

The smoothing factor $\alpha[n]$ is designed to favor eyelids idleness (i.e., small values of the difference of $d[n]$) but not when the eye is nearly closed eyelids (i.e., $d[n]$ below its median value). It is also designed to be flexible enough to allow for the baseline signal $b[n]$ to quickly adapt when the gaze direction changes.

Segmentation of blinks

A blink is composed of three phases: the closing phase, the closed phase, and the opening phase. The segmentation of blinks consists therefore in identifying the three time segments (each defined as a set of contiguous time indices) corresponding to the three phase of each blink. We proceed in four steps.

In the first step, we identify candidate time segments by applying experimentally-set thresholds on the backward difference of $s[n]$, $\nabla s[n] \equiv s[n] - s[n-1]$. More specifically, we identify three classes of candidate segments:

1. the candidate closing segments: $\{n \in \mathbb{N} : n_1 \leq n \leq n_2 \text{ and } \nabla s[n] \leq \lambda_1\}$,
2. the candidate opening segments: $\{n \in \mathbb{N} : n_1 \leq n \leq n_2 \text{ and } \nabla s[n] \geq \lambda_2\}$,
3. the candidate plateau segments: $\{n \in \mathbb{N} : n_1 \leq n \leq n_2 \text{ and } \lambda_1 < \nabla s[n] < \lambda_2\}$,

where $n_1 \in \mathbb{N}$ is the time index of the start of the candidate segment, $n_2 \in \mathbb{N}$ is the time index of the end, and $\lambda_1 = -13.2\text{e}{-3}$ and $\lambda_2 = 5.5\text{e}{-3}$ are the experimentally-set thresholds. We define a plateau as an eye state that is either continuously closed or continuously open. Note that each and every time index n belongs to exactly one candidate segment because (1) n_1 can equal n_2 (making the candidate segment consisting of one sample at time index n_1), and (2) the conditions on $\nabla s[n]$ are mutually exclusive.

This first step results in a sequence of candidate segments, denoted by $a[k]$ where $k \in \mathbb{N}$ is the candidate segment index. If we denote the three classes of candidate segments respectively by "c", "o", and "-", one would expect each blink to be associated with a subsequence of $a[k]$, denoted by $(a[k_1], a[k_1+1], \ldots, a[k_2])$, similar to (c, o) for short blinks, or to (c, -, o) for long blinks. However, given the noise in $s[n]$ and—by extension—in $\nabla s[n]$, candidate segments can be misclassified, thereby resulting in blinks associated with subsequence of $a[k]$ similar to (c, -, c, o), (c, o, -, o), or (c, -, c, -, o, -, o, -, o). Therefore, in the next steps, we aim at finding the best combination of (potentially misclassified) candidate segments so as to obtain the true closing segment and the true opening segment of each blink.

In the second step, we clean $a[k]$ by discarding (i.e., classifying as candidate plateau segments) the candidate closing segments and the candidate opening segments that do not satisfy the following conditions: $\min(s[n_1], s[n_2]) \leq 0.81$ and $|s[n_1] - s[n_2]| > 0.13$. Then, to avoid having two contiguous candidate segments of the same class, we combine contiguous candidate segments together, and denote this new sequence of candidate segments by $a'[k']$. For examples, the following subsequence of $a[k]$ (c_1, -, c_2, -, o_1, -, o_2) may become (-, c_2, -, o_2), (c_1, -, o_1, -, o_2), (-, c_2, -), or even (-) as subsequence of $a'[k']$.

In the third step, we parse $a'[k']$ into subsequences each corresponding to one blink. We consider that a candidate closing segment, "c", cannot be misclassified as an candidate opening segment, "o", and vice-versa. Therefore, the subsequence of a blink (1) always starts by a "c" that is preceded by an "o" or by (o, -), and (2) always ends by an "o" that is followed by a "c" or by (-, c). Obviously, we make some exceptions at the boundaries of

$a'[k']$, e.g., the full sequence of $a'[k']$ (-, c, o, -, c, -, c, o, c, -, o, -, o) is parsed into the three subsequences (c, o), (c, -, c, o), and (c, -, o, -, o).

In the fourth step, we analyze each subsequence of a blink so as to identify the three phases of this blink. We consider that the true closing (opening, respectively) segment is composed of successive "c"s (successive "o"s, resp.) that can be interspersed by "-"s. For example, the true closing segment of (c_1, -, c_2, -, o_1, -, o_2) can be (c_1), (c_2), or (c_1, -, c_2). Therefore, we search, over all pair combinations of true closing segments (successive "c"s) and true opening segments (successive "o"s), the one that best represents the blink, i.e., that maximizes the following quantity:

$$\begin{aligned} B\left(n_1, n_2, n_3, n_4\right) = & \exp\left(-|\Delta h_p|\right) \\ & * \exp\left(-|\Delta h_c - \Delta h_o|\right) \\ & * H\left(\Delta h_c - 0.1\right) \\ & * H\left(\Delta h_o - 0.1\right), \end{aligned} \tag{4.3}$$

where n_1 (n_2, resp.) is the time index of the start (end, resp.) of the first (last, resp.) "c" in the true closing segment, n_3 (n_4, resp.) is the time index of the start (end, resp.) of the first (last, resp.) "o" of the true opening segment, $\Delta h_p = s[n_3] - s[n_2]$ is the "height" of the true closed segment, $\Delta h_c = s[n_1] - s[n_2]$ is the "height" of the true closing segment, $\Delta h_o = s[n_4] - s[n_3]$ is is the "height" of the true opening segment. In other words, we consider a pair of a true closing segment and a true opening segment to be probable if Δh_f is close to zero, Δh_c is close to Δh_o, Δh_c is greater than 0.1, and Δh_o is greater than 0.1. Note that, if the subsequence of a blink is composed of N_c candidate closing segments ("c"s) and N_o candidate opening segments ("o"s), we maximize B over $(N_c!) * (N_o!)$ pair combinations. In the case where B equals 0 for every pair combinations, we discard the corresponding blink.

In such a manner, for each blink, we identify (1) the true closing segment, (2) the true opening segment, and (3) the true closed segment, which is defined as the time segment between (1) and (2).

Extraction of ocular features

We consider 15 ocular features, all computed for a contiguous time window W that ends at the present time index. (The length of W is specified below.) Ten features are related to the histogram of the values of $s[n]$ in W, and five are related to the segmented blinks in W.

The 10 histogram-related features are the 10 proportions of elements in each the 10 successive bins of a 10-bins histogram, in which all values of $s[n]$ in W are arranged, with all values of $s[n]$ above 1 placed in the last bin. The 10 histogram-related features, i.e., proportions, sum to 1. We denote these histogram-related features by $H_{[0,0.1]}$, $H_{[0.1,0.2]}$, $H_{[0.2,0.3]}$, $H_{[0.3,0.4]}$, $H_{[0.4,0.5]}$, $H_{[0.5,0.6]}$, $H_{[0.6,0.7]}$, $H_{[0.7,0.8]}$, $H_{[0.8,0.9]}$, and $H_{[0.9,1]}$.

The 5 blink-related features are average metrics computed over the segmented blinks that end within W. The blink-related features are the average blink duration, $\overline{D}_{blink}$; the average closing duration, $\overline{D}_{closing}$; the average closed duration, $\overline{D}_{closed}$; the average opening duration, $\overline{D}_{closing}$; and the number of microsleeps, $N_{\mu sleeps}$, where a microsleep is defined as a blink with a duration greater or equal than 500ms. The durations are expressed in milliseconds. The closing duration is the time duration to go from 70% to 0% of the amplitude of the closing segment. The closed duration is the time duration to go from 10% of the amplitude of the closing segment to 10% of the amplitude of the opening segment.

The opening duration is the time duration to go from 0% to 70% of the amplitude of the opening segment. The blink duration is the time duration to go from 70% of the amplitude of the closing segment to 70% of the amplitude of the opening segment.

We extract these 15 ocular parameters for 7 time windows with lengths of $\{30, 35, 40, 45, 50, 55, 60\}$ seconds, which results in a vector of 105 ocular features.

4.2.4 "Drowsiness" module

The "drowsiness" module characterizes drowsiness from the vector of 105 ocular features. We consider four ways for characterizing drowsiness that are formulated as four problems:

1. estimative regression of the mean RT of the most-recent minute;
2. predictive regression of the mean RT of the next minute;
3. estimative binary classification of the Level of Drowsiness (LoD) of the most-recent minute;
4. predictive binary classification of the LoD of the next minute.

For regression problems, the "drowsiness" module consists of a Support Vector Regression (SVR) model. For binary classification problems, the "drowsiness" module consists of a Support Vector Machine (SVM) model. We provide a technical background on SVM and SVR in Appendix B. Note that the adjectives "estimative" and "predictive" are non-standard terms, but are used here to indicate that the ground truth has been produced, respectively, from the current present minute (estimate of the present) and from the following minute (prediction of the future).

4.3 Training of the system

Two of the modules require training: the "face landmarks" module and the "drowsiness" module. We trained these two modules using the data from 14 subjects of our sleep-deprivation dataset (detailed in Chapter 3). We emphasize that the "eyelids distance" module and the "ocular features" module do not require any training.

4.3.1 "Face landmarks" module

Rather than training one "face landmarks" module, we trained 14 subject-specific modules, i.e., one specific CLM per subject. The reason was to ensure that the face landmarks were located as precisely as possible, so that further analyses of the "drowsiness" module were as least as possible affected by tracking errors. Note that this choice makes the system non-automatic. However, we could make it automatic by training a generic "face landmarks" module.

Dataset

For training the 14 "face landmarks" modules, we built the "face landmarks" dataset. This dataset is composed of 720 manually-selected key frames spread across the 14 subjects (about 51 frames per subject) and judged to represent rather completely the various appearances of the face of each subject. For each of these frames, we manually annotated 68 landmarks both in 2D (in the two image coordinates, expressed in pixels) and in 3D

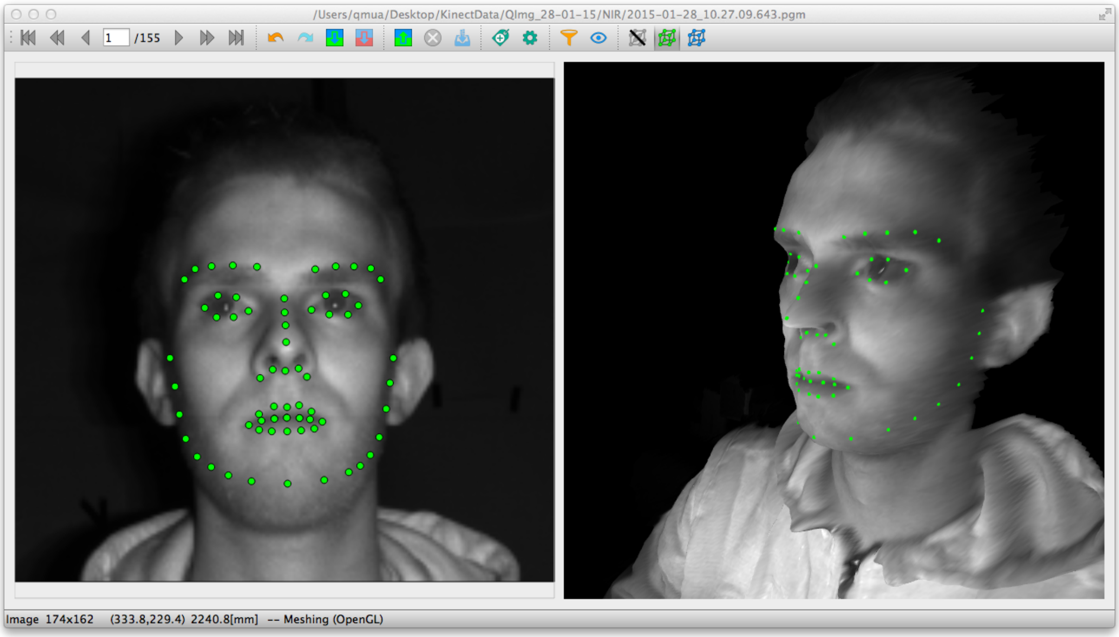

/Users/qmua/Desktop/KinectData/QImg_28-01-15/NIR/2015-01-28_10.27.09.643.pgm
1
/155
Image 174x162 (333.8,229.4) 2240.8[mm] -- Meshing (OpenGL)

Figure 4.5 – The software we developed for annotating a set of face landmarks in 2D on the image (left), and in 3D on the reconstructed 3D mesh (right).

(in the three camera coordinates, expressed in millimeters) by the means of a specifically-developed annotation software. This annotation software, illustrated in Figure 4.5, automatically builds a 3D surface of the face from the depth map $\mathcal{D}$ with the texture sampled from the (co-registered) NIR image $\mathcal{I}$. By looking jointly at $\mathcal{I}$ and the rotatable, textured 3D surface, we pinpointed the 2D location of each visible landmark in the video frame. Since $\mathcal{I}$ and $\mathcal{D}$ are co-registered, the location of each such landmark was also immediately known and annotated in 3D.

Training

We provide in Appendix A.5 the details for training the subject-specific CLMs, i.e., the subject-specific deformable shape models and the subject-specific multimodal appearance models.

4.3.2 "Drowsiness" module

Dataset

For training the "drowsiness" module, we built the "drowsiness" dataset from 14 subjects who performed a total of 36 PVTs. In particular, we sampled the 1-min sequences of video frames that end at the occurrence time of every PVT stimulus (except for the PVT stimuli that occurred within the first minute of the PVT). This sampling strategy leads to 3064 samples, with an average of about 85 samples per PVT, and of about 219 samples per subject. We pre-processed each 1-min sequence with the subject-specific "face landmarks" module, the "eyelids distance" module, and the "ocular features" module, so as to obtain a vector of 105 ocular features per sample, which is the input to the "drowsiness" module.

Ground truth of drowsiness

Since we consider four distinct problems for the baseline drowsiness characterization system, we define a distinct ground truth of drowsiness for each problem.

- For the estimative (predictive, respectively) regression problem, the ground truth of drowsiness consists of the pre-stimulus (post-stimulus, respectively) mean RT (expressed in milliseconds), where the mean is computed over the 1-min period before (after, respectively) the stimulus occurrence.

- For the estimative (predictive, respectively) classification problem, the ground truth of drowsiness consists of a binary LoD based on whether or not the pre-stimulus (post-stimulus, respectively) mean RT is above some fixed threshold, where the mean is computed over the 1-min period before (after, respectively) the stimulus occurrence. We set the threshold to 500ms since an RT (observed during a PVT) above this value is conventionally interpreted as a lapse [13, 44], i.e., an error of omission. The "drowsy" label ("positive" class, an LoD of 1) corresponds to a mean RT $\geq$ 500ms, whereas the "alert" label ("negative" class, an LoD of 0) corresponds to a mean RT $<$ 500ms. Out of the 3064 samples, 448 are labeled "drowsy" and 2616 "alert" for the estimative problem, and 383 are labeled "drowsy" and 2681 "alert" for the predictive problem.

Training and optimization

For each of the four problems, we trained 14 models (SVMs for classification problems, and SVRs for regression problems) following a leave-one-subject-out cross-validation strategy of 14 folds, i.e., one test set for each subject. For each fold, we validated the hyper-parameters (i.e., $\{C\}$ for classification problems, and $\{C, \epsilon\}$ for regression problems) via an inner leave-one-subject-out cross-validation strategy of 13 folds, i.e., all subjects but the one in the test set of the outer cross-validation. Upon determination of the optimal set of hyper-parameters, we trained the final model on all 13 subjects of the training set (of the outer cross-validation). We individually scaled each ocular feature such that each feature of training samples ranges within $[0, 1]$. For classification, we weighted the two classes (i.e., "alert" and "drowsy") in the SVM optimization routine with the reciprocal of the number of their occurrence in the training set. We performed no data augmentation. We performed training and inference of SVRs and SVMs with the LIBLINEAR library [47] and LIBSVM library [26].

4.4 Experimental results and performance

4.4.1 Evaluation of performance

We evaluated the performance of our baseline system by aggregating the results of the 14 test sets, each associated with one trained model, before computing the performance metrics. We did not average the performance metrics across the 14 subjects because (1) the amount of data is not identical for all subjects (some PVTs were missing), and (2) the proportion of "drowsy"/"alert" samples varies significantly between subjects. We trained either with a linear kernel for interpretability, or with an RBF kernel for performance. Table 4.1 summarizes the results obtained, and these are further discussed below, first for the linear kernel, and then for the RBF kernel.

		Estimation		Prediction	
		Linear	**RBF**	**Linear**	**RBF**
Regression	**PCC**	0.58	0.62	0.54	0.60
	RMSE	108	104	120	114
Classification	**TNR**	77.40%	87.73%	77.37%	86.51%
	TPR	79.11%	75.46%	75.89%	75.45%
	Acc.	77.61%	86.19%	77.15%	84.89%

Table 4.1 – Performance metrics of the baseline system with each kernel function and for each problem. The RMSE is expressed in milliseconds.

Linear kernel

1. For the estimative regression problem, we obtained a Pearson correlation coefficient (PCC) of 0.58 and a Root Mean Square Error (RMSE) of 108ms.

2. For the predictive regression problem, we obtained a PCC of 0.54 and an RMSE of 120ms.

3. For the estimative classification problem, we obtained a specificity (true negative rate or TNR) of 77.40%, a sensitivity (true positive rate or TPR) of 79.11%, and a global accuracy of 77.61%.

4. For the predictive classification problem, we obtained a TNR of 77.37%, a TPR of 75.89%, and a global accuracy of 77.15%.

RBF kernel

1. For the estimative regression problem, we obtained a PCC of 0.62 and an RMSE of 104ms.

2. For the predictive regression problem, we obtained a PCC of 0.60 and an RMSE of 114ms.

3. For the estimative classification problem, we obtained a TNR of 87.73%, a TPR of 75.46%, and a global accuracy of 86.19%.

4. For the predictive classification problem, we obtained a TNR of 86.51%, a TPR of 75.45%, and a global accuracy of 84.89%.

4.4.2 Comparison of performance between estimation and prediction problems

Overall, the estimation performance is greater than the prediction performance, this for both regression and classification problems. This difference in performance is consistent with expectations. Indeed, even though the model could incorporate the knowledge that impairment of RT increases with time-on-task (see Chapter 3), prediction is a challenge as the subject's physiological state can change in unpredictable ways. It would be even more the case in operational, real-life settings, where the external conditions also change in unpredictable ways, thereby further impacting the subject in even less predictable ways.

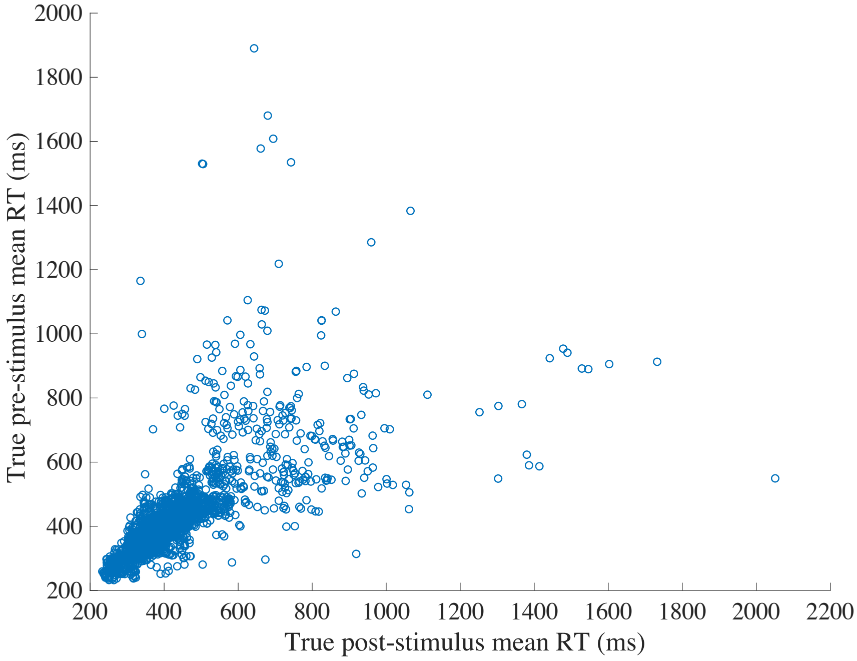


Figure 4.6 – Scatter plot of the true pre-stimulus mean RT vs the true post-stimulus mean RT. We observe that most samples have a mean RT that is similar when computed from the minute before the stimulus (pre-stimulus) and from the minute after (post-stimulus).

Interestingly, the difference in performance between estimation and prediction is not that large. A possible reason is that the ground truth for estimation (based on the pre-stimulus mean RT) and the one for prediction (based on the post-stimulus mean RT) differ little from each other. If this reason is true, the trained models for estimation and prediction would then be nearly identical, and so would their performance. We check this possibility by comparing the two ground truths, this separately for regression problems and for classification problems. Figure 4.6 shows the pre-stimulus and post-stimulus mean RT that we used to produce the ground truth for estimation and prediction, respectively. We observe that most samples have a pre-stimulus mean RT equivalent to the post-stimulus one, which explains why the performance metrics for estimation problems and for prediction problems are similar to each other. On a side note, we also observe that the mean RT can significantly vary in the lapse of just one minute, e.g., the mean RT can vary from 450ms (pre-stimulus) to 1050ms (post-stimulus). This demonstrates the significant time-varying aspect of the level of drowsiness.

For regression problems, the ground truth consists of a mean RT (computed pre-stimulus for estimation, and post-stimulus for prediction). We observe that the two types of mean RT increasingly differ from each other with increasing values of mean RT, resulting in a PCC of 0.69 between the two. However, the estimative regressors and the predictive regressors are almost identical. With an RBF kernel, we observe (1) a PCC of 0.97 between the output of the two regressors and (2) a fitted linear regression model of $E\left(x\right) = -22.1364 + 1.0659P\left(x\right)$ where x is a vector of scaled ocular features, $E\left(x\right)$ is the output of an estimative regressor, and $P\left(x\right)$ is the output of the corresponding predictive regressor (i.e., of the same fold). With a linear kernel, we observe (1) a PCC of 0.99 and (2) a fitted linear regression model of $E\left(x\right) = -5.9271 + 1.0283P\left(x\right)$. These observations and results support the fact that the future of mean RTs is unpredictable; the best strategy for predicting the future mean RT (post-stimulus) is therefore to output an estimate of the present mean RT (pre-stimulus). Figure 4.7 shows scatter plot results for the estimative regressor and the predictive regressor, both with an RBF kernel. We observe that, in both problems, the regression errors are the largest at high values of the mean RT, i.e., where the ground truths differ the most. The regression errors are thus similar, but not distributed similarly among samples.

For classification problems, the binary ground truth is produced by thresholding the mean RT at 500ms. Between estimation and prediction, only 203 out of 3064 samples have a ground truth labeled differently. More specifically, 69 samples are labeled as "alert" after the stimulus but labeled as "drowsy" before, 134 samples are labeled as "drowsy" after the stimulus but "alert" before, 314 samples are labeled as "drowsy" before and after the stimulus, and 2547 samples are labeled as "alert" before and after the stimulus. Hence, considering that the two types of ground truth differ little from each other, it is natural that the estimation performance and the prediction performance differ little from each other.

4.4.3 Importance of both window lengths and types of ocular features

Given that we trained 14 linear models per problem, we can produce, for each problem, importance scores for the 105 features by summing the model linear weights, **w**, over the 14 trained linear models. We can then produce importance scores for the 7 window lengths by summing over the 15 types of ocular features. Similarly, we can produce importance scores for the 15 types of ocular features by summing over the 7 window lengths. Considering that alertness corresponds to the "negative" class and drowsiness to the "positive" one, we can associate negative importance scores to alertness and positive ones to drowsiness.

Window lengths

Table 4.2 shows the importance scores of the 7 time windows for the four problems. Overall, we observe that longer window lengths have higher importance scores, i.e., have more importance in the model decision/approximation function. For the estimative classification problem, the importance score (2.9667) of the 60-s window is more than twice the importance score (1.3973) of the 30-s window. However, for the predictive regression problem, we observe that the 30-s window is more important than the 35-s, 40-s, and 45-s windows, but is less important than the 50-s, 55-s, and 60-s windows.

		Regression		Classification		Normalized
		Estim.	**Pred.**	**Estim.**	**Pred.**	**sum**
Window lengths	**30s**	458.9	1324.5	1.3973	1.6842	0.3616
	35s	664.8	1137.2	1.5917	1.8714	0.3878
	40s	978.4	1206.3	1.7800	1.9778	0.4434
	45s	1316.1	1165.6	2.3572	2.5241	0.5426
	50s	1725.2	1568.2	2.5751	2.7171	0.6463
	55s	2321.1	1913.3	2.7816	2.8467	0.7582
	60s	2793.4	2249.6	2.9667	3.0456	0.8600

Table 4.2 – Importance scores of the window lengths for the four problems. The last column contains the sum of importance scores that were normalized, i.e., scaled such that the sum over each of the four columns equals one.

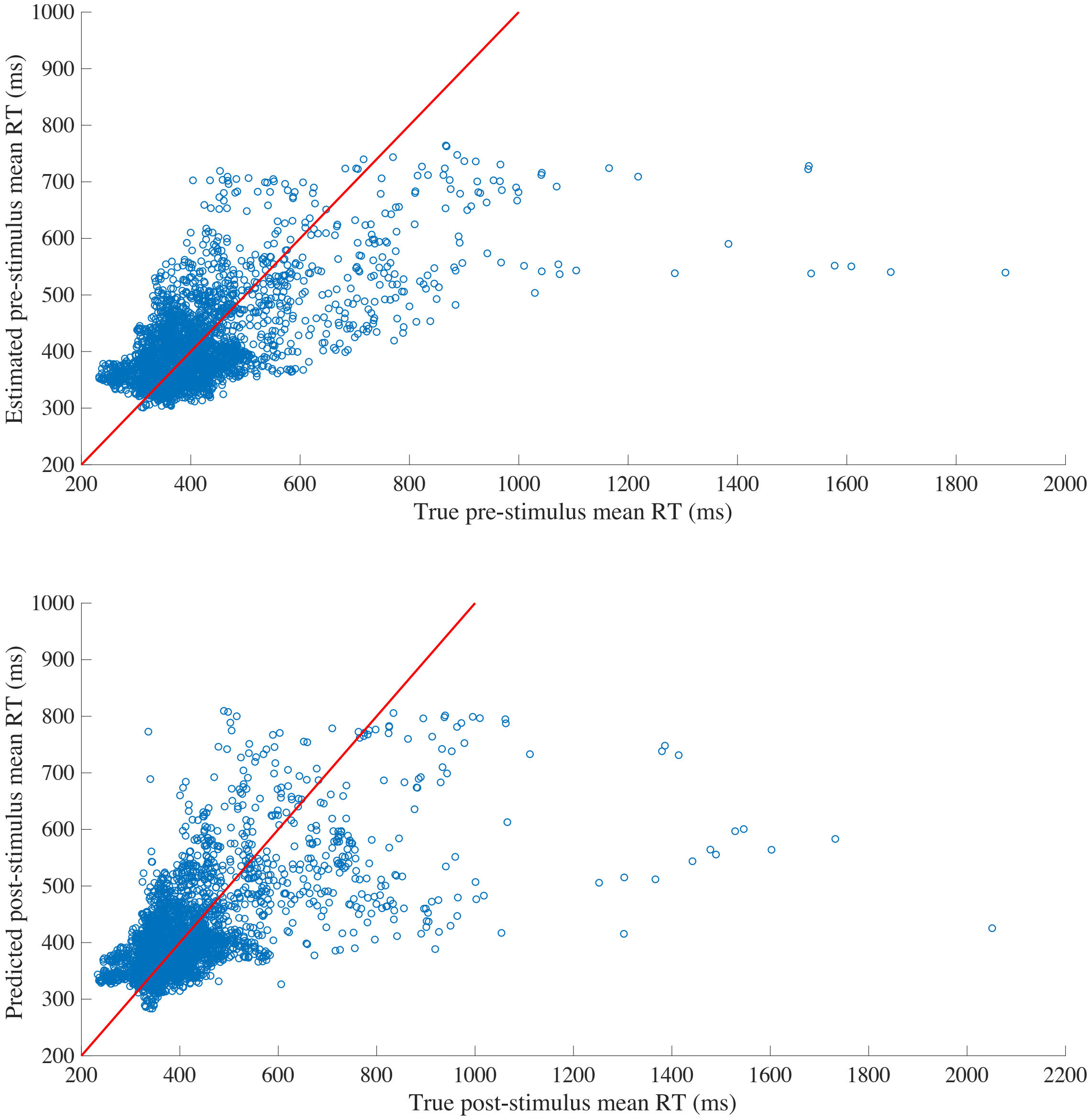


Figure 4.7 – Scatter plot of the estimated pre-stimulus mean RT (top) and predicted post-stimulus mean RT (bottom) produced with an RBF kernel vs their true value. The red line is the perfect regressor. We observe similar distributions of errors in both graphs, with the large errors at high values of mean RT.

Types of ocular features

Table 4.3 shows the importance scores of the 15 types of ocular features for the four problems, and it highlights the five most-important features for drowsiness (in red) and the two most-important features for alertness (in blue). We observe that $H_{[0.9,1]}$, i.e., the proportion of normalized eyelids distances within $[0.9, 1]$, has a large negative importance score for each problem. To a lesser extent, $H_{[0.7,0.8]}$ has also a negative importance score for each problem. Overall, we observe that $N_{\mu sleeps}$, $\overline{D}_{blinks}$, $\overline{D}_{opening}$, $\overline{D}_{closed}$, and $H_{[0,0.1]}$ are the most-important ocular features for drowsiness. These observations are coherent with the literature since drowsiness is characterized by slower and longer blinks, and thus by reduced times of eyes openness.

		Regression		**Classification**		**Normalized**
		Estim.	**Pred.**	**Estim.**	**Pred.**	**sum**
Ocular features	$\overline{D}_{blink}$	2261.3	2523	2.9487	3.3997	0.8541
	$\overline{D}_{closing}$	−496.4	−485.7	1.4725	1.7917	0.1084
	$\overline{D}_{closed}$	2005.5	2827.9	1.4419	2.9961	0.6698
	$\overline{D}_{opening}$	2505.6	2353.0	2.6904	1.8874	0.8209
	$N_{\mu sleeps}$	2508.3	2121.6	3.6297	3.6669	0.9003
	$H_{[0,0.1]}$	965.8	1099.8	−2.8398	3.0365	0.5643
	$H_{[0.1,0.2]}$	647.7	457.9	−3.5860	2.8229	0.5080
	$H_{[0.2,0.3]}$	−980.8	−118.7	2.0314	1.9696	0.1428
	$H_{[0.3,0.4]}$	442.0	150.1	1.2400	1.3297	0.2173
	$H_{[0.4,0.5]}$	−24.8	−253.1	1.3936	0.9592	0.1214
	$H_{[0.5,0.6]}$	1358.9	270.8	1.3545	1.1048	0.3121
	$H_{[0.6,0.7]}$	−431.6	−59.2	0.0839	−0.0442	−0.0449
	$H_{[0.7,0.8]}$	−522.9	−259.5	−0.6205	−0.3818	−0.1386
	$H_{[0.8,0.9]}$	703.8	724.6	0.0596	0.7764	0.1876
	$H_{[0.9,1]}$	−684.6	−787.9	−8.7020	−8.6480	−1.2234

Table 4.3 – Importance scores of the ocular features for the four problems. The last column contains the sum of importance scores that were normalized, i.e., scaled such that the sum over each of the four columns equals one. We highlight the five most-important features related to drowsiness in red, and the two most-important features related to alertness in blue.

4.4.4 Correlation of standard ocular features with standard drowsiness measures

For the development of our baseline system, we only considered the mean RT as the ground truth of drowsiness. Therefore, it is interesting to perform a statistical analysis of several standard ocular features (produced in a similar manner than with our baseline system) with different standard measures of drowsiness (recorded in our sleep-deprivation dataset).

We consider the following seven ocular features: $\overline{D}_{blink}$, $\overline{D}_{closing}$, $\overline{D}_{closed}$, $\overline{D}_{opening}$, $N_{\mu sleeps}$, N_{blinks} (the number of blinks), and $PERCLOS$ (the percentage of eye closure, where an eye closure is defined as a normalized eyelids distance below 0.7). Note that we did not use N_{blinks} in our baseline system, and that $PERCLOS$ is equivalent to a linear combination of the histogram features, i.e., $PERCLOS = H_{[0,0.7]}$. The time window, over which these ocular features are computed, depends on the time resolution of the standard measure of drowsiness, with which these ocular features are compared to.

We consider the following four measures of drowsiness: (1) the KSS score, with one value per 10-min PVT; (2) the KDS score, with one value per 20-s epoch, available only for four subjects; (3) the mean RT, with one value per 1-min epoch; and (4) the number of lapses, with one value per minute 1-min epoch. Therefore, the ocular features compared with (1), (2), (3), and (4) are computed over a window with a length of 10min, 20s, 1min, and 1min, respectively.

Table 4.4 shows the PCC of each ocular feature with each drowsiness measure. The number of blinks, N_{blinks}, is weakly correlated to the KDS score (Pearson correlation coefficient, PCC, of 0.39), but uncorrelated to the other drowsiness measures (PCCs of 0.19, 0.05, and 0.07). Indeed, we noticed during the development of our baseline system that N_{blinks} is more a function of the subject than a function of the level of drowsiness. This could explain the correlation of N_{blinks} with the KDS score since it is computed from four subjects. The other ocular features are well correlated to the drowsiness measures, where the most-correlated is $\overline{D}_{blink}$ (average PCC of 0.65) followed by $PERCLOS$ and $N_{\mu sleeps}$ and $\overline{D}_{opening}$ (0.59), then by $\overline{D}_{closed}$ (0.56), and then by $\overline{D}_{closing}$ (0.49).

	Ocular features						
	$\overline{D}_{blink}$	$\overline{D}_{closing}$	$\overline{D}_{closed}$	$\overline{D}_{opening}$	N_{blinks}	$N_{\mu sleeps}$	$PERCLOS$
KSS score	0.59	0.54	0.46	0.67	0.19	0.51	0.53
KDS score	0.7	0.46	0.64	0.6	0.39	0.61	0.65
Mean RT	0.69	0.48	0.64	0.54	0.05	0.66	0.58
Nb. of lapses	0.6	0.48	0.5	0.56	0.07	0.59	0.59
Average	0.65	0.49	0.56	0.59	0.17	0.59	0.59

Table 4.4 – Pearson correlation coefficients (PCC) between 7 standard ocular features (listed on the second line) and 4 standard measures of drowsiness (listed in the first column).

Figure 4.8 shows the distribution of $\overline{D}_{blink}$ and $PERCLOS$ as a function of the KSS score, the KDS score, and the number of lapses. Overall, we observe that these two ocular features increase when the measures of drowsiness increase, although not always in a close-to-linear fashion.

4.5 Conclusion

In this chapter, we presented a baseline drowsiness characterization system. Our baseline system processes a 1-min sequence of face images with four successive modules, extracts standard ocular features, and characterizes drowsiness in four different ways. Compared to the systems that we present in the following chapters, the baseline system uses a person-specific “face landmarks” module. Therefore, the baseline system is not thoroughly generic and automatic. However, it allows us to study the relationship between eye closure dynamics and performance impairments without worrying about the performance of the computer-vision modules of our system. In the next chapter, with the goal of comparing the performance of our baseline system with those of our other systems, we make modifications to the baseline system which make it generic and automatic.

We evaluated our baseline system in controlled, laboratory conditions on 14 subjects via a leave-one-subject-out cross-validation. The results show that the estimation performance of estimative problems are higher that the performance of predictive problems. For the estimative regression problem, the baseline system achieves a Pearson correlation coefficient (PCC) of 0.62 and a Root Mean Square Error (RMSE) of 104ms. For the estimative

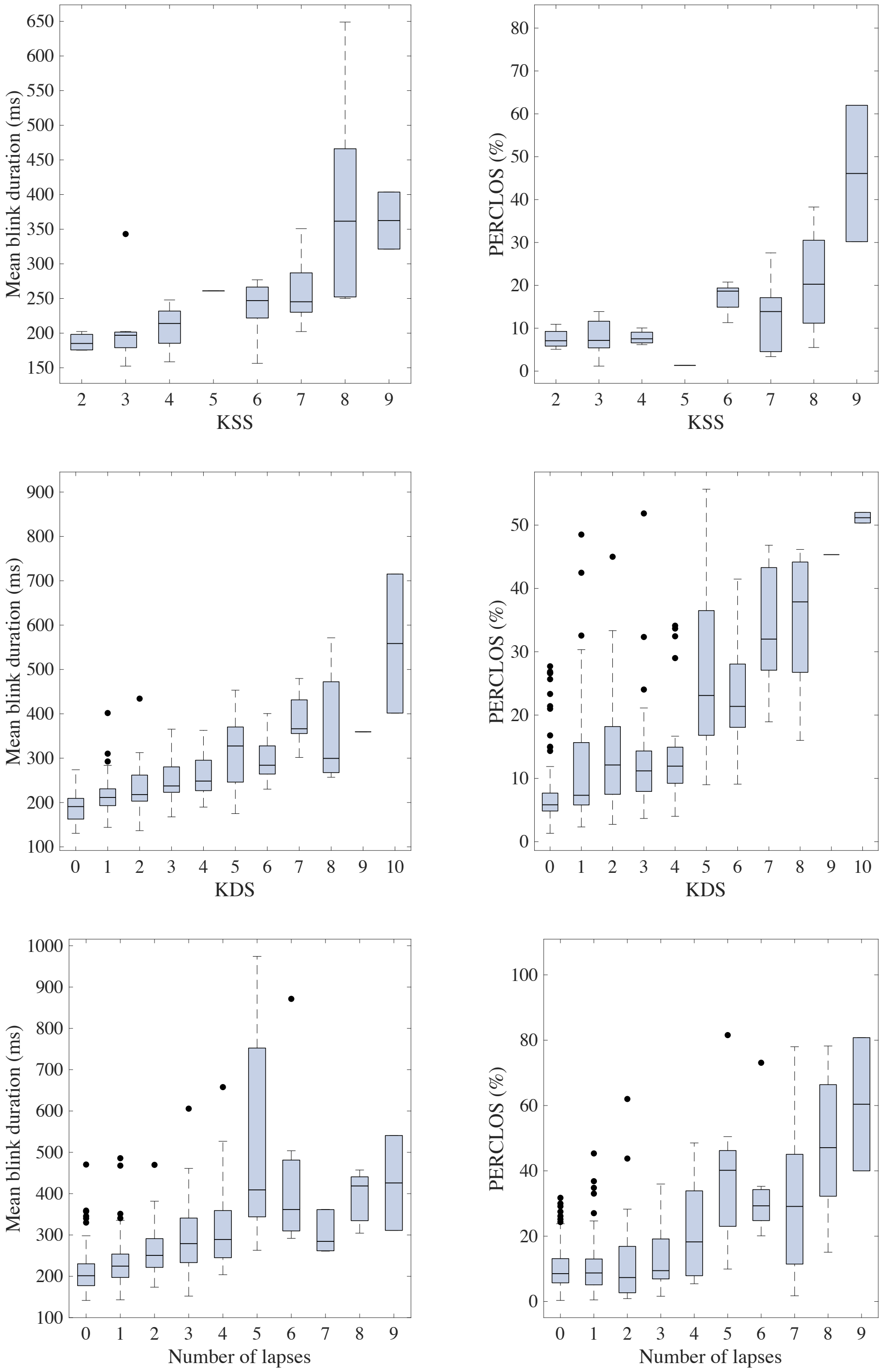


Figure 4.8 – Box plots of the mean blink duration (left column) and the PERCLOS (right column) as a function of KSS scores (first row), KDS scores (second row), and the number of lapses (third row). From left to right and top to bottom, the PCCs are 0.59, 0.53, 0.7, 0.69, 0.6, and 0.59.

classification problem, the baseline system achieves a true detection of alertness of 87.73%, a true detection of drowsiness of 75.46%, and a global accuracy of 86.19%.

We found that the number of microsleeps ($N_{\mu sleeps}$), the average blink duration ($\overline{D}_{blinks}$), the average opening duration ($\overline{D}_{opening}$), the average closed duration ($\overline{D}_{closed}$), and the proportion of normalized eyelids distance below 10% ($H_{[0,0.1]}$) are the most-important ocular features related to impairments of performance due to drowsiness. When comparing to other standard measures of drowsiness, we found the following ocular features to be the most correlated with drowsiness: $\overline{D}_{blink}$ (average PCC of 0.65) followed by $PERCLOS$ and $N_{\mu sleeps}$ and $\overline{D}_{opening}$ (0.59), then by $\overline{D}_{closed}$ (0.56), and then by $\overline{D}_{closing}$ (0.49).

Chapter 5

Multi-timescale drowsiness characterization system

This chapter presents a multi-timescale drowsiness characterization system that aims at dealing with the trade-off between accuracy and responsiveness. Section 5.1 introduces and motivates our multi-timescale system. Section 5.2 describes our system. Section 5.3 details the training of our system. Section 5.4 reports experimental results, and evaluates the performance. Section 5.5 investigates the combination of the binary LoDs into a single LoD, which is more convenient to use operationally. Section 5.6 concludes this chapter. This chapter is based on the following published journal article [93]: Q. Massoz, J. Verly, and M. Van Droogenbroeck. Multi-timescale drowsiness characterization based on a video of a driver's face. Sensors, 18(9):1–17, August 2018.

5.1 Introduction

In the scientific literature, drowsiness characterization systems typically make use of eye closure dynamics by averaging blink-related features (e.g., blink duration) over a time window of fixed length (e.g., one minute). However, systems using this strategy suffer from a trade-off between accuracy and responsiveness. Indeed, a system based on a short time window (of eye closure dynamics) will be very responsive to sudden changes in eye closure dynamics and, therefore, to brief episodes of drowsiness such as lapses and microsleeps, but it will not characterize drowsiness with high accuracy. By contrast, a system based on a long time window will be more accurate, but less responsive. Ideally, drowsiness characterization systems should be both accurate and responsive.

With the goal of satisfying both accuracy and responsiveness, we present a novel multi-timescale drowsiness characterization system that is data-driven, automatic, real-time, and generic. Our system extracts, via convolutional neural networks (CNNs), data-driven features related to eye closure dynamics at four timescales, i.e., four time windows of increasing lengths (5s, 15s, 30s, and 60s) and all extending up to the present, so as to infer four binary Levels of Drowsiness (LoDs). We design a novel multi-timescale ground truth of drowsiness in such a manner that (1) an LoD inferred at a low timescale is an early and responsive, but noisy estimate of drowsiness, and (2) an LoD inferred at a high timescale is an accurate, but less responsive estimate of drowsiness. To obtain such multi-timescale ground truth, we produce four binary ground-truth LoDs (one per inferred LoD) based on the median values—computed over time windows of increasing lengths—of the reaction times (RTs) performed during standard Psychomotor Vigilance Tasks (PVTs). In such a manner, our system produces, from any 1-min sequence of face images, four binary LoDs

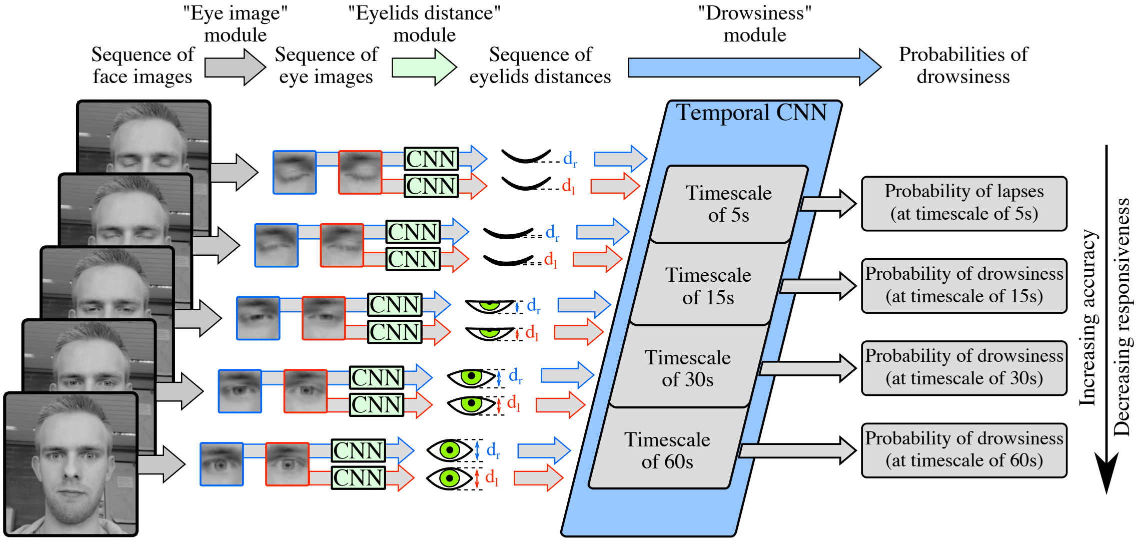

"Eye image" module
"Eyelids distance" module
"Drowsiness" module
Sequence of face images
Sequence of eye images
Sequence of eyelids distances
Probabilities of drowsiness
Temporal CNN
CNN
d_r
d_l
Timescale of 5s
Timescale of 15s
Timescale of 30s
Timescale of 60s
Probability of lapses (at timescale of 5s)
Probability of drowsiness (at timescale of 15s)
Probability of drowsiness (at timescale of 30s)
Probability of drowsiness (at timescale of 60s)
Increasing accuracy
Decreasing responsiveness

Figure 5.1 – Overview of the multi-timescale drowsiness characterization system operating on any given 1-min sequence of face images. First, from each face image, the “eye image” module produces two eye images (left and right) via off-the-shelf algorithms. Second, from each eye image, the “eyelids distance” module produces the eyelids distance via a convolution neural network (CNN). Third, from the 1-min sequence of eyelids distances and via a temporal CNN, the “drowsiness” module (1) extracts data-driven ocular features at four timescales, i.e., for the four most-recent time windows with increasing lengths of 5s, 15s, 30s, and 60s, and (2) produces four probabilities of drowsiness of increasing accuracy, but decreasing responsiveness.

with diverse trade-offs between accuracy and responsiveness.

5.2 Multi-timescale system

Our multi-timescale drowsiness characterization system is composed of three modules operating in cascade: the “eye image” module, the “eyelids distance” module, and the “drowsiness” module. Figure 5.1 depicts the multi-timescale system and its three modules.

5.2.1 “Eye image” module

The “eye image” module is composed of off-the-shelf algorithms and extracts, for each frame and for each eye, an eye image of size 24 × 24 pixels, this in four successive steps. First, we detect the face region using the OpenCV [20] implementation of the Viola and Jones algorithm [138]. Second, within the detected face region, we localize 68 face landmarks using the dlib [82] implementation of the Kazemi and Sullivan algorithm [81]. Third, from the 12 eyelids landmarks, we compute the eye center positions of the right and left eye, c_r and c_l, respectively, and the rotation angle needed to align them horizontally, α. Fourth (and last), we extract the right and left eye images using affine warping so as to obtain a right (respectively left) eye image centered on c_r (respectively c_l), rotated by an angle of α around c_r (respectively c_l), scaled at 24% of the face region width (from the first step), and with size of 24 × 24 pixels. Figure 5.2 depicts the extraction of both eye images.

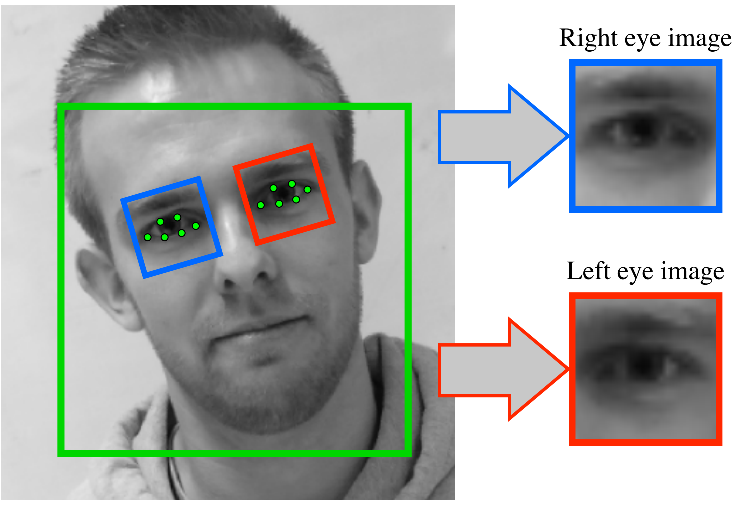


Figure 5.2 – Illustration of the steps of the "eye image" module. In succession, we detect the face (green square), we align the eyelids landmarks (green dots), and we geometrically extract the right eye image (blue square) and the left eye image (red square) with a common size of 24 × 24 pixels.

5.2.2 "Eyelids distance" module

The "eyelids distance" module is a spatial CNN taking, as input, a grayscale eye image, and producing, as output, an estimate of the eyelids distance (i.e., a real number) in pixels (referenced in the eye image, not in the original frame). The architecture of the module is very similar to the VGGNet architecture [126]. We provide a short technical background on CNNs in Appendix C.

The eye image is sequentially processed by (1) eight 3 × 3 convolutional layers (stride of 1, padding of 2, depths of 32, 32, 64, 64, 128, 128, 256, and 256, respectively, followed by the ReLU non-linearity then batch normalization [74]) interspersed with three 2 × 2 max pooling layers (padding of 2) positioned after every two convolutional layers, (2) a global max pooling, and (3) a fully connected layer (1 output neuron) so as to output the eyelids distance, i.e., a real number. Figure 5.3 depicts the architecture of the "eyelids distance" module.

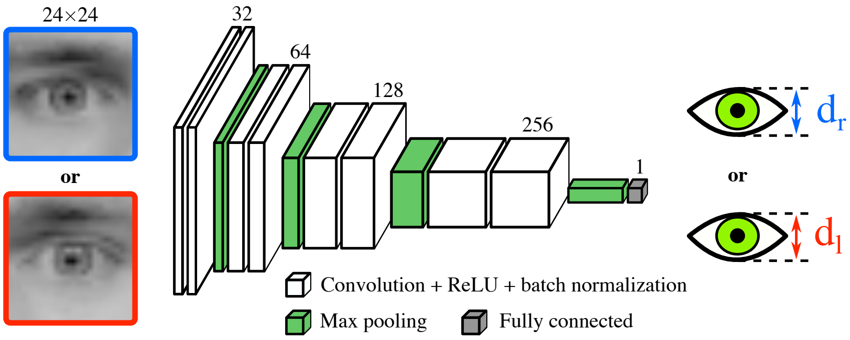


Figure 5.3 – Architecture of the "eyelids distance" module. The CNN produces an estimate of the right (or left, respectively) eyelids distance (i.e., a real number) from the right (or left, respectively) eye image of size 24 × 24 pixels. Note that one can process both eye images simultaneously in a batch of size 2.

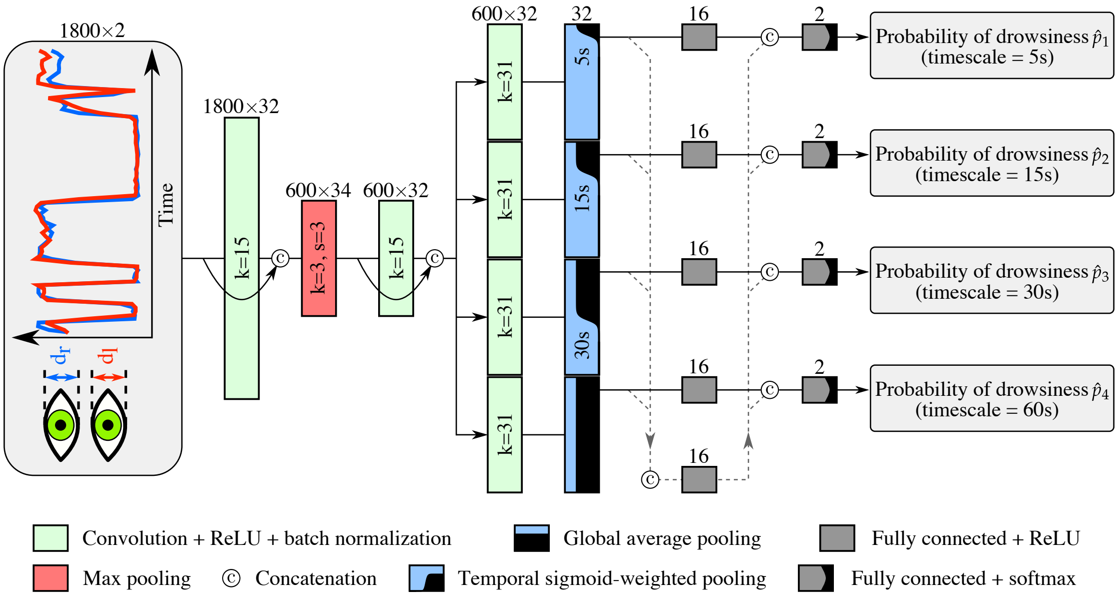

1800×2
Time
dr
dl
1800×32
k=15
600×34
k=3, s=3
600×32
k=15
600×32
k=31
k=31
k=31
k=31
32
5s
15s
30s
16
16
16
16
16
2
2
2
2
Probability of drowsiness $\hat{p}_1$ (timescale = 5s)
Probability of drowsiness $\hat{p}_2$ (timescale = 15s)
Probability of drowsiness $\hat{p}_3$ (timescale = 30s)
Probability of drowsiness $\hat{p}_4$ (timescale = 60s)
Convolution + ReLU + batch normalization
Global average pooling
Fully connected + ReLU
Max pooling
Concatenation
Temporal sigmoid-weighted pooling
Fully connected + softmax

Figure 5.4 – Architecture of the “drowsiness” module. The temporal CNN processes a 1-min sequence of eyelids distances using multiple time windows extending up to the present (via global pooling) to characterize drowsiness at multiple timescales.

5.2.3 “Drowsiness” module

The “drowsiness” module is a temporal CNN taking, as input, a 1-min sequence of eyelids distances related to both eyes (1800 × 2 values, at a framerate of 30 frames per second), and producing, as output, four binary LoDs associated to four timescales, i.e., 5s, 15s, 30s, and 60s. The processing is depicted in Figure 5.4, and is as follows.

First, the module processes the input sequence with two temporal convolutional layers (depth of 32, receptive field of 15, stride of 1, padding of 7, followed by ReLU then batch normalization) separated by a max pooling layer (receptive field k of 3, and stride s of 3). These two convolutional layers are densely connected [71], meaning that their outputs are concatenated with their inputs via a skip connection, leading to output sequences with dimensions of 34 and 66, respectively.

Second, the module forwards the resulting sequence (with depth of 66) to four branches, each tasked to produce one of the four estimated probabilities of drowsiness $\hat{p}_i$. Each branch consists of (1) a temporal convolutional layer (depth of 32, receptive field k of 31, stride s of 1, padding of 15, followed by ReLU then batch normalization, and without skip connection), (2) a global pooling layer (different for each branch, see below), (3) a first fully connected layer (depth of 16, and followed by ReLU), and (4) a last fully connected layer (depth of 2) followed by the softmax function.

Because the ground-truth LoD varies rapidly in time at a low timescale (see Section 5.3.2), the estimation of drowsiness should be mostly based on a short time window so as to be responsive to sudden changes in the eye closure dynamics. Therefore, the global pooling of the first three branches (timescales of 5s, 15s, and 30s) focus their attention over the recent past of varying length n_0 (of 5s, 15s, and 30s, respectively) via a “temporal sigmoid-weighted pooling” layer, represented in Figure 5.4, and defined as

$$\mathbf{a}(n_0) = \sum_{n=1}^{600} \frac{\sigma\left(\frac{3}{2}\left(n - 600 + 10n_0\right)\right)}{\sum_{k=1}^{600} \sigma\left(\frac{3}{2}\left(k - 600 + 10n_0\right)\right)} \mathbf{v}_n, \tag{5.1}$$

where $\mathbf{a}$ is the output feature vector, $\mathbf{v}_n$ is the feature vector at the nth position in the input sequence, $\sigma(x)$ is the sigmoid function expressed as $(1 + e^{-x})^{-1}$, and n_0 is the cut-off time (expressed in seconds) of the attention weights. We chose the sigmoid function to have the temporal weights decrease sharply, yet smoothly, at n_0. The global pooling of the fourth branch (timescale of 60s) corresponds to a global average pooling.

Furthermore, we add what we call "multi-timescale context" to each branch: the outputs of the global pooling layer of each branch are concatenated together, processed by a fully connected layer (depth of 16, and followed by ReLU), and then concatenated back into each branch with the output of their respective first fully connected layer. This is equivalent to adding dependencies between the branches, which we will show to be crucial to obtain strong performance for estimating drowsiness at low timescales.

5.3 Training of the system

We trained the "eyelids distance" module and the "drowsiness" module sequentially.

5.3.1 "Eyelids distance" module

Dataset

We built the "eyelids distance" dataset for training and evaluating the performance of the "eyelids distance" module. This dataset consists of the Multi-PIE (MPIE) face dataset [59] augmented with a subset of near-infrared face images (834) from our sleep-deprivation dataset (denoted SDD). We chose the MPIE dataset because of its variety in subjects, illumination conditions, head poses (from frontal to near-profile head poses), and types of eyeglasses (when present).

For each face image (of both sub-datasets), we extracted two eye images, i.e., one for each eye, by making use of the 68 manually-annotated face landmarks. For each eye image, we computed the ground-truth eyelids distance (i.e., the regression target) as the average of the two inter-eyelid Euclidean distances (referenced in the eye image) between the two face landmarks positioned on the upper eyelid, and the two on the lower eyelid.

Training and optimization

We split the "eyelids distance" dataset into a training set, a validation set, and a test set intended for training the model parameters, validating its hyper-parameters (via random search), and evaluating its performance, respectively. Table 5.1 contains the number of subjects and samples in these three sets, and from each of the two sub-datasets (MPIE or SDD). We randomly split the subjects so that the training, validation, and test sets have (1) an approximate ratio of 70/10/20 for both the numbers of subjects and samples, and (2) no overlap in subjects between them. We doubled the amounts of training, validation, and test data by flipping horizontally each eye image.

We trained the "eyelids distance" module with the Mean Squared Error (MSE) loss function using the RMSProp [133] optimization routine with a smoothing constant α of 0.9886, a batch size of 32, and a learning rate of 0.001428. We normalized the eye images by subtracting the average pixel value computed from the training set. We doubled the

		MPIE	**SDD**	**Total**
Number of subjects	**Training set**	242	11	253
	Validation set	28	2	31
	Test set	67	3	70
Number of samples	**Training set**	6438	1090	7528
	Validation set	794	182	976
	Test set	1924	396	2320

Table 5.1 – Numbers of subjects and samples in the training, validation, and test sets, and from each sub-datasets (MPIE and SDD) of the “eyelids distance” dataset. For each set, horizontal flipping of every eye image doubles the number of samples contained in this table.

number of samples of the training, validation, and test sets by horizontally flipping every eye images. We performed no other data augmentation.

5.3.2 “Drowsiness” module

Dataset

Out of the 88 PVTs of our sleep-deprivation dataset, we only used 82 PVTs (from 29 subjects, 18 females and 11 males) for the development of the multi-timescale system. The reason is that the PVT1 data, which are necessary for the inter-subject normalization of the RTs, were missing for 3 subjects.

Inter-subject normalization of the reaction times (RTs)

While performing a PVT, the RT achieved by a subject depends on various factors including drowsiness, time-on-task (i.e., fatigue), and individual skills. Drowsiness is the state that we wish to characterize, time-on-task is considered to have minor impact given the short PVT duration of 10 minutes, and individual skills can be mitigated by inter-subject normalization. Considering that the reciprocal of the RT (i.e., the reaction speed) of an individual follows relatively well a normal distribution [24], we normalize each RT from each subject according to

$$x' = \left(\frac{1}{x} - \mu_k + \frac{1}{29} \sum_{i=1}^{29} \mu_i \right)^{-1}, \tag{5.2}$$

where k is the subject index, x is an observed RT from subject k, x' is the corresponding normalized RT for subject k, and μ_k is the mean of the reciprocal of all RTs recorded during PVT1 of subject k. This normalization shifts the RT distribution of a subject in an alert state (i.e., in the first morning, during PVT1) to the population average (estimated from the 29 subjects).

Multi-timescale ground truth of drowsiness

We want to develop a system that operates both at long timescales (leading to accurate estimation of drowsiness) and at short timescales (leading to responsive estimation of drowsiness). Therefore, we need to produce the appropriate ground-truth LoDs of increasing accuracy and of decreasing responsiveness. Given that drowsiness is characterized by impairments of performance, i.e., overall longer RTs while performing a PVT, a ground-truth LoD could be generated by thresholding either (1) a single RT, which is perfectly time-localized (resulting in a responsive, but noisy estimate of drowsiness) or (2) a metric computed from a set of RTs within a time window (resulting in a more accurate, but less responsive estimate of drowsiness).

Accordingly, we define four metrics of performance, which we call "median RTs", denoted by m_i with $i \in \{1, 2, 3, 4\}$. The first median RT, m_1, corresponds to a single RT that either (1) occurs within the $[-1\text{s}, +1\text{s}]$ time window or (2) is a linear interpolation between the previous RT and the next RT. The other median RTs, m_2, m_3, and m_4, are computed as the harmonic means (equivalent to the medians of the reciprocal normal distributions) of the RTs that occur within the $[-15\text{s}, +5\text{s}]$, $[-30\text{s}, +5\text{s}]$, and $[-60\text{s}, +5\text{s}]$ time windows, respectively. Each median RT can be considered as being a continuous signal that varies in time at a specific timescale, induced by its corresponding sliding time window (i.e., with a specific length). These time windows are allowed to be non-causal since they are used for producing the ground-truth LoDs, and thus not for operational use.

By thresholding these four median RTs, we obtain four binary ground-truth LoDs, each varying at a distinct timescale, and each associated with a ground-truth likelihood score of drowsiness (loosely referred to as a probability of drowsiness from here on), denoted by p_i and defined as

$$p_i = \begin{cases} 0 & \text{if } m_i \leq 400\text{ms} \\ 0.5 & \text{if } m_i \in]400, 500[\text{ ms} \\ 1 & \text{if } m_i \geq 500\text{ms}. \end{cases} \text{, for each } i \in \{1, 2, 3, 4\} \tag{5.3}$$

The above thresholds of 400ms and 500ms were chosen empirically, yet pertinently. Indeed, the threshold of 400ms corresponds to about the 98–99th percentile of the distribution of m_4 during PVT1 (i.e., in non-sleep deprived conditions), whereas the threshold of 500ms corresponds to the value above which a RT (such as m_1) is conventionally interpreted as a lapse [13, 44]. From here on, each ground-truth LoD is referenced either by its index (from 1 to 4), or by the timescale at which the classifier estimating it operates (i.e., 5s, 15s, 30s, and 60s, respectively).

Loss function

We trained the "drowsiness" module with the average of four binary relative entropies, each associated with one of the four probabilities of drowsiness. The loss function is given for one sample by

$$L(\hat{\mathbf{p}}, \mathbf{p}) = \frac{-1}{4} \sum_{i=1}^{4} \left[p_i \ln\left(\frac{\hat{p}_i}{p_i}\right) + (1 - p_i) \ln\left(\frac{1 - \hat{p}_i}{1 - p_i}\right) \right], \tag{5.4}$$

where $\hat{p}_i$ is the ith estimated probability of drowsiness produced by the "drowsiness" module, and p_i is the ith ground-truth probability of drowsiness defined in Equation 5.3.

Training and optimization

Given the limited number of subjects (29), we trained 29 models following a leave-one-subject-out cross-validation strategy of 29 folds. Each fold consists of a training set of 23 subjects, a validation set of 5 subjects, and a test set of 1 subject. Moreover, each subject appears in an equal number of folds (23, 5, and 1, respectively) for each of the three sets, and with no overlap in subjects between sets of the same fold. The "eye image" module and the “eyelids distance” module were shared across folds. The samples (i.e., 1-min sequences of face images) composing each set are obtained as follows.

For the training set, we adopted a stratified random sampling strategy, where each training epoch consists of an equal number (256) of 1-min sequences randomly drawn from each of five groups (also known as strata). All possible 1-min sequences (of the training set, at a frame level) were divided into five strata based on the number of their four median RTs (noted m_i) that are greater than or equal to 470ms, with this number ranging from 0 to 4 for the five strata, respectively.

For the validation set and test set, we sampled the 1-min sequences that end at the occurrence time of every PVT stimulus (except for the PVT stimuli that occurred within the first minute of the PVT). In this way, the first ground-truth LoD is perfectly time-synchronized with the 1-min sequence. This deterministic sampling strategy leads to an average of about 85 samples per PVT.

We validated the hyper-parameters via random search so as to minimize the average validation loss across the 29 folds. Moreover, while we balanced the training sets (at an epoch level) via stratified random sampling, we balanced the validation sets (across folds) by weighting each sample in the ith relative entropy loss function (i.e., the ith term of the sum constituting the loss function in Equation 5.4) based on whether the median RT m_i (of the sample) is lower or greater than 470ms. This results in eight weights (two per timescale, shared across folds) with values that equal half of the reciprocal of the occurrence frequencies at a specific timescale (indexed by i), and across folds. Table 5.2 shows the eight computed occurrence frequencies, and the eight resulting weight values.

Timescale index i	**Occurrence frequency f (%)**		**Weight value $w = f^{-1}/2$**	
	$m_i < 470$ms	$m_i \geq 470$ms	$m_i < 470$ms	$m_i \geq 470$ms
1	87.29	12.71	0.5728	3.9318
2	92.97	7.03	0.5378	7.1173
3	94.79	5.21	0.5275	9.5810
4	95.75	4.25	0.5222	11.7821

Table 5.2 – The computed occurrence frequencies, f, and resulting weight values, w, that are used to balance the alert/drowsy samples in the average validation loss (across the 29 folds). Note that both f and w are functions of (1) the timescale index i, and (2) whether the median RT m_i is lower or greater than 470 ms.

We trained the 29 models (one per fold) using the Adam [83] optimization routine with a first moment coefficient of 0.9, a second moment coefficient of 0.999, a batch size of 32, and a learning rate of 0.0016029. We used dropout [128] with probabilities of 0.35, 0.7, and 0.35 respectively at three positions: (1) right after the concatenation of the second convolutional layer, (2) right after each global pooling layer, and (3) right before each last fully connected layer. Independently for each fold, we normalized the eyelids distances by subtracting the average eyelids distance computed from the training set. We augmented the data by randomly swapping (with a probability of 0.5) the right and left sequences of eyelids distances.

5.4 Experimental results and performance

5.4.1 "Eye image" module

We evaluated the performance of the "eye image" module on the held-out test set of the "eyelids distance" dataset. We computed the Root Mean Square Error (RMSE) between (1) the true eye positions obtained from the manually-annotated eye landmarks, and (2) the estimated eye positions obtained from the eyelids landmarks of the "eye image" module. We discarded samples with large errors (distances above 80 pixels) in estimated eye positions, i.e., when the algorithm did not converge. The reason is that, when processing a sequence of face images, we can easily detect such large errors (e.g., with a threshold on the variation in eye positions), and then estimate better eye positions by either interpolating or extrapolating them from the eye positions of other frames.

Following this evaluation scheme, we obtained an RMSE of 1.2 pixels, which is low enough for the eye to be always entirely contained within the eye image produced by the "eye image" module.

5.4.2 "Eyelids distance" module

We evaluated the performance of the "eyelids distance" module on the held-out test set composed of 4640 eye images from 70 subjects, and obtained an RMSE of 0.523 pixel. Figure 5.5 shows a scatter plot of the estimated eyelids distances versus their ground-truth value. We observe that the absolute error remains below 2, 1, and 0.5 pixel(s) for 99.9%, 93.1%, and 70.5% of the test samples, respectively.

For purposes of comparison, we also produced the eyelids distances directly from the eyelids landmarks localized by the "eye image" module, scaled them to be referenced in the coordinates of the eye image (rather than those of the face image), and obtained an RMSE of 1.152 pixels on the same held-out test set. This significant difference of a $1.152/0.523 = 2.2$ factor in performance clearly motivates the use of a specialized module, i.e., the "eyelids distance" module, for producing the eyelids distances.

Indeed, face alignment techniques, such as the one used in the "eye image" module, aim at localizing landmarks positioned on the entire face, rather than only those positioned on the eyelids. Because of this, the localization of eyelids landmarks significantly depends on the positions of other landmarks. This inter-landmark dependency is crucial for good coarse localization of the eyelids landmarks, but limits the fine localization of these landmarks since these are few in number (about $\sim 20\%$ of all face landmarks). On the contrary, the "eyelids distance" module aims at directly producing an estimate of the eyelids distance from the eye image, which can be efficiently carried out with a CNN.

5.4.3 "Drowsiness" module

We evaluated the performance of the "drowsiness" module by aggregating the results of the 29 test sets, each associated to one trained model, before computing the performance metrics. We did not average the performance metrics across the 29 subjects because (1) the amount of data was not identical for all subjects (some PVTs were missing), and (2) the proportion of fast/slow RTs varied significantly between subjects.

In addition, we discarded, at each timescale i independently, the samples with a ground-truth probability of drowsiness p_i of 0.5. That is, we only kept the samples of which the median RT m_i is below 400ms ($p_i = 0$, the sample is labeled as alert, the negative class), or above 500ms ($p_i = 1$, the sample is labeled as drowsy, the positive class). This discarding resulted, for the 1st, 2nd, 3rd, and 4th timescales respectively, in aggregated (across folds)

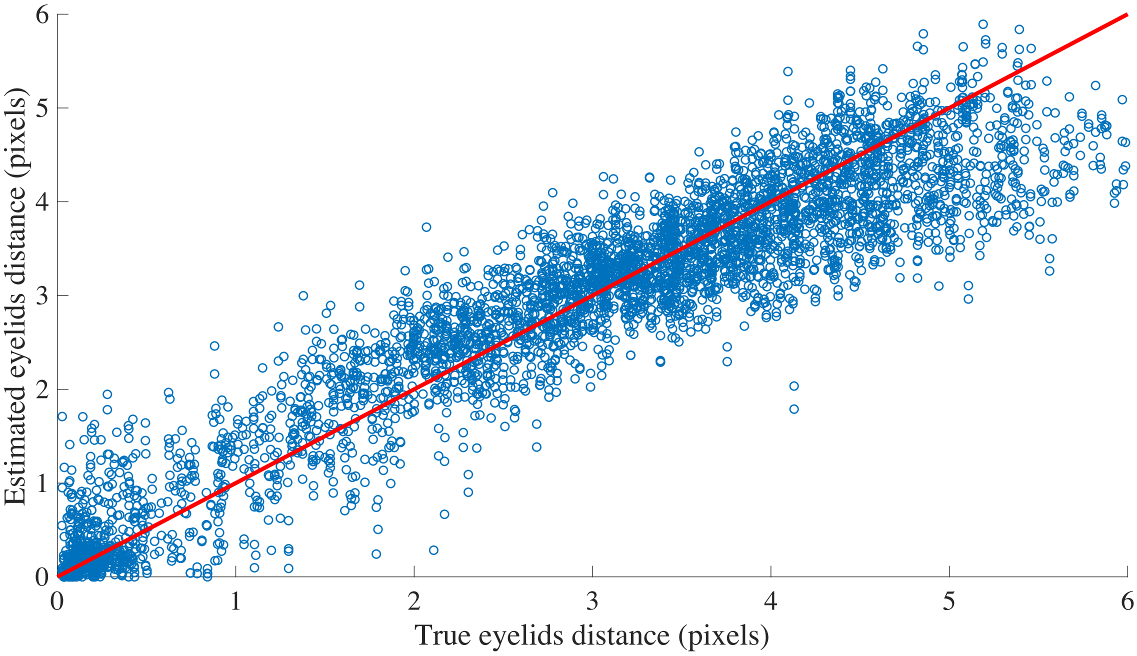


Figure 5.5 – Scatter plot of the estimated eyelids distance produced by our "eyelids distance" module vs. its ground-truth value (both expressed in pixels). The red line is the perfect regressor.

System	Timescale (s)	TNR (%)	TPR (%)	Accuracy (%)
Multi-timescale system	5	72.26	58.69	70.68
	15	86.29	71.84	85.45
	30	90.44	75.76	89.82
	60	94.80	74.19	94.22

Table 5.3 – Classification performance of the multi-timescale system. The negative class corresponds to the "alert" label, and the positive class to the "drowsy" label.

numbers of alert/drowsy (i.e., negative/positive) samples of 4845/639, 5100/316, 5221/231, and 5345/155, respectively.

The obtained results are shown in Table 5.3. The multi-timescale system achieved, for the 1st, 2nd, 3rd, and 4th timescales respectively, a specificity (i.e., true negative rate, TNR) of 72.26%, 89.29%, 90.44%, and 94.80%; a sensitivity (i.e., true positive rate, TPR) of 58.69%, 71.84%, 75.76%, and 74.19%; and a global accuracy of 70.69%, 85.45%, 89.82%, and 94.22%. Overall, we observe that all performance metrics increase with the timescale at which the LoD is inferred. The most significant increase in accuracy (of 14.77%) is found from the 1st timescale to the 2nd timescale. These results could be explained by the fact that, as the timescale increases, the characterization of drowsiness becomes less challenging because (1) the associated ground-truth LoD estimates more accurately the level of drowsiness, and (2) the data-driven features (related to eye closure dynamics) becomes less noisy as they are averaged over a longer time window.

5.4.4 Processing times

We evaluated the processing time of each module on a computer equipped with a Nvidia GeForce GTX TITAN X (Maxwell architecture) and an Intel i7-6700. The "eye image" module processes one video frame in 12ms. The "eyelids distance" module processes one pair of eye images (i.e., the left one and the right one) in 1.2ms. The "drowsiness" module processes an 1-min sequence of eyelids distances in 2.5ms, 13ms, or 62ms when using 1, 6, or 29 models, respectively. Note that, although the "eye image" module and the "eyelids distance" module have to be applied at each and every new frame, i.e., at 30 times per second, the "drowsiness" module can be applied at a lower rate, e.g., at 10 times per second. In this way, real-time constraints can be satisfied with an adjustable, comfortable margin.

5.4.5 Impact on performance of the "multi-timescale context"

We study the impact on performance of the "multi-timescale context" (defined in Section 5.2.3) by training, validating the hyper-parameters, and evaluating the 29 models without this context, i.e., by removing the auxiliary branch that is concatenated into each of the four main branches. We doubled the depth of the first fully connected layer to compensate for the reduced number of parameters.

The results in Table 5.4 show that the accuracy significantly drops at the 1st timescale (from 70.68% to 61.94%) accompanied with an increase in sensitivity (from 58.69% to 65.26%), and that the sensitivity drops at the 2nd, 3rd, and 4th timescales (by 3.49%, 7.79%, and 3.87%, respectively). This mostly shows that the context (of eye closure dynamics) from the higher timescales is crucial for good performance at the lower timescales. This makes sense since a single long blink is more probably associated with a lapse of attention if the driver has been experiencing long blinks for the last minute than if he has not.

System	Timescale (s)	TNR (%)	TPR (%)	Accuracy (%)
Multi-timescale system with "multi-timescale context"	5	72.26	58.69	70.68
	15	86.29	71.84	85.45
	30	90.44	75.76	89.82
	60	94.80	74.19	94.22
Multi-timescale system without "multi-timescale context"	5	61.51	65.26	61.94
	15	82.94	68.35	82.09
	30	91.02	67.97	90.04
	60	94.78	70.32	94.09

Table 5.4 – Comparison of the performance of the system with and without "multi-timescale context".

5.4.6 Comparison of performance with the baseline system

As seen in Chapter 2, the fair comparison between systems of different studies is infeasible. The principal reason is that the data and the ground truth of drowsiness differ between studies. To provide comparisons that are as fair as possible, we compare the performance of our multi-timescale system with our baseline system (presented in Chapter 4), which is representative of a large panel of systems of other studies. We modified, re-trained, and re-evaluated the baseline system to provide a fair comparison with our multi-timescale system. Details follow.

Modifications to the baseline system

We made modifications to each module of the baseline system.

1. We replaced the “face landmarks” and “eyelids distance” modules of our baseline system with the “eye image” and “eyelids distance” modules of our multi-timescale system. In this way, the baseline system (1) processes the same sequence of eyelids distances as the one produced by our multi-timescale system, and (2) is subject-generic, i.e., can be applied to each of the 29 subjects.

2. We modified the “ocular features” module of our baseline system to extract six standard ocular features from four time windows with lengths of 5s, 15s, 30s, and 60s, resulting in a vector of 24 ocular features. We used the six following standard ocular features: the mean blink duration, $\overline{D}_{blink}$; the mean closing duration, $\overline{D}_{closing}$; the mean closed duration, $\overline{D}_{closed}$; the mean opening duration, $\overline{D}_{opening}$; the number of microsleeps, $N_{\mu sleeps}$; and the percentage of eye closure below 70%, $PERCLOS$. In this way, the baseline system uses (1) ocular features typically found in other studies, and (2) the same time windows as the ones of our multi-timescale system.

3. We modified the “drowsiness” module of our baseline system such that it is composed of four SVM classifiers (one per timescale). We considered both a linear kernel and a radial basis function (RBF) kernel. Note that, by feeding, as input, ocular features computed from four time windows, each SVM characterizes drowsiness with “multi-timescale context”.

Training of the baseline system

We trained each SVM, i.e., each timescale, separately. At each timescale, we trained 29 models following a leave-one-subject-out cross-validation strategy of 29 folds. However, considering the significantly faster training time of SVMs compared to CNNs, we validated the hyper-parameters (C for the linear kernel, C and γ for the RBF kernel) via an inner leave-one-subject-out cross-validation strategy of 28 folds, i.e., all subjects (29) but the one (1) in the test set of the outer cross-validation. Upon determination of the optimal values of hyper-parameters, we trained the final model on all 28 subjects of the training set (of the outer cross-validation).

We obtained all samples of the training, validation, and test sets in the same manner, i.e., by sampling the 1-min sequences that end at the occurrence time of every PVT stimulus (except for the PVT stimuli that occurred within the first minute of the PVT). We discarded samples with a ground-truth probability of drowsiness p_i of 0.5, for all three sets and at each timescale i independently (as in Section 5.4.3). We individually scaled each feature so as to be within the range $[0, 1]$ for the samples of the training set. We weighted the classes (i.e., “alert” and “drowsy”) in the SVM optimization routine with the reciprocal of the number of their occurrence in the training set. We performed training and inference with the LIBLINEAR library [47] and the LIBSVM library [26] for the linear kernel and the RBF kernel, respectively. We performed no data augmentation.

Evaluation of the performance of the baseline system

We evaluated the performance of the baseline system by aggregating the results of the 29 test sets, each associated to one trained model, before computing the performance metrics. The obtained results are shown in Table 5.5, with a comparison with our multi-timescale system. With a linear kernel, our baseline system achieved, for the 1st, 2nd, 3rd, and 4th timescales respectively, a specificity (i.e., TNR) of 64.43%, 78.71%, 81.06%, and 84.49%; a sensitivity (i.e., TPR) of 61.03%, 65.81%, 60.34%, 64.52%; and a global accuracy of 64.03%, 77.97%, 80.18%, and 83.93%. With an RBF kernel, our baseline system achieved,

System	Timescale (s)	TNR (%)	TPR (%)	Accuracy (%)
Multi-timescale system	5	72.26	58.69	70.68
	15	86.29	71.84	85.45
	30	90.44	75.76	89.82
	60	94.80	74.19	94.22
Baseline system (linear kernel)	5	64.43%	61.03%	64.03%
	15	78.71%	65.81%	77.97%
	30	81.06%	60.34%	80.18%
	60	84.49%	64.52%	83.93%
Baseline system (RBF kernel)	5	75.56%	53.83%	73.03%
	15	85.74%	69.36%	84.80%
	30	87.71%	74.14%	87.14%
	60	88.61%	82.58%	88.44%

Table 5.5 – Comparison of performance between the multi-timescale system and the baseline system (with a linear kernel or an RBF kernel).

for the 1st, 2nd, 3rd, and 4th timescales respectively, a TNR of 75.56%, 85.74%, 87.71%, and 88.61%; a TPR of 53.83%, 69.36%, 74.14%, and 82.58%; and a global accuracy of 73.03%, 84.80%, 87.14%, and 88.44%.

Compared to our multi-timescale system, the baseline system has a greater specificity (+3.3%) and a greater global accuracy (+2.35%) at the first timescale, and a greater sensitivity (+8.39%) at the fourth timescale. On all the other performance metrics, the multi-timescale system outperforms the baseline system, which demonstrates the appropriateness of using a temporal CNN architecture to process a sequence of eyelids distances so as to characterize drowsiness.

5.5 Combination of multi-timescale decisions

Up to now, we attained the above results and observations by considering the four binary LoDs individually. When considered together, the four LoDs have 2^4 (16) possible outcomes. Interestingly, whereas the (combined) ground-truth LoD takes its value from all of the 16 possible outcomes, the (combined) inferred LoD takes its value only from 5 outcomes: “0000”, “1000”,”1100”, “1110”, and “1111”. This means, that if the system detects drowsiness at one timescale (e.g., 30s), it will consequently detect drowsiness at all lower timescales (e.g., 5s and 15s). As a corollary, it also means that the detection of drowsiness at one timescale (e.g., 5s) will happen before (or, at worst, at the same time) than the detections at higher timescales (e.g., 15s and above).

This suggests that the system has “learned” some form of internal timescale hierarchy as a result of the fact that we have trained the four classifiers together. However, it is also possible that this behavior of the system simply stems from the built-in hierarchy of the time windows (of 5s, 15s, 30s, and 60s) at the global pooling stage of the “drowsiness” module.

One could thus build a unified classifier by adding the binary decisions of each classifier so as to output a combined LoD ranging from 0 to 4 (with the lower levels being more responsive, and the higher ones more accurate). In real-world applications, one can conveniently feed such combined LoD back to the driver and/or to a semi-autonomous driving system. Indeed, when the (combined) LoD reaches 1, the driver would take notice early that he/she might be starting to be drowsy. At this time, the driver should determine the plausibility of drowsiness by answering whether he/she has been driving for a long time, and whether he/she had enough sleep. When the LoD reaches 2–3, drowsiness becomes more and more probable, and the driver can start taking early safety actions. When the LoD reaches 4, drowsiness is most probable, and the driver would have had enough time to decide the best safety actions to take, such as pulling to the nearest rest area to switch drivers, take a 15-min nap, and/or consume a caffeinated beverage [70]. Note that, whereas a driver may become too drowsy to take any safety actions, a semi-autonomous driving system would always be ready to take the actions necessary to prevent any accidents, including autonomously bringing the vehicle to the nearest rest area.

5.6 Conclusion

In this chapter, we presented a multi-timescale drowsiness characterization system that is novel, data-driven, automatic, real-time, and generic. Our multi-timescale system processes a 1-min sequence of face images with three successive modules, extracts data-driven features related to eye closure dynamics at distinct timescales (5s, 15s, 30s, and 60s), and outputs four binary LoDs with diverse trade-offs between accuracy and responsiveness. To train our system, we introduced a multi-timescale ground truth of drowsiness that consists of four ground-truth LoDs based on thresholded, normalized median RTs computed from time windows of different lengths.

We evaluated our multi-timescale system in controlled, laboratory conditions on 29 subjects via a leave-one-subject-out cross-validation. The results show that the system achieves overall strong performance, with the highest performance (specificity of 94.80%, and sensitivity of 74.19%) at the 4th timescale (of 60s). We showed that the system outperforms our baseline system based on a vector of multi-timescale, standard ocular features being fed to timescale-specific SVMs, which is representative of a wide range of systems found in other studies.

In real-world applications, the driver (or a monitoring system and/or a semi-autonomous driving system) could combine these four estimated LoDs (of increasing accuracy, and of decreasing responsiveness) to assess the driver's physiological state of drowsiness, and then decide—with full knowledge—to take safety actions.

Chapter 6

Parametric drowsiness characterization system

This chapter presents a parametric drowsiness characterization system that aims at estimating the parameters of the instantaneous probability density function of drowsiness-induced reaction times. Section 6.1 introduces and motivates our parametric system. Section 6.2 describes our system. Section 6.3 details the training of our system. Section 6.4 reports experimental results, and evaluates the performance. Section 6.5 provides a visual interpretation, and a first analytic analysis, of the data-driven features learned by our system, and related to eye closure dynamics. Section 6.6 concludes this article. This chapter is based on the following published conference paper [92]: Q. Massoz and J. Verly. Vision-based system for monitoring vehicle operator responsiveness from face images. In International Conference on Managing Fatigue, pages 1–3, San Diego, CA, USA, March 2017.

6.1 Introduction

In the two previous chapters, we defined the ground truth (of drowsiness) in the terms of the mean/median reaction time (RT) computed over a time window. For our baseline system, the ground truth is (1) the mean RT computed over a 1-min window for regression problems, or (2) the thresholded mean RT computed over a 1-min window for classification problems. For our multi-timescale system, the multi-timescale ground truth is four thresholded median RTs computed over four windows with lengths of 5s, 15s, 30s, and 60s.

In other words, we defined a function that maps a set of recent RTs into a continuous or discrete ground-truth quantity. Such mapping functions are non-injective, i.e., different sets of RTs can be mapped to the same ground-truth quantity. For instance, let's consider the following two sets of RTs: (A) $\{330, 350, 370, 750\}$ms and (B) $\{430, 440, 450, 480\}$ms. These sets have both a mean RT of 450ms. However, these sets are most probably not sampled from the same distribution. Indeed, one would expect the probability of observing a RT of 500ms to be higher for the underlying distribution from which B is sampled than for the distribution from which A is sampled. Therefore, mapping the set of RTs to a single quantity results in a loss of information.

In this chapter, we make directly use of the set of observed RTs to train a parametric system that estimates the underlying probability density function (pdf) of these drowsiness-induced RTs. Mathematically, the RT, x, is an observation of a random process, $X(\omega, t) : \Omega, T \subset \mathbb{R} \to \mathbb{R}^+$. For a fixed sample path ω, this process is a function of time, where the set of RTs observed during a 10-min Psychomotor Vigilance Task (PVT) is a trajectory of this process. For a fixed time t, this process is a random variable associated with an instantaneous pdf, $f_X^t(x)$. This instantaneous pdf is what we are interested in as it represents the distribution of how fast an individual may react to a sudden event happening at time t. However, in general settings such as operating a vehicle, $X(\omega, t)$ is non-stationary as $f_X^t(x)$ depends on many time-varying variables including the task's difficulty, the individual's skill, the level of distraction, and the level of drowsiness. The non-stationarity property renders the estimation of $f_X^t(x)$ from sparsely-observed RTs infeasible as we would need multiple RTs observed at the same time t.

Nevertheless, in the controlled settings of a PVT, $f_X^t(x)$ depends mainly on the level of drowsiness because (1) the task's difficulty is fixed, (2) the individual's skill can be normalized (e.g., with a baseline), (3) the level of distraction is fixed, and (4) impairments of performance are mostly driven by an increase in the level of drowsiness (induced by sleep deprivation). In such settings, because the level of drowsiness should not vary drastically in time, we assume $X(\omega, t)$ to be locally stationary [36], i.e., stationary in a reasonably small time segment. By also assuming ergodicity in the mean and variance, one can estimate $f_X^t(x)$ via stationary methods on the RTs observed in a small time segment centered around time t. Note that this estimation procedure causes a bias which depends on the degree of non-stationarity of the random process in the time segment, but that we assume to be negligible. In controlled PVT settings, it has been empirically shown that the distribution of RTs is well approximated by a reciprocal normal ("recinormal") distribution [24, 86], i.e., the reciprocal of the RT is normally distributed:

$$f_X^t(x) \approx \mathcal{R}_{[\mu,\sigma^2]}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\frac{1}{x}-\mu\right)^2}{2\sigma^2}\right), \tag{6.1}$$

where $\mathcal{R}_{[\mu,\sigma^2]}(x)$ is the recinormal pdf, and μ and σ^2 are respectively the mean and variance parameters.

With the goal of estimating the instantaneous pdf of drowsiness-induced RTs, we present a novel parametric drowsiness characterization system that is data-driven, automatic, real-time, and generic. Our system produces, via convolutional neural networks (CNNs), an estimate of the parameters of $f_X^t(x)$, $\hat{\mu}$ and $\hat{\sigma}^2$, based on eye closure dynamics extracted from a face video. To train such systems, we use the locally-observed, skill-normalized RTs as ground truth to maximize the likelihood of the two estimated parameters, $\hat{\mu}$ and $\hat{\sigma}^2$. In such a manner, our system produces, from any 1-min sequence of face images, an estimate of the parameters of the instantaneous recinormal pdf of drowsiness-induced RTs, $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}(x)$.

Note that, if used in general settings, our system would not produce an absolute estimate of the instantaneous pdf of RTs, but would instead produce an estimate of the instantaneous pdf of drowsiness-induced RTs that would be observed during a PVT performed in controlled settings. Estimating such pdf is still useful in general settings as one can use it as a proxy to indirectly estimate the instantaneous level of drowsiness, i.e., characterize drowsiness.

6.2 Parametric system

Our parametric drowsiness characterization system is composed of two successive modules operating in cascade: the "eyelids distance" module and the "drowsiness" module. The "eyelids distance" module estimates the eyelids distance from each face image. The "drowsiness" module estimates, from the most-recent 1-min sequence of eyelids distances, the two parameters of $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}(x)$, $\hat{\mu}$ and $\hat{\sigma}^2$. Figure 6.1 depicts the parametric system and

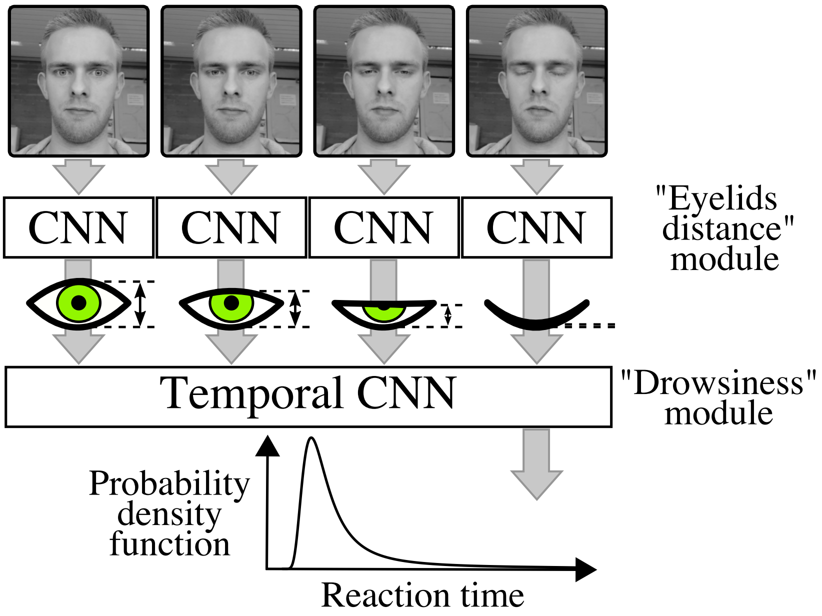


Figure 6.1 – Overview of our parametric drowsiness characterization system. First, the "eyelids distance" module estimates, for each frame and via a convolutional neural network (CNN), the eyelids distance (a real number expressed in pixels). Then, the "drowsiness" module estimates, via a temporal CNN, the parameters of the recinormal probability density function (pdf) of drowsiness-induced reaction times (RTs) based on the sequence of eyelids distances of the past minute.

its two modules. Note that the face image is extracted from the complete frame using the OpenCV [20] implementation of the Viola and Jones algorithm [138].

6.2.1 "Eyelids distance" module

For measuring the eyelids distance, the common approach consists to (1) locate face landmarks in 2D/3D via face alignment, and then (2) compute the Euclidean distance between the located upper and lower eyelid landmarks. As seen in Chapter 4, one may perform face alignment using parametric deformable models iteratively fitted with the Gauss-Newton algorithm [31, 121], iterative additive position updates learned by cascaded regression models [23, 81, 114, 149], or landmarks position heat-maps jointly generated by CNNs [22].

Our "eyelids distance" module uses a more straightforward way of measuring the eyelids distance: we estimate the eyelids distance directly from the face image using a spatial CNN. This approach has the advantages of not requiring the definition of a deformable shape model and of being fast (with the adequate hardware), and turned out to be reasonably robust to occlusions (including glasses and hands) and other facial expressions.

More precisely, the "eyelids distance" module is a spatial CNN taking, as input, a 128×128 grayscale face image, and producing, as output, an estimate of the eyelids distance (a real number expressed in pixels) of the most-opened visible eye. Focusing on the most-opened visible eye increases the system's robustness by allowing the module to (1) generate a single value per face; to (2) naturally deal with some common occlusions of the eyes, such as self-occlusions (induced by large head rotations); and to (3) ignore the cases where one eye blinks but not the other, which we consider to be the consequence of conscious behaviors and therefore not of interest to characterize drowsiness. Technically, we achieve

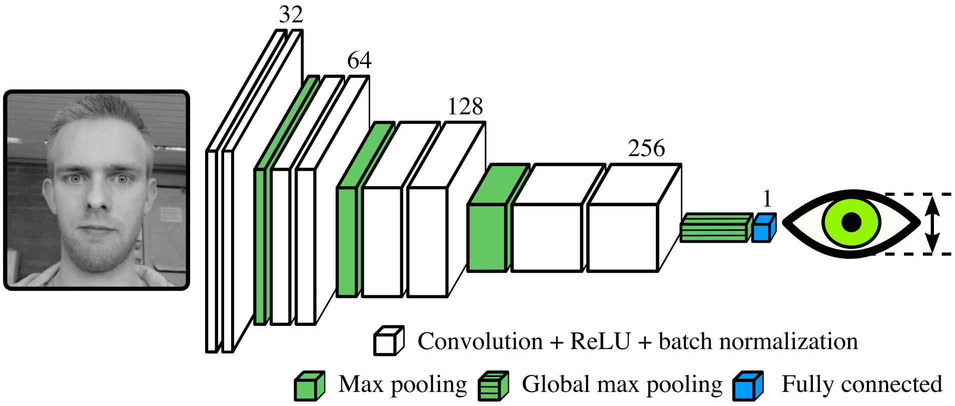


Figure 6.2 – Architecture of the "eyelids distance" module. The spatial CNN estimates, from a 128 × 128 grayscale face image, the eyelids distance, i.e., a real number expressed in pixels.

this focus on the most-opened eye by using a global max pooling in the model architecture, and by selecting the maximum (annotated) eyelids distance as the ground-truth target.

The model architecture is a stack of eight convolutional layers disposed in a similar fashion to the architecture of VGGNet [127]. Convolutional layers all have a receptive field of 3×3, a stride of 1, and a padding of 1. They are followed by the ReLU activation function then batch normalization [74]. Max pooling is performed after every two convolutional layers, over a 2×2 window and with a stride of 2. The depth of the convolutional layers is doubled after every max pooling, thus ranging from 32 for the first two layers to 256 for the last two layers. The model architecture ends with a global max pooling followed by a fully connected layer (1 output neuron). Figure 6.2 depicts this model architecture.

6.2.2 "Drowsiness" module

The "drowsiness" module consists of a temporal CNN taking, as input, a 1-min sequence of eyelids distances (that is 1800 values, at a framerate of 30 frames per second), and producing, as output, an estimate of the log-mean, $\ln(\hat{\mu})$, and log-variance, $\ln\left(\hat{\sigma}^2\right)$, parameterizing $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}(x)$. The reason why we chose the log-mean and log-variance as the two outputs is to enforce the positivity constraint on the mean and variance parameters.

The model architecture is composed of (1) a temporal convolution layer (depth of 32, receptive field of 15, a stride of 1, and a padding of 7), (2) a max pooling layer (receptive field of 3, and a stride of 3), (3) a temporal convolutional layer (depth of 32, receptive field of 31, a stride of 1, and a padding of 15), (4) a global average pooling layer, (5) a fully connected layer (32 neurons) followed by the ReLU activation function, and (7) a final fully connected layer (2 output neurons) that outputs $\ln(\hat{\mu})$ and $\ln\left(\hat{\sigma}^2\right)$. Both temporal convolutional layers are followed by the ReLU activation function then batch normalization. Figure 6.3 depicts this model architecture.

6.3 Training of the system

We trained the "eyelids distance" module and the "drowsiness" module sequentially.

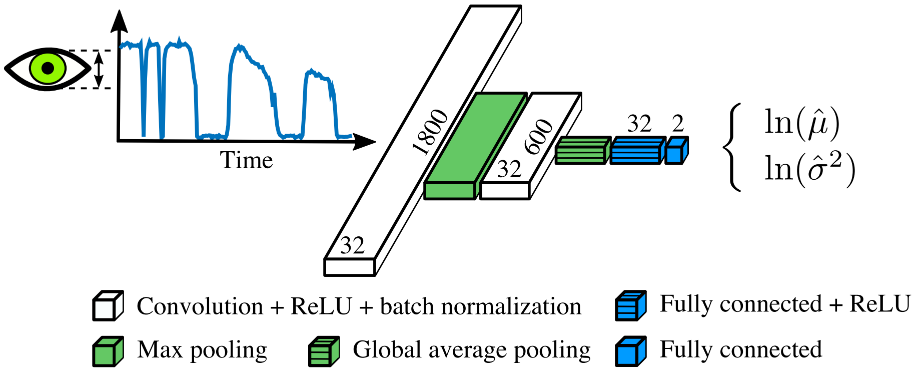


Figure 6.3 – Architecture of the "drowsiness" module. The temporal CNN estimates, from the most-recent 1-min sequence of eyelids distances, the log-mean and log-variance parameters of the recinormal distribution of the reactions times.

6.3.1 "Eyelids distance" module

Dataset

We built the "eyelids distance" dataset for training and evaluating the performance of the "eyelids distance" module. This dataset is composed of face images (i.e., inputs) and ground-truth eyelids distances (i.e., targets) aggregated from three separate datasets: (1) the CMU "MPIE" Face dataset [58] that is used for its variety in subjects and facial expressions, (2) the "PUT" Face dataset [80] used for its variety in head pose, and (3) a subset of near-infrared images from our sleep-deprivation dataset (denoted "SDD").

We produced the face images and ground-truth eyelids distances by making extended use of the provided, manually annotated face landmarks (with a number of 68, 30, and 68 face landmarks per image, respectively for MPIE, PUT, and SDD). More specifically, we automatically extracted, from each image of our "eyelids distance" dataset, a grayscale face image of size 128 × 128 pixels that is centered on the average position of the face landmarks (i.e., the face center), and scaled so that the vertical range of face landmarks (in most cases, between the chin landmarks and the eyebrows landmarks) equals 70 pixels. Likewise, we automatically produced the ground-truth eyelids distances by taking the maximum eyelids distance between the right eye and the left eye. For one eye, we computed the eyelids distance as the average of the two inter-eyelid Euclidean distances between the two face landmarks positioned on the upper eyelid, and the two on the lower eyelid. We discarded 6 and 705 images we found to be either erroneously or poorly annotated from the MPIE and PUT datasets, respectively.

Training and optimization

We split the "eyelids distance" dataset into a training set, a validation set, and a test set intended for training the model parameters, validating its hyper-parameters (via random search), and evaluating its performance, respectively. Table 6.1 contains the number of subjects and samples in these three sets, and from each of the three datasets (MPIE, PUT, and SDD). We randomly split the subjects so that the training, validation, and test sets have (1) an approximate ratio of 70/10/20 for both the numbers of subjects and samples, and (2) no overlap in subjects between them. We doubled the amounts of training, validation, and test data by a step of data augmentation consisting in flipping each face image horizontally.

		MPIE	PUT	SDD	Total
Number	**Training set**	242	69	11	322
of	**Validation set**	28	11	2	41
subjects	**Test set**	67	20	3	90
Number	**Training set**	3219	6380	545	10144
of	**Validation set**	397	1043	91	1531
samples	**Test set**	962	1843	198	3003

Table 6.1 – Numbers of subjects and samples in the training, validation, and test sets, and from each datasets (MPIE, PUT, and SDD) of the "eyelid distance" dataset. For each set, horizontal flipping of every face image doubles the number of samples contained in this table.

We optimized the hyper-parameters via random search. We trained the "eyelids distance" module via the Mean Squared Error (MSE) loss function using the RMSProp [133] optimization routine with a smoothing constant α of 0.9811, a batch size of 32, and a learning rate of 0.001195. We used dropout [128] with a probability of 0.5 after the global max pooling. We normalized the face images by subtracting the global mean pixel value of the face images computed from the training set. We augmented the training data by randomly rotating the training face images uniformly between -30 and 30 degrees, and this with different rotation angle for each sample at each epoch.

6.3.2 "Drowsiness" module

Dataset

Out of the 88 PVTs of our sleep-deprivation dataset, we only use 82 PVTs (from 29 subjects, 18 females and 11 males) for the development of the parametric system. The reason is that the PVT1 data, which are necessary for the inter-subject normalization of the RTs, are missing for 3 subjects.

Inter-subject normalization of the reaction times (RTs)

While performing a PVT, the instantaneous pdf, $f_X^t(x)$, depends on some time-varying variables, including the level of drowsiness, the time-on-task (i.e., fatigue), and the individual skill. Drowsiness is the state we wish to characterize, time-on-task is considered to have minor impact given the short PVT duration of 10 minutes, and individual skill can be mitigated by inter-subject normalization. Therefore, given that $f_X^t(x)$ is well approximated by a recinormal distribution, we normalize each RT from each subject according to

$$x' = \left(\frac{\frac{1}{x} - \mu_k}{\sigma_k} \sigma^* + \mu^* \right)^{-1}, \tag{6.2}$$

where k is the subject index, x is a RT observed from subject k, x' is the corresponding normalized RT for subject k, μ_k (resp. σ_k) is the "subject mean" (resp. "subject SD") defined as the mean (resp. SD) of the reciprocal of all RTs observed during PVT1 of subject k, and μ^* (resp. σ^*) is the "population mean" (resp. "population SD") defined as the average of all 29 "subject means" (resp. "subject SDs"). In the present study, we measured a "population mean", μ^*, of about $\frac{1}{346}$ ms^{-1} and a "population SD", σ^*, of about $\frac{1}{2228}$ ms^{-1}.

This inter-subject normalization aligns the normal pdf of the reciprocal RTs of each subject in an alert state (in the first morning, during PVT1) to the population average (estimated from 29 subjects). Note that, since (1) $f_X^t(x)$ is recinormal (i.e., $f_X^t(x^{-1})$ is normal) and (2) x'^{-1} is linear with respect to x^{-1}, the pdf of the normalized RTs, $f_X^t(x')$, is also recinormal (i.e., $f_X^t(x'^{-1})$ is normal) .

Instantaneous ground truth of drowsiness

We want to develop a system that automatically estimates the two parameters, $\hat{\mu}$ and $\hat{\sigma}^2$, of the instantaneous recinormal pdf of drowsiness-induced RTs, and this based on the recent eye closure dynamics. Because we assume the random process $X(\omega, t)$ from which the RTs are sampled is (1) locally stationary, i.e., stationary in a reasonably small time segment, and (2) ergodic in its mean and variance, we can use the RTs locally observed in time as an instantaneous ground truth of drowsiness. Accordingly, for a fixed time t, we define the "ground-truth set", S_{GT}, as the set of skill-normalized RTs that are locally observed within the $[t - 60s, t + 30s]$ time window: $S_{GT} = \{x' : x' \text{ occurred within } [t - 60\text{s}, t + 30\text{s}]\}$. The size of the ground-truth set, $|S_{GT}| = n$, varies with time since the inter-stimulus interval varies randomly between 2 and 10 seconds. On average, the ground-truth set is composed of $\overline{n} = 14$ RTs. Note that the ground-truth set is a non-causal set of recently-observed RTs. This is not a problem since the ground-truth set is only used for training the system, thus not for operational use.

Furthermore, we define the ground-truth mean parameter, μ_{GT}, and the ground-truth variance parameter, σ^2_{GT}, that are computed, respectively, as the mean and the variance of the reciprocal of the n RTs of S_{GT}:

$$\begin{cases} \mu_{GT} = \frac{1}{n} \sum\limits_{j=1}^{n} \frac{1}{x'_j} \\ \sigma^2_{GT} = \frac{1}{n-1} \sum\limits_{j=1}^{n} \left(\frac{1}{x'_j} - \mu_{GT} \right)^2, \end{cases} \tag{6.3}$$

where x'_j is the jth normalized RT in the ground-truth set.

Loss function

For training the "drowsiness" module, we need to optimize with respect to a loss function that measures how well/poorly $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}(x)$, i.e., the output of our system, fits the observations of the ground-truth set, i.e., $S_{\text{GT}} = \{x'_1, \ldots, x'_n\}$. There are mostly two approaches to do so:

1. minimize the MSE function between the estimated parameters of $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}(x)$ ($\hat{\mu}$ and $\hat{\sigma}^2$) and the parameters computed from the observations of S_{GT} (μ_{GT} and σ^2_{GT}),
2. maximize the likelihood function for the estimated parameters of $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}(x)$ given the observations of S_{GT}.

The first approach amounts to a regression of the two pdf parameters. Empirically, we found that the first approach tends to quickly overfit the training data, thereby leading to poor generalization performance. Therefore, we use the second approach. More specifically, we use a combination of two loss functions:

$$L = L_{NLL} + \lambda L_{BCE}, \tag{6.4}$$

where L_{NLL} is an average negative log-likelihood (NLL) loss function, L_{BCE} is a binary cross-entropy (BCE) loss function, and λ is a constant weight empirically set to 0.2. Details about the NLL function and the BCE function follow.

NLL: The NLL loss function is a regression loss that relates to the plausibility of the estimated distribution parameters, i.e., $\ln(\hat{\mu})$ and $\ln(\hat{\sigma}^2)$, given the observed RTs in the ground-truth set. By assuming the RTs in the ground-truth set independent and identically distributed, the likelihood function for the distribution parameters is expressed for one data sample as

$$\begin{aligned} \mathcal{L}\left(\hat{\mu},\hat{\sigma}^2 \mid S_{GT}\right) &= \prod_{j=1}^{n} \mathcal{L}\left(\hat{\mu},\hat{\sigma}^2 \mid x'_j\right) \\ &= \prod_{j=1}^{n} \mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}\left(x'_j\right) \\ &= \prod_{j=1}^{n} \tfrac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\tfrac{1}{2\hat{\sigma}^2}\left(\tfrac{1}{x'_j}-\hat{\mu}\right)^2\right) \\ &= \left(2\pi\hat{\sigma}^2\right)^{-\frac{n}{2}} \exp\left(-\tfrac{1}{2\hat{\sigma}^2}\sum_{j=1}^{n}\left(\tfrac{1}{x'_j}-\hat{\mu}\right)^2\right), \end{aligned} \tag{6.5}$$

where n is the size of the ground-truth set, x'_j is the jth normalized RT in S_{GT}. By taking the Napierian logarithm of equation (6.5), we obtain the log-likelihood function:

$$\ln\mathcal{L}\left(\hat{\mu},\hat{\sigma}^2 \mid S_{GT}\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\hat{\sigma}^2\right) - \frac{1}{2\hat{\sigma}^2}\sum_{j=1}^{n}\left(\frac{1}{x'_j}-\hat{\mu}\right)^2. \tag{6.6}$$

The NLL loss function is the average negative log-likelihood:

$$\begin{aligned} L_{NLL}\left(\hat{\mu},\hat{\sigma}^2,S_{GT}\right) &= -\tfrac{1}{n}\ln\mathcal{L}\left(\hat{\mu},\hat{\sigma}^2 \mid S_{GT}\right) \\ &= \tfrac{1}{2}\ln(2\pi) + \tfrac{1}{2}\ln\left(\hat{\sigma}^2\right) + \tfrac{1}{2n\hat{\sigma}^2}\sum_{j=1}^{n}\left(\tfrac{1}{x'_j}-\hat{\mu}\right)^2. \end{aligned} \tag{6.7}$$

Since the "drowsiness" module output the log-mean and log-variance, we rewrite equation (6.7) with a change of variables $a = \ln(\hat{\mu})$ and $b = \ln\left(\hat{\sigma}^2\right)$:

$$L_{NLL}(a,b,S_{GT}) = \frac{1}{2}\ln(2\pi) + \frac{b}{2} + \frac{\exp(-b)}{2n}\sum_{j=1}^{n}\left(\frac{1}{x'_j}-\exp(a)\right)^2. \tag{6.8}$$

The first partial derivatives of L_{NLL} *w.r.t.* a and b are

$$\frac{\partial L_{NLL}(a,b,S_{GT})}{\partial a} = -\frac{\exp(a-b)}{n}\sum_{j=1}^{n}\left(\frac{1}{x'_j}-\exp(a)\right), \tag{6.9}$$

$$\frac{\partial L_{NLL}(a,b,S_{GT})}{\partial b} = \frac{1}{2} - \frac{\exp(-b)}{2n}\sum_{j=1}^{n}\left(\frac{1}{x'_j}-\exp(a)\right)^2. \tag{6.10}$$

BCE: The BCE loss function is a classification loss that penalizes large errors, i.e., samples that are misclassified *w.r.t.* a confidence score of drowsiness computed from the mean parameter. For one data sample, L_{BCE} is expressed as

$$L_{BCE}(\hat{\mu},\mu_{GT}) = -s_d(\mu_{GT})\ln\big(s_d(\hat{\mu})\big) - (1-s_d(\mu_{GT}))\ln\big(1-s_d(\hat{\mu})\big), \tag{6.11}$$

where $s_d(\mu) = \text{sigmoid}(\alpha\ln(\mu)+\beta)$ is a function that maps a log-mean parameter, $\ln(\mu)$, to a confidence score of drowsiness, $\hat{\mu}$ is the estimated mean parameter produced by the "drowsiness" module, and μ_{GT} is the ground-truth mean parameter computed from S_{GT}. We set the parameters α and β such as to satisfy $s_d\left(\frac{1}{450ms}\right) = 0.5$ and $s_d\left(\frac{1}{500ms}\right) = 0.95$, resulting in $\alpha \approx -29.44$ and $\beta \approx -179.91$.

Training and optimization

Given the limited number of subjects (29), we trained 29 models following a leave-one-subject-out cross-validation strategy of 29 folds. For each fold, we randomly split the 29 subjects into a training set of 24 subjects, a validation set of 4 subjects, and a test set of 1 subject. Furthermore, we made sure that every subject appears, across folds, in only 1 test set, in 4 validation sets, and in 24 training sets. The samples, i.e., the 1-min sequences of eyelids distances and their associated ground-truth set of RTs, composing each set are obtained as follows.

For the training set, we adopted a stratified random sampling strategy where each training epoch consists of an equal number (1056) of 1-min sequences randomly drawn from each of three groups (also known as strata). The first stratum contains all the 1-min sequences (from the training set, at a frame level) with a reciprocal ground-truth mean parameter, μ_{GT}^{-1}, below 400ms (i.e., normal responsiveness), the second stratum those with one between 400ms and 500ms (i.e., slow responsiveness), and the third stratum those with one above 500ms (i.e., very slow responsiveness). Note that, given the recinormal distribution of the RTs, the reciprocal of the mean parameter is equivalent to the median RT.

For the validation set and the test set, we extracted 37 samples from each 10-min PVT using a deterministic sliding window strategy with a step of 15s.

We validated the hyper-parameters via random search so as to minimize the average validation loss across the 29 folds. We trained the 29 models (one per fold) using the Adam [83] optimization routine with a first moment coefficient β_1 of 0.9, a second moment coefficient β_2 of 0.999, a batch size of 32, no dropout, and a learning rate of 0.008745. Independently for each fold, we normalized the sequences of eyelids distances by subtracting the average eyelids distance computed from the training set.

6.4 Experimental results and performance

6.4.1 "Eyelids distance" module

We evaluated the performance of the "eyelids distance" module on the held-out test set composed of 6006 samples from 90 subjects, and obtained a Root Mean Square Error (RMSE) of 0.582 pixel. Figure 6.4 shows a scatter plot of the 6006 estimated eyelids distances versus their ground-truth value. We observe that the eyelids distance ranges from 0 to 6 pixels for 96% of samples. Furthermore, we observe that the absolute error remains below 2, 1, and 0.5 pixel(s) for 99.7%, 91.5%, and 64.9% of the test samples, respectively.

For the purpose of comparison, we evaluated the performance of a face alignment algorithm. In particular, we used the dlib [82] implementation of the Kazemi and Sullivan algorithm [81], localized 68 face landmarks on each face image of the test set, and then geometrically produced each maximum eyelids distance from these localized face landmarks. This algorithm obtained an RMSE of 0.99 pixel, which is significantly higher than the one obtained by our "eyelids distance" module.

Figure 6.5 illustrates some input-output results, including six of the largest errors in the last row (noted e). Although the interpretation of CNNs is complex, results empirically appear unaffected by glasses, hands, facial expressions, and moderate head pose, thereby suggesting that the spatial CNN has learned to focus on the small regions of the eyes and to discard the superfluous information. Interestingly, some large error results (in row e) might be explained by the challenging conditions in the input. For instance, because of the

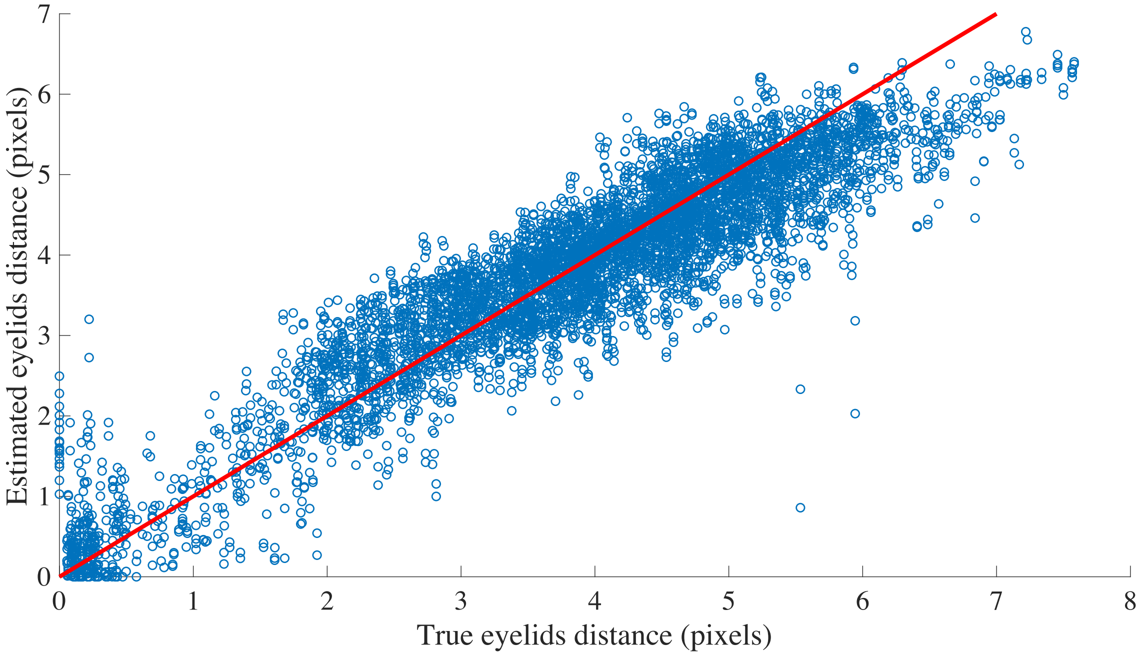


Figure 6.4 – Scatter plot of the estimated eyelids distance produced by our "eyelids distance" module vs its ground-truth value. The red line is the perfect regressor.

subject's long eyelashes, the closed eyes in e1 is visually similar to the downward look in d1, which could explain why our "eyelids distance" module incorrectly estimates an eyelids distance of ∼3 pixels instead of ∼0 pixel. Another, but less likely, possible explanation is that the CNN has learned that, when the head is tilted downwards, there is a higher probability that the subject is looking downward (i.e., small eyelids distance) than closing his/her eyes (i.e., minimum eyelids distance).

6.4.2 "Drowsiness" module

We evaluated the performance of the "drowsiness" module by aggregating the results of the 29 test sets, each associated to one trained model, before computing the performance metrics. We did not average the performance metrics across the 29 subjects because (1) the amount of data was not identical for all subjects (some PVTs were missing), and (2) the range of observed RTs varied greatly between subjects.

Figure 6.6 shows a scatter plot of the results aggregated from the 29 test sets, for both the estimated log-mean and log-variance parameters vs their corresponding ground-truth value computed from S_{GT}. Note that the reciprocal of the mean parameter is equivalent to the median RT, therefore a lower (log-)mean parameter corresponds to a higher median RT. We observe a Pearson correlation coefficient (PCC) of 0.52 between $\ln(\hat{\mu})$ and $\ln(\mu_{GT})$, and a PCC of 0.15 between $\ln\left(\hat{\sigma}^2\right)$ and $\ln\left(\sigma_{GT}^2\right)$. The log-mean parameter is thus well correlated with its ground-truth value, whereas the log-variance parameter is not. This lack of correlation for the log-variance parameter may be explained by two facts: (1) the low amount of RT observations contained in the ground-truth set (with an average of 14 RTs/S_{GT}), resulting in a noisy ground-truth variance, and (2) the absence of discriminative features in the temporal sequence of eyelids distances to produce a good estimate of the log-variance parameter. This is further evidenced by the fact that $\ln\left(\hat{\sigma}^2\right)$ is nearly constant, i.e., remains close to its average value of -15, whereas $\ln\left(\sigma_{GT}^2\right)$ ranges from -19 to -13.

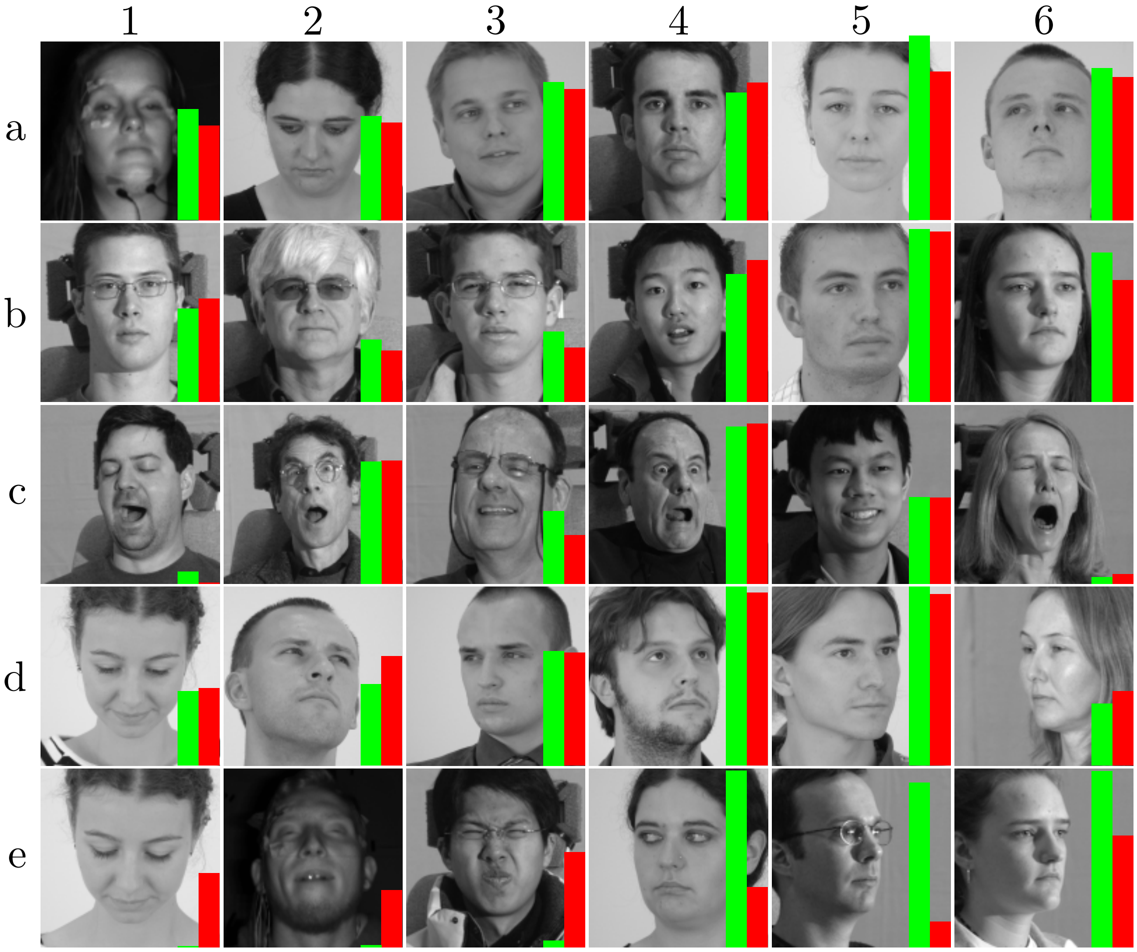


Figure 6.5 – Examples of input-output results of the "eyelids distance" module, sampled from the test set. The red bar represents the estimated eyelids distance, and the green bar the ground-truth value. The bar height is proportional to the eyelids distance; a bar height that equals the image height corresponds to an eyelids distance of 6 pixels. For ease of reference, we index the lines with letters and the columns with numbers.

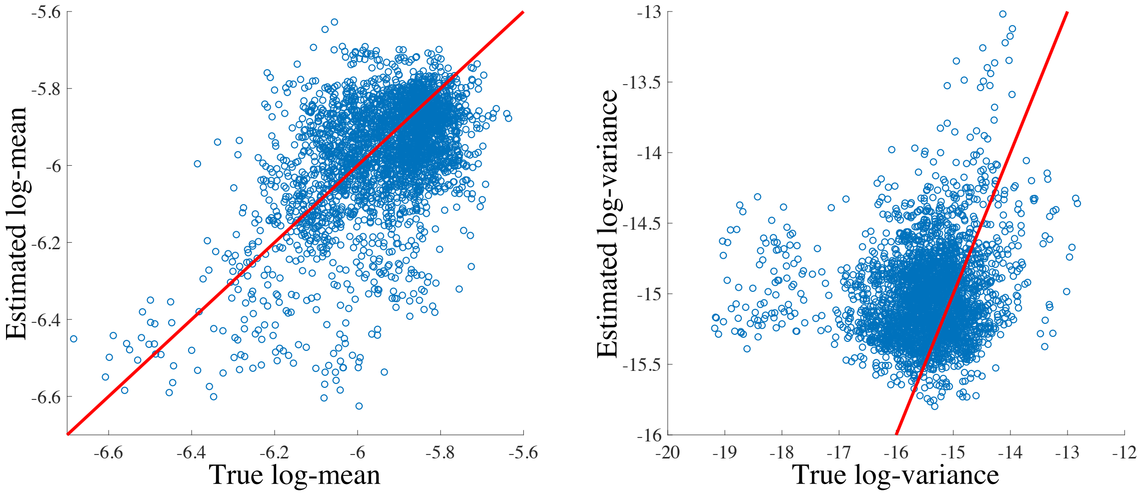


Figure 6.6 – Scatter plot of the estimated log-mean (left) and log-variance (right) parameters, $\ln(\hat{\mu})$ and $\ln\left(\hat{\sigma}^2\right)$, produced by the "drowsiness" module vs their ground-truth value, $\ln(\mu_{GT})$ and $\ln\left(\sigma^2_{GT}\right)$. The red lines are perfect regressors.

Intra-subject analysis

Note that the correlation coefficients reported above correspond to inter-subject correlations. It would therefore be interesting to analyze the intra-subject correlations. Table 6.2 contains the intra-subject PCCs for the log-mean and log-variance parameters, as well as the intra-subject range of median RTs ($= \mu_{GT}^{-1}$) observed over the three PVTs. Out of 29 subjects, we observe that 18 subjects (62%) have a PCC above 0.4 between $\ln(\hat{\mu})$ and $\ln(\mu_{GT})$, and that only 5 subjects (17%) have a PCC above 0.4 between $\ln\left(\hat{\sigma}^2\right)$ and $\ln\left(\sigma^2_{GT}\right)$. Let us focus our analysis on the log-mean parameter. We observe negative correlations (PCC < 0) for 5 subjects, and low correlations (PCC $\in [0, 0.4]$) for 6 subjects. Out of these 11 subjects, 6 subjects have relatively low maximum median RT (i.e., 410ms, 414ms, 375ms, 379ms, 373ms, and 423ms for subjects #3, #10, #12, #13, #25, and #26, respectively), which could explain their bad correlations, whereas 5 subjects have a maximum median RT close to 500ms (i.e., 506ms, 510ms, 507ms, 502ms, and 504ms for subjects #8, #15, #18, #21, and #28, respectively). Further explaining these linear intra-subject correlations is challenging as each can stem from the subject (1) sustaining minimal/moderate performance impairments from sleep deprivation, and/or (2) displaying eye closure dynamics that are not representative of the performance impairments.

Classification of the level of drowsiness

We can convert our parametric system into a binary classifier so as to produce a binary Level of Drowsiness (LoD). To do so, we consider the subject to be "drowsy" if the probability of his/her RT taking on a value above 500ms is greater than 0.5, i.e., $\int_{500}^{+\infty} \mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}(x)\,\mathrm{d}x \geq 0.5$, and we consider the subject to be "alert" otherwise. This is equivalent to thresholding the log-mean parameter: we consider the subject to be "drowsy" if $\ln(\hat{\mu}) \leq \ln\left(\frac{1}{500ms}\right) \approx -6.215$, and we consider the subject to be "alert" otherwise. We chose this particular threshold since a RT greater than 500ms is conventionally interpreted as a lapse [13, 44]. Following this procedure, the aggregated 29 test sets contained 2931 samples labeled as "alert" (negative class), and 103 samples labeled as "drowsy" (positive class).

The binary classifier achieved a specificity (i.e., true negative rate, TNR) of 95.19%, a

Subject	**PCC for**		**Median**
	log-mean	**log-var**	**RT (ms)**
1	0.746	0.147	311–450
2	0.632	0.189	320–401
3	−0.338	−0.081	317–410
4	0.825	0.574	325–593
5	0.864	0.692	330–658
6	0.870	0.392	309–734
7	0.891	0.291	323–516
8	0.286	0.089	318–506
9	0.634	−0.459	308–431
10	0.161	0.218	318–414
11	0.470	−0.051	323–367
12	−0.440	0.446	296–375
13	−0.592	0.115	280–379
14	0.598	0.199	324–800
15	0.104	0.183	307–510
16	0.701	0.165	316–498
17	0.701	0.116	323–465
18	0.361	0.346	313–507
19	0.459	0.138	298–384
20	0.614	0.397	304–476
21	−0.316	−0.221	326–502
22	0.599	0.484	331–491
23	0.414	0.008	324–460
24	0.720	0.606	320–593
25	0.265	0.231	308–373
26	0.046	0.203	313–423
27	0.527	0.133	299–384
28	−0.671	−0.131	326–504
29	0.409	−0.307	325–453
All	0.515	0.148	280–800

Table 6.2 – Intra-subject Pearson correlation coefficients (PCCs) between the estimated distribution parameters ($\ln(\hat{\mu})$ and $\ln\left(\hat{\sigma}^2\right)$) and the ground-truth distribution parameters ($\ln(\mu_{GT})$ and $\ln\left(\sigma^2_{GT}\right)$). To feed our analysis, we added (1) the intra-subject ranges of median RT ($= \mu_{GT}^{-1}$) observed over the three PVTs and (2) the inter-subject PCC and range of median RT ("All" line).

sensitivity (i.e., true positive rate, TPR) of 73.79%, and an accuracy of 94.46%. Furthermore, by varying the classification threshold on the log-mean parameter, we produced a ROC curve, and obtained an Area Under the ROC Curve (AUC) of 0.744.

6.4.3 Processing times

We evaluated the processing time of each module on a computer equipped with a Nvidia GeForce GTX TITAN X (Maxwell architecture) and an Intel i7-6700. As pre-processing for one frame, the Viola and Jones algorithm [138] extracts the face image in 8–10ms. The "eyelids distance" module processes one face image in 1.6ms. The "drowsiness" module processes a 1-min sequence of eyelids distances in 0.1ms, 2.1ms, and 9.5ms when using 1, 6, and 29 models, respectively. Therefore, our system satisfies real-time constraints.

6.4.4 Comparison of performance with the baseline system

Similarly to the previous chapter, we compare the performance of our parametric system with the performance of our baseline system. We modified, re-trained, and re-evaluated the baseline system to provide fair comparison with our parametric system. Details follow.

Modifications to the baseline system

We made modifications to each module of the baseline system.

1. We replaced the "face landmarks" and "eyelids distance" modules of our baseline system with the "eyelids distance" module of our parametric system.

2. We modified the "ocular features" module of our baseline system to extract six standard ocular features from a single time window with a length of 60s. We used the six

following standard ocular features: the mean blink duration, $\overline{D}_{blink}$; the mean closing duration, $\overline{D}_{closing}$; the mean closed duration, $\overline{D}_{closed}$; the mean opening duration, $\overline{D}_{opening}$; the number of microsleeps, $N_{\mu sleeps}$; and the percentage of eye closure below 70%, *PERCLOS*.

3. We modified the "drowsiness" module of our baseline system to regress the median RT, i.e., $\hat{\mu}^{-1}$, with one Support Vector Regression (SVR) model. We considered both a linear kernel and a radial basis function (RBF) kernel.

Training of the baseline system

We trained 29 SVRs following a leave-one-subject-out cross-validation strategy of 29 folds. However, considering the significantly faster training time of SVRs compared to CNNs, we validated the hyper-parameters (C and ϵ for the linear kernel; C, ϵ, and γ for the RBF kernel) via an inner leave-one-subject-out cross-validation strategy of 28 folds, i.e., all subjects (29) but the one (1) in the test set of the outer cross-validation. Upon determination of the optimal values of hyper-parameters, we trained the final model on all 28 subjects of the training set (of the outer cross-validation).

We obtained all samples of the training, validation, and test sets in the same manner, i.e., by extracting 37 samples from each 10-min PVT using a deterministic sliding window strategy with a step of 15s. We individually scaled each feature so as to be within the range $[0, 1]$ for the samples of the training set. We performed training and inference with the LIBSVM library [26]. We performed no data augmentation.

Evaluation of the performance of the baseline system and comparison

We evaluated the performance of the baseline system by aggregating the results of the 29 test sets, each associated to one trained model, before computing the performance metrics between $\hat{\mu}^{-1}$ and μ_{GT}^{-1}. The obtained results are shown in Table 6.3, with a comparison with the parametric system.

System	PCC	RMSE (ms)	Linear regression model (ms)
Parametric system	0.55	60.88	$\hat{\mu}^{-1} = 0.67\mu_{GT}^{-1} + 140$
Baseline system (linear kernel)	0.48	48.40	$\hat{\mu}^{-1} = 0.22\mu_{GT}^{-1} + 301$
Baseline system (RBF kernel)	0.49	48.41	$\hat{\mu}^{-1} = 0.21\mu_{GT}^{-1} + 306$

Table 6.3 – Comparison of performance between the parametric system and the baseline system (with a linear kernel or an RBF kernel).

With a linear kernel, the baseline system achieved an inter-subject PCC of 0.48 and an RMSE of 48.40ms. With an RBF kernel, the baseline system achieved an inter-subject PCC of 0.49 and an RMSE of 48.41ms. In comparison, the parametric system achieved, between $\hat{\mu}^{-1}$ and μ_{GT}^{-1}, an inter-subject PCC of 0.55 and an RMSE of 60.88ms. Therefore, the baseline system outperforms the parametric system in terms of RMSE, but falls behind it in terms of PCC. We can disambiguate these results by fitting a linear regression model with the form of $\hat{\mu}^{-1} = a\mu_{GT}^{-1} + b$, where a and b are the slope and bias, respectively. We obtained, as the fitted models, $\hat{\mu}^{-1} = 0.22\mu_{GT}^{-1} + 301$ for the baseline system with a linear kernel, $\hat{\mu}^{-1} = 0.21\mu_{GT}^{-1} + 306$ for the baseline system with an RBF kernel, and $\hat{\mu}^{-1} = 0.67\mu_{GT}^{-1} + 140$ for the parametric system. Figure 6.7 displays, for each system, the fitted model over the scatter plot of the median RT. We observe that the output ($\hat{\mu}^{-1}$) of the baseline system mostly ranges between 350–450ms, whereas the output of the parametric system mostly ranges between 300–700ms. As a consequence, considering that the ground-truth median RT mostly ranges between 300–450ms, the baseline system naturally achieves a smaller RMSE than the parametric system does. Evidenced by a larger PCC and a greater slope a, we conclude that the parametric system (based on a temporal CNN model) outperforms the baseline system (based on an SVR model).

6.5 Interpretation of the learned features

Interpretability of automatic systems is crucial, especially for safety-related applications where human lives are at stake. In this thesis, we designed our systems with interpretability in mind: we chose to decompose our system into successive modules with intermediate representations that are interpretable. For the parametric system, we chose to use the sequence of eyelids distances (1D) as the intermediate representation. This choice facilitates the interpretation of the data-driven features learned by the “drowsiness” module, which is the module we focus our analysis on. We perform the interpretation visually with the procedure detailed below. Note that, at the current stage of theoretical knowledge of CNNs, visual interpretation is a common and interesting approach for gaining insights into how the model operates.

6.5.1 Procedure

We interpret the learned features (of the “drowsiness” module) by visually comparing each of them side-by-side with the input that activates them the most. More specifically, we installed an individual in front of a video camera, processed his face images in real-time with our parametric system, and visually compared the sequence of eyelids distances (the input) with the learned features at two different positions in the temporal CNN.

1. The first position is the output of the ReLU layer after the first temporal convolution. At this position, the 32 features have the same temporal resolution as the sequence of eyelids distances; we call them the “local features”. We will show that the local features are equivalent to temporally segmenting the different phases of a blink (opened, closing, closed, and opening).

2. The second position is the output of the global average pooling layer. At this position, the 32 features consist of one scalar for the whole sequence of eyelids distances; we call them the “global features”. We will show that the global features include ocular features typically found in the literature such as the PERCLOS or the number of (long) blinks, as well as novel ones discovered by the training algorithm.

In such a manner, the individual in front of the video camera could control the input (by opening/closing his eyes) and find the patterns that activate the most the learned feature to interpret. We restricted our interpretation to the trained model of the 6th fold (out of the 29 folds). The reason is that this model performed well on its test set, which has the second largest range of median RTs (i.e., from 309ms to 734ms).

6.5.2 Interpretation of local features

The local features are the outputs of the ReLU layer after the first temporal convolution. By definition, they are sequences with the same temporal resolution as the input sequence

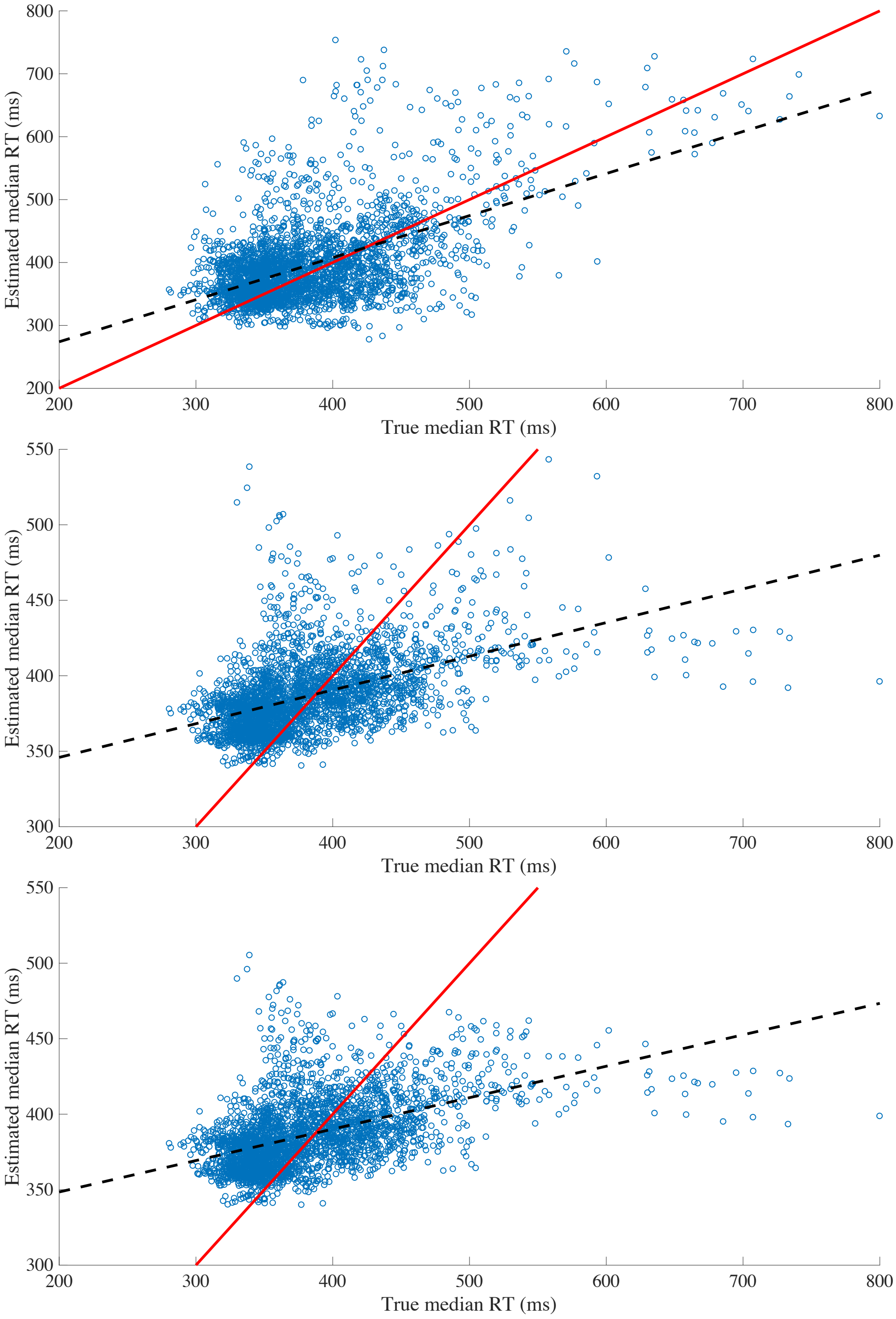


Figure 6.7 – Scatter plots of the estimated median RT, $\hat{\mu}^{-1}$, produced by the parametric system (top), the baseline system with a linear kernel (middle), and the baseline system with an RBF kernel (bottom) vs its ground-truth value, μ_{GT}^{-1}. The red lines are the perfect regressors. The black dashed lines are the fitted linear regression models.

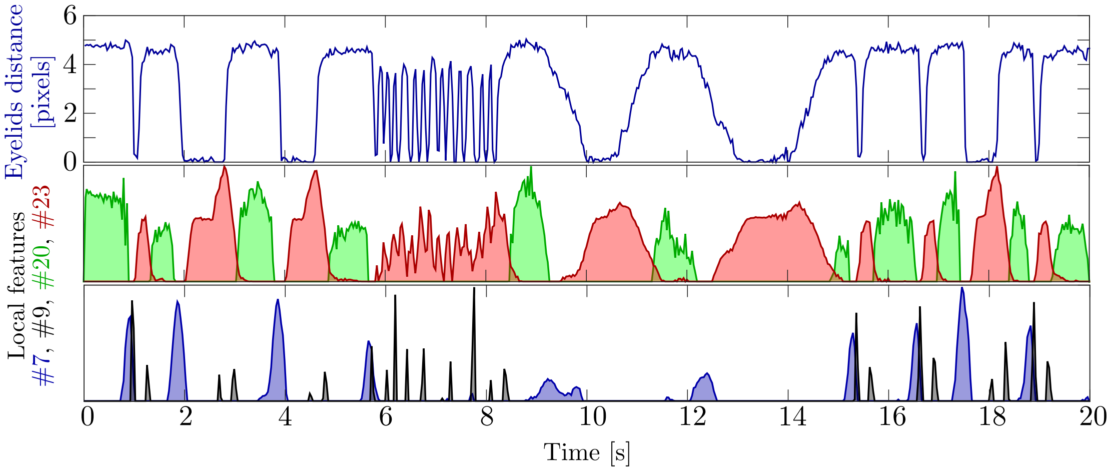


Figure 6.8 – Examples of four local features (bottoms) time-synchronized with a 20-s sequence of eyelids distances (top). Feature #20 (green) activates with opened eyes, #23 (red) with closed eyes and sharp eye opening, #7 (blue) with sharp eye closing, and #9 (black) with sharp eye opening. The amplitude of each local feature is normalized for visualization purposes. This figure is best viewed in color.

of eyelids distances. As a result, their visual interpretation is straightforward since their "activations", i.e., values greater than zero, are temporally localized, i.e., time-synchronized with the input. Figure 6.8 illustrates four local features time-synchronized with the same input sequence of eyelids distances.

As shown in Table 6.4, we distinguish five conditions under which a local feature may activate. We observed that 17 local features activate when the eye is opened, 9 activate when the eye is closed, 11 when the eye sharply closes, 17 when the eye sharply opens, and 6 when the eye shortly blinks. Notice that these numbers do not sum to 32, i.e., the total number of local features. The reason is that one local feature can activate because of more than one condition. For example, as is visible in Figure 6.8, the local feature #23 activates when the eye is closed, but also when the eye sharply opens.

Activating condition	**Local feature index**	**Total**
Eye opened	$1, 3, 5, 6, 8, 10, 12, 13, 15,$ $16, 18, 19, 20, 22, 24, 27, 31$	17
Eye closed	$2, 4, 11, 17, 23, 25, 26, 29, 32$	9
Eye (sharp) closing	$2, 7, 13, 14, 17, 25, 26, 28, 29, 30, 31$	11
Eye (sharp) opening	$2, 3, 4, 8, 9, 11, 12, 14, 15,$ $16, 18, 21, 22, 23, 24, 27, 30$	17
Short blink	$1, 5, 6, 10, 19, 32$	6

Table 6.4 – Summary of the visual interpretation of the local features of our "drowsiness" module.

6.5.3 Interpretation of global features

The global features are the outputs of the global average pooling layer. By definition, each of them consists of one real number computed from the whole 1-min sequence of eyelids

Global feature index		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Category	Integrated closed	+	+*	+*		+*					+*	+*
	Integrated opened											
	Integrated droopy				+			+		+		
	PERCLOS											
	Number of blinks		+*	+*			+*					
	Other								+			
PCC with analytic implement.		**0.97**	**0.82**	0.39	**0.95**	**0.97**	0.21	**0.9**	0.27	**0.9**	**0.93**	**0.88**

Global feature index		**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Category	Integrated closed						+	$\overset{*}{+}$				
	Integrated opened		+	+	+	+				+		
	Integrated droopy											
	PERCLOS	+							+			
	Number of blinks							+*			+	+*
	Other											
PCC with analytic implement.		**0.81**	**0.99**	**0.99**	**0.99**	**0.99**	**0.91**	0.23	**0.76**	**0.99**	0.56	0.52

Global feature index		**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**Total**
Category	Integrated closed	Constant value	+*									9
	Integrated opened					+				+	+	8
	Integrated droopy			+	+				−			6
	PERCLOS											2
	Number of blinks						+*					7
	Other							+				2
PCC with analytic implement.			0.32	**0.82**	**0.85**	**0.97**	**0.74**	/	**−0.73**	0.18	**0.93**	

Table 6.5 – Summary of the visual interpretation of the global features of our "drowsiness" module. The symbol "+" corresponds to a positive relationship, the symbol "−" to a negative one, and the symbol "*" indicates that the corresponding global feature is a "variant", i.e., that behaves differently than the other features within the same category.

distances. To facilitate the visual interpretation of a global feature, we construct its "global feature sequence", corresponding to the sequence of partial sums (i.e., the cumulative sum) of the input sequence of the global average pooling layer. In this way, we can visually associate an increase/decrease in these global feature sequences with an "activation event" occurring in the sequence of eyelids distances (e.g., a blink or a value). Note that the global feature sequences and the eyelids distance sequence are time-synchronized, but have different temporal resolutions (of 10 and 30 frames per second, respectively). Figure 6.9 illustrates eight global feature sequences each time-synchronized with an eyelids distance sequence that highly activates each of them.

As shown in Table 6.5, we observed that each global feature belongs to at least one of six categories. We observed 9 global features that are equivalent to an integration of a closed eye signal, 8 equivalent to an integration of an opened eye signal, 6 equivalent to an integration of a slightly closed, droopy eye signal, 2 equivalent to a PERCLOS of long blinks, 7 similar to the number of blinks, and 2 outsiders which are sensitive to opening/closing activation events. Furthermore, we noticed 11 "variants", which we define as global features that have activation events slightly different than the other features within the same category. For instance, a variant in the "number of blinks" category might be related to the number of blinks longer than 400ms rather than the number of blinks of any duration. Details about each category follow.

1. Global features in the "integrated closed" category activate (i.e., have their sequence increase) when, and by an amount proportional to how much, the eyelids distance is below a particular threshold. We estimated the thresholds to equal 2.6, 2.7, 3, 3.4, 3.6, 4.1, 4.3, 4.4, and 5.1 pixels for features #3, #2, #10, #1, #5, #17, #11,

#18, and #24, respectively. Variants activate when the eyelids distance is below the threshold during a long eye closure, but not during a short blink.

2. Global features in the "integrated opened" category activate when, and by an amount proportional to how much, the eyelids distance is above a particular threshold. We estimated the thresholds to equal 3.8, 4, 4.2, 4.2, 4.3, 4.3, 4.7, and 5.3 for features #20, #13, #14, #27, #15, #16, #32, and #31, respectively. We observed no variants.

3. Global features in the "integrated droopy" category activate when the eyelids distance is between two particular thresholds, and by an amount proportional to how close this eyelids distance is to the middle of these thresholds. We estimate the pair of thresholds to equal $\{2.8, 4.6\}$, $\{3.2, 6.1\}$, $\{2.8, 4.5\}$, $\{0, 4.7\}$, $\{3.5, 5\}$, and $\{3.8, 7.9\}$ for features #4, #7, #9, #25, #26, and #30, respectively. We observed no variants, except feature #30, which activates negatively (its sequence decreases rather than increasing) because of a negative multiplicative weight in the batch normalization layer.

4. Global features in the "PERCLOS" category activate (by a constant amount) when the eyelids distance is below a particular threshold (4.5 pixels for #12, and 4.3 pixels for #19) during a long blink. We observed no variants.

5. Global features in the "number of blinks" category activate (by a step) whenever a blink occurs. The variants of this category add some constraints on the blinks. Variants #2, #6, and #22 only activate for blinks longer than ∼250ms, ∼400ms, and ∼1300ms, respectively. Variants #3 and #18 only activate for short blinks that are preceded by at least ∼2.5s of eye openness. Variant #28 activates for (1) medium blinks of duration between 100ms and 250ms, and (2) short blinks preceded by another blink at most ∼2.5s before.

6. Global features #8 and #29 do not fit in any category and are considered as outsiders. Feature #8 activates (by a step) when the eyelids distance rapidly increase from $[3, 4.5]$ pixels to above 5 pixels. This activation event may correspond to a sudden recovery in alertness, which is a common behavior following a micro-sleep. Feature #29 activates when the eye slowly closes within a particular range of eyelids distances. Note that global feature #23 "died" during training, and always returns a constant value.

Note that we visually performed these interpretations of global features via trials and errors, which may lead to observation biases. Therefore, we checked our interpretation of most global features by (1) analytically implementing them as best as possible (based on our visual interpretation), and by (2) computing the correlation between their true, observed value and their analytic value. As shown in Table 6.5, a total of 22 global features (out of 31) are strongly correlated (PCC above 0.7) with their analytic value.

6.5.4 Interpretation of inner-workings

By putting together the interpretations of local and global features, we can point out some similarities between (1) the inner-workings of our parametric system and (2) the inner-workings of systems typically found in other studies, i.e., based on pre-defined, hard-coded ocular features. Indeed, other systems would first segment the blinks into four eye states (opened, closing, closed, and opening) based on the sequence of eyelids distances and/or

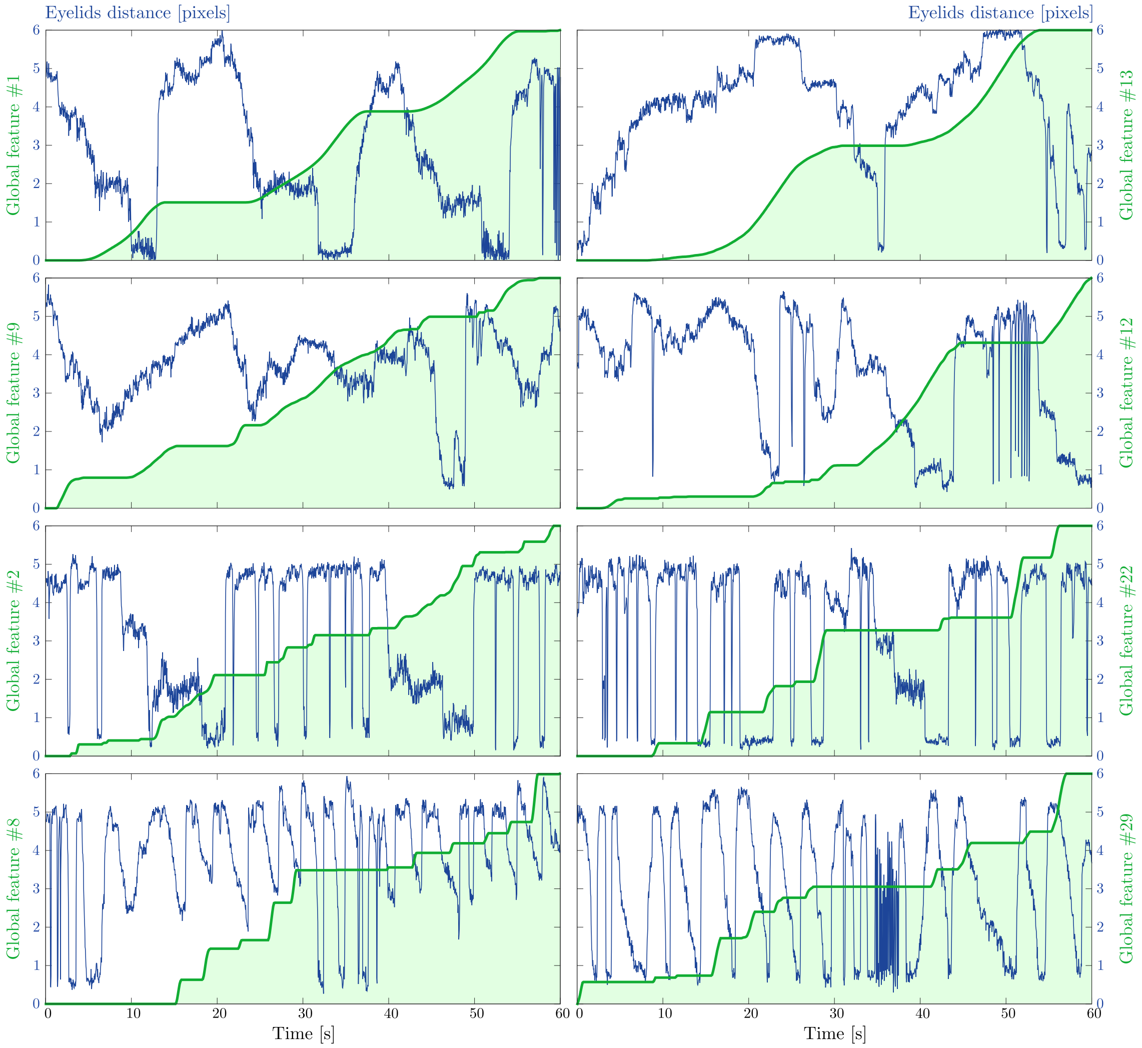


Figure 6.9 – Examples of the cumulative sum of eight global features (green) time-synchronized with a 1-min sequence of eyelids distances that highly activates each of them (blue). The amplitude of each global feature is normalized for visualization purposes. Feature #1 is similar to an integration of a closed eye signal (below ∼3.7 pixels), feature #13 to an integration of an opened eye signal (above 4 pixels), feature #9 to an integration of a slightly closed, droopy eye signal (between 3 and 4.3 pixels), feature #12 to the PERCLOS of long blinks, feature #2 to a combination of (1) an integration of a long closed eye signal and (2) the number of blinks longer than 100ms, feature #22 to the number of very long blinks (longer than ∼1300ms), feature #8 to the number of sharp eye opening from $[3, 4.5]$ pixels to above 5 pixels, and feature #29 to the number of slow eye closing.

its derivative, just like our local features do. Then, these other systems would compute some ocular features such as the PERCLOS, the mean closing/opening speed, or the mean blink duration, just like our global features do. To process these ocular features, other systems would finally use machine learning models (e.g., an artificial neural network, an SVM, and a logistic regression model) to characterize drowsiness, just like the last two fully connected layers of the "drowsiness" module of our parametric system.

While there are some clear similarities in inner-workings, it is important to note that our parametric system is data-driven, i.e., its inner-workings has been tweaked and shaped from data rather than from prior knowledge. One could argue that the choice of architecture we made for the temporal CNN constitutes prior knowledge. This is true to some extent, but such choice is significantly less restricting than the choice of using hard-coded ocular features. As proof, our parametric has learned to extract some global features that relate to some typical hard-coded ocular features, as well as some novel global features.

6.6 Conclusion

In this chapter, we presented a parametric drowsiness characterization system that is novel, data-driven, automatic, real-time, and generic. Our parametric system processes a 1-min sequence of face images with two successive modules, extracts data-driven features related to eye closure dynamics, and estimates the recinormal pdf of drowsiness-induced RTs, $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}\left(x\right)$. To train our system, we introduced (1) a ground-truth set of skill-normalized RTs that were locally observed, S_{GT}, and (2) a loss function that measures how well $\mathcal{R}_{[\hat{\mu},\hat{\sigma}^2]}\left(x\right)$ fits the observations of S_{GT}.

We evaluated our parametric system in controlled, laboratory conditions on 29 subjects via leave-one-subject-out cross-validation. The results show that our system produces a meaningful estimate of the log-mean parameter, $\ln\left(\hat{\mu}\right)$, but produces a poor estimate of the log-variance parameter, $\ln\left(\hat{\sigma}^2\right)$. Indeed, the estimated log-mean parameter is well correlated (inter-subject PCC of 0.52) to its ground-truth value, but the estimated log-variance parameter is not (inter-subject PCC of 0.15). Out of 29 subjects, we observed that 18 subjects (62%) have an intra-subject PCC above 0.4 between $\ln\left(\hat{\mu}\right)$ and $\ln\left(\mu_{GT}\right)$, and that only 5 subjects (17%) have a PCC above 0.4 between $\ln\left(\hat{\sigma}^2\right)$ and $\ln\left(\sigma^2_{GT}\right)$.

We converted our system into a binary classifier by thresholding the median RT at 500ms, and achieved a specificity of 95.19%, a sensitivity of 73.79%, an accuracy of 94.46%, and an AUC of 0.744. We showed that our parametric system outperforms our baseline system based on a vector of standard ocular features being fed to an SVR, which is representative of a wide range of systems found in other studies.

We conducted a visual, and partly analytic, interpretation of the features learned by our system and related to eye closure dynamics. We found that the model has learned to extract some (global) features that are closely related to those typically found in the literature such as the PERCLOS and the number of long blinks, as well as to some novel ones such as the integration of a "droopy eye" signal and the number of "sudden recovery in alertness" events.

In real-world applications, the driver (or a monitoring system and/or a semi-autonomous driving system) could take advantage of this distribution (or any derived metrics, such as the median RT), realize there is a high (or higher than usual) probability that he/she will not be able to react fast enough to sudden and potentially dangerous situations, and then decide to take safety actions with full knowledge.

Chapter 7

Conclusion

Drowsiness is a complex physiological state associated with a difficulty of staying awake, a strong inclination toward falling asleep. During the performance of a task, drowsiness impairs the ability of an individual to make sound decisions and to complete the task at normal performance. When the task is critical, drowsiness therefore becomes a danger that puts human lives at risk. In facts, drowsiness is a major cause of fatal accidents, in particular in the transportation sector where it is estimated to be responsible for 20–30% of them. There is thus a clear need for automatic, real-time drowsiness characterization systems that aim at preventing such accidents by issuing timely drowsiness warnings to vehicle operators.

The main goal of this thesis is the development of novel, automatic, and real-time drowsiness characterization systems that (1) operate on a video stream of face images, (2) focus on the analysis of eye closure dynamics, and (3) are trained with a ground truth of drowsiness based on impairments of psychomotor performance. To this end, we collected a dataset composed of 32 subjects who each performed three Psychomotor Vigilance Tasks (PVTs) under increasing acute sleep deprivation conditions (see Chapter 3). From this dataset, we developed three novel systems operating each on a 1-min sequence of face images: a baseline system in Chapter 4, a multi-timescale system in Chapter 5, and a parametric system in Chapter 6.

The baseline system characterizes drowsiness from a set of pre-defined ocular features, which is a typical approach used by most systems of other studies. As output, this system estimates (or predicts) a binary Level of Drowsiness (LoD) or the mean reaction time (RT). The baseline system allowed us to study the relationship between eye closure dynamics and performance impairments. We found that the number of microsleeps ($N_{\mu sleeps}$), the average blink duration ($\overline{D}_{blinks}$), the average opening duration ($\overline{D}_{opening}$), the average closed duration ($\overline{D}_{closed}$), and the proportion of normalized eyelids distance below 10% ($H_{[0,0.1]}$) are the most-important ocular features related to impairments of performance induced by drowsiness. We evaluated the performance of the baseline system at four different problems on 14 subjects using a leave-one-subject-out cross-validation. For the estimative regression problem, the baseline system achieved an inter-subject Pearson correlation coefficient (PCC) of 0.62 and an inter-subject Root Mean Square Error (RMSE) of 103ms. For the estimative classification problem, the baseline system achieved a true detection rate of alertness of 87.73%, a true detection rate of drowsiness of 75.46%, and a global accuracy of 86.19%. For the predictive classification (regression, respectively) problem, the baseline system achieved slightly lower performance than for the estimative classification (regression, respectively) problem. Upon further analysis, we found that the baseline system has learned predictive models that are almost identical to their equivalent estimative models, which stems from the fact that the future of performance impairments is somewhat unpredictable—the best strategy for predicting the future is therefore to use the estimate of the present.

The multi-timescale system characterizes drowsiness from four sets of data-driven ocular features, with each set extracted at a different timescale, i.e., window length. As output, this system estimates four binary LoDs with diverse trade-offs between accuracy and responsiveness. When combining these four LoDs, the system is able to (1) detect drowsiness onsets further in advance (at the cost of accuracy), and to (2) detect drowsiness onsets with high accuracy (at the cost of responsiveness). To train such a system, we introduced a multi-timescale ground truth of drowsiness based on skill-normalized RTs produced from four sliding time windows of diverse lengths. We evaluated the performance of the multi-timescale system on 29 subjects using a leave-one-subject-out cross-validation. The multi-timescale system achieved, for the 1st, 2nd, 3rd, and 4th timescales respectively, a true detection rate of alertness of 72.26%, 89.29%, 90.44%, and 94.80%; a true detection rate of drowsiness of 58.69%, 71.84%, 75.76%, and 74.19%; and a global accuracy of 70.69%, 85.45%, 89.82%, and 94.22%. We showed that context from the higher timescales is crucial for obtaining strong performance at the short timescale. This makes sense since a single long blink is more probably associated with a brief episode of drowsiness if the driver has been experiencing long blinks for the last minute than if he has not. Furthermore, we showed that the multi-timescale system outperforms the baseline system. This demonstrates the appropriateness of using a temporal convolutional neural network (CNN) model to characterize drowsiness from a sequence of eyelids distances.

The parametric system characterizes drowsiness from a set of data-driven ocular features. As output, this system estimates the two parameters (mean and variance) of the instantaneous reciprocal normal (“recinormal”) probability density function (pdf) of drowsiness-induced RTs. The ground truth of drowsiness consists of a set of skill-normalized RTs that occurred recently within a sliding time-window, and does therefore not consist of a single continuous or discrete quantity like the ground truths of the previous systems do. We evaluated the performance of the parametric system on 29 subjects using a leave-one-subject-out cross-validation. We showed that the parametric system estimates well the log-mean parameter (PCC of 0.52), but estimates poorly the log-variance parameter (PCC of 0.15) of the recinormal pdf. We converted our parametric system into a binary classifier by thresholding at 0.5 the probability of observing a RT above 500ms, which achieved a true detection rate of alertness of 95.19%, a true detection rate of drowsiness of 73.79%, a global accuracy of 94.46%, and an Area Under the ROC Curve (AUC) of 0.744. Furthermore, we showed that the parametric systems outperforms the baseline system. This again demonstrates the appropriateness of using a temporal CNN model to characterize drowsiness from a sequence of eyelids distances. With the goal of understanding the inner-workings of such temporal CNN, we conducted a visual interpretation—and a first analytic analysis—of the data-driven ocular features that the temporal CNN learned to extract from the sequence of eyelids distances. We pointed out strong similarities between the ocular features learned parametric system and the typical ocular features used by other systems in the literature. We found that the CNN learned to extract “local features” that are equivalent to a temporal segmentation of the different phases of a blink (i.e., opened, closing, closed, and opening), or of a combination of these phases (e.g., a combination of opened and opening phases). The CNN also learned to extract “global features” that are closely related to those typically found in the literature such as the PERCLOS and the number of long blinks, as well as to some novel ones such as the integration of a “droopy eye” signal and the number of “sudden recovery in alertness” events. Interestingly, whereas most global features take into account every blink, we observed some variants that take into account only the blinks of specific durations, e.g., above 400ms, or within the range 100–250ms.

Throughout this thesis, we discussed various aspects of the development of drowsiness characterization systems. In Chapter 2, we proposed a classification of indicators of drowsiness, provided a list of their associated standard measures, and indicated which indicators are most suited to be used as (1) an input to a system or as (2) a ground truth to train and evaluate a system. Moreover, we provided a comprehensive review of the systems of other studies in the scientific literature. In Chapter 3, we discussed various choices for designing the protocol of the sleep-deprivation dataset, which is essential for the development of drowsiness characterization systems. Furthermore, we discussed the ecological validity of a laboratory dataset, i.e., the extent to which the conclusions and findings drawn from a laboratory dataset can be generalized to real-life, operational settings. Although we strongly advocate for further research on this particular topic, laboratory sleep-deprivation datasets appear to have high relative—though not absolute—ecological validity. Indeed, laboratory protocols are designed to (1) favor drowsiness via monotonous tasks—for practical and conditions-controlling reasons, and to (2) not expose the subject to danger—for ethical and safety reasons. Therefore, drowsiness reaches generally higher levels in a laboratory study than in real-life settings, but increases with a similar pattern in both cases. One may thus develop a drowsiness characterization system with laboratory data, but should probably adjust the system's thresholds for producing drowsiness warnings in operational settings.

Furthermore, we proposed novel algorithms for the extraction of the eyelids distance from a face image. In this respect, we introduced in Chapter 4 (and detailed in Appendix A.4) several adjustments to the classic formulation of constrained local models (CLMs) and the regularized landmark mean-shift (RLMS) fitting strategy. Although we did not perform a quantitative evaluation, these adjustments led to a gain in alignment performance for landmarks with multi-modal appearance (e.g., eyes and mouth), and an increase in robustness to occlusions based on the discrepancy between the depth map and the 3D shape model. However, we still did not manage to obtain a generic CLM-based model that is able to align the eyelids landmarks with satisfactory accuracy. For this reason, we explored the use of CNNs to straightforwardly estimate (1) the eyelids distance from an eye image in Chapter 5, and (2) the maximum eyelids distance from a face image in Chapter 6. Both of these approaches achieved strong performance, i.e., an RMSE of about 0.5 pixel, have the advantages of being simple and fast (with the adequate hardware), and turned out to be reasonably robust to occlusions (including glasses and hands) and other facial expressions.

The systems developed in this thesis have some limitations. First, our sleep-deprivation dataset suffers from a small amount of data. Indeed, our dataset is composed of 32 subjects—29 out of which are usable for developments—who each performed three 10-min PVTs. We recommend using a greater number of subjects in order to develop and validate a drowsiness characterization system that generalizes well to a wide range of individuals in operational settings. In addition, we recommend using a greater number of PVTs, thereby increasing the representativity in the time-of-day (circadian rhythm) which is known to be a crucial determinant of drowsiness. Second, our sleep-deprivation dataset suffers from a limited representativity in subjects. Indeed, the subjects of our dataset were relatively young, i.e., with ages in the range of 19–34 years. We recommend incorporating in the dataset older subjects and patients suffering from sleep disorders.

In addition, there is still some room for further improvements that one could implement. For example, one could incorporate in the decision of the system the time-of-day and the time-on-task, i.e., two important and easily-obtainable determinants of drowsiness. One could compute/learn a baseline of eyelids distance or a baseline of (pre-defined or data-driven) ocular features, which would enable some kind of normalization, and potentially improve characterization performance. Also, one could develop a way to adapt automatically or semi-automatically the threshold of the drowsiness warnings, which would be a way to personalize the characterization of drowsiness to the vehicle operator.

Lastly, there are still some key questions that we have not addressed in this thesis. What is the legal and public acceptance of automatic, real-time drowsiness characterization systems based on face images? What would be the behavior of a vehicle operator in the presence of a system monitoring them? Would they accept the warnings and adopt adequate countermeasures? Or would they try to fool the monitoring system to keep driving anyway? Should the drowsiness warnings be a sound, a vibration, a message, an indication via a gauge, or via another mean? Nevertheless, what we do know is that efficiently mitigating drowsy driving will require a combination of multi-disciplinary efforts: (1) the development of reliable drowsiness characterization systems and other vehicular technologies such as collision avoidance systems, (2) new legislations about liability of drowsy driving and monitoring technologies, (3) public education and awareness about the problem and its current solutions, and (4) further fundamental sleep research to discover new and more reliable indicators of drowsiness.

List of publications

Paper in peer-reviewed academic journal

- Q. Massoz, J. Verly, and M. Van Droogenbroeck. Multi-timescale drowsiness characterization based on a video of a driver's face. Sensors, 18(9):1–17, August 2018.

Papers in scientific congresses and symposia as first author

- Q. Massoz and J. Verly. Vision-based system for monitoring vehicle operator responsiveness from face images. In International Conference on Managing Fatigue, pages 1–3, San Diego, CA, USA, March 2017.

- Q. Massoz, T. Langohr, C. François, and J. Verly. The ULg multimodality drowsiness database (called DROZY) and examples of use. In IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–7, Lake Placid, NY, USA, March 2016.

- Q. Massoz, T. Langohr, and J. Verly. Preliminary development and test of a new automatic drowsiness quantification system using range and intensity images obtained from a dashboard-mounted near-infrared 3D range sensor. In International Symposium on Somnolence and Safety (SomnoSafe), Brussels, Belgium, February 2016.

Papers in scientific congresses and symposia as co-author

- C. François, Q. Massoz, T. Hoyoux, J. Wertz, and J. Verly. First adaptation of a validated drowsiness monitoring system to process face images instead of eye images. In International Conference on Managing Fatigue, pages 1–3, San Diego, CA, USA, March 2017.

- C. François, V. Bosch, Q. Massoz, B. Fortemps de Loneux, R. Poirrier, and J. Verly. Development of an automated reference approach for quantifying drowsiness using polysomnographic signals. In International Symposium on Somnolence and Safety (SomnoSafe), Brussels, Belgium, February 2016.

- C. François, V. Bosch, Q. Massoz, B. Fortemps de Loneux, R. Poirrier, and J. Verly. Development and validation of an automatic reference polysomnographic system for quantifying drowsiness. In World Congress of the World Sleep Federation (Worldsleep), Istanbul, Turkey, November 2015.

Appendix A

Face alignment with constrained local models

Face alignment is the problem of localizing a set of face landmarks in a target image $\mathcal{I}$, i.e., the problem of fitting a face shape to a target image. Fundamentally, face alignment is challenging as it involves an optimization in a high-dimensional space, where the appearance of the face can vary greatly between instances due to rigid (i.e., rotation and translation in 3D) and non-rigid (i.e., facial expressions and identity) deformations of the face, lightning conditions, image noises, and occlusions. Face alignment requires three main components: (1) a deformable shape model that constrains the fitting procedure to lead to valid face shape configurations, (2) an appearance model that drives the fitting procedure across the image content, and (3) a strategy, i.e., algorithm, for the fitting procedure that puts components (1) and (2) together. Unless stated otherwise, the equations in this appendix are adapted from Saragih *et al.* [121] and from the PhD thesis of Tadas Baltrušaitis [9].

A.1 Deformable shape model

The deformable shape model, also known as the point density model (PDM), parameterizes how the face rigidly and non-rigidly deforms. More specifically, the PDM is a generative function producing a 2D face shape, $\mathbf{x}$, from a set of parameters, $\mathbf{p}$. Note that the PDM generates a valid face shape only when the parameter values are valid, i.e., statistically consistent with those computed from a set of annotated face shapes. The PDM mapping is a composition of three operations: (1) the application of non-rigid deformations to a 3D average, neutral face shape within a 3D, zero-centered, reference space; (2) the application of rigid deformations to place the 3D face shape into the 3D camera coordinates; and (3) the projection of the 3D face shape into the 2D image coordinates, usually using weak perspective projection. In practice, non-rigid deformations are linearly approximated by applying a principal component analysis (PCA) on a set of annotated face shapes that were rigidly aligned into the reference space. The number of non-rigid parameters, k, is generally chosen such that the k PCA's basis vectors account for a portion, e.g., 99%, of the variance of non-rigid deformations. Rigid deformations and the weak perspective projection are often combined so as to re-parameterize the 6 rigid parameters. Mathematically, the 2D location of the PDM's ith landmark, $\mathbf{x}_i \in \mathbb{R}^2$, is given by

$$\mathbf{x}_i = P_{wp,i}(\mathbf{p}) = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \end{bmatrix} \mathbf{R}(\theta_p, \theta_y, \theta_r) \left(\bar{\mathbf{X}}_i + \mathbf{\Phi}_i \mathbf{q} \right) + \begin{bmatrix} t_x \\ t_y \end{bmatrix}, \quad \forall i \in [1, n], \tag{A.1}$$

where n is the PDM's number of face landmarks; $P_{wp,i} : \mathbb{R}^{(6+k)} \rightarrow \mathbb{R}^2$ denotes the weak perspective projection of the ith 3D face landmark of the PDM; $\bar{\mathbf{X}}_i \in \mathbb{R}^3$ is the 3D location of the neutral face's ith landmark; $\mathbf{\Phi}_i \in \mathbb{R}^{3 \times k}$ is the PCA's sub-matrix of basis of non-rigid deformations, $\mathbf{\Phi} \in \mathbb{R}^{3n \times k}$, associated to the ith landmark; $\mathbf{q} \in \mathbb{R}^k$ is the vector of non-rigid parameters; $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ is the 3D rotation matrix which is parameterized by the pitch θ_p, yaw θ_y, and roll θ_r rotation angles; t_x and t_y are translations in the x- and y-coordinates, respectively; s is a scaling factor; and $\mathbf{p} = [\theta_p; \theta_y; \theta_r; t_x; t_y; s; \mathbf{q}] \in \mathbb{R}^{(6+k)}$ is the vector shape parameters.

Using such a deformable shape model, face alignment translates into finding an optimal and valid set of $6 + k$ shape parameters $\mathbf{p}^* = [\theta_p^*; \theta_y^*; \theta_r^*; t_x^*; t_y^*; s^*; \mathbf{q}^*]$ that minimizes the misalignment error

$$\begin{aligned} \mathbf{p}^* &= \arg\min_{\mathbf{p}} \mathcal{Q}(\mathbf{p}) \\ &= \arg\min_{\mathbf{p}} \mathcal{R}(\mathbf{p}) + \sum_{i=1}^{n} \mathcal{M}_i(\mathbf{x}_i; \mathcal{I}), \end{aligned} \tag{A.2}$$

where $\mathcal{Q}(\mathbf{p})$ is a regularized misalignment error function, $\mathcal{R}(\mathbf{p})$ penalizes deformations that are not valid (*a.k.a.* the regularization term or the prior term), and $\mathcal{M}_i(\mathbf{x}_i; \mathcal{I})$ measures the misalignment error that the ith landmark is experiencing at position $\mathbf{x}_i$ (which is a function of $\mathbf{p}$) in the image $\mathcal{I}$ (*a.k.a.* the data term or the likelihood term). Assuming that the PCA-embedded non-rigid parameters are distributed according to a Gaussian probability density function, the prior on $\mathbf{q}$ can be expressed as

$$\mathcal{R}(\mathbf{q}) = \mathbf{q}^T \mathbf{\Lambda}^{-1} \mathbf{q} = \|\mathbf{q}\|^2_{\mathbf{\Lambda}^{-1}}, \tag{A.3}$$

where $\mathbf{\Lambda}^{-1} = \text{diag}\{[\lambda_1^{-1}; \ldots; \lambda_k^{-1}]\} \in \mathbb{R}^{k \times k}$ is the inverse covariance matrix, i.e., a diagonal matrix with elements λ_i denoting the eigenvalue of the ith PCA mode of non-rigid deformations. Given that rigid deformations are all considered equally likely, and are thus generally not penalized, the prior on $\mathbf{p}$ is given by

$$\mathcal{R}(\mathbf{p}) = \mathbf{p}^T \tilde{\mathbf{\Lambda}}^{-1} \mathbf{p}, \tag{A.4}$$

where $\tilde{\mathbf{\Lambda}}^{-1} = \text{diag}\{[0; 0; 0; 0; 0; 0; \lambda_1^{-1}; \ldots; \lambda_k^{-1}]\} \in \mathbb{R}^{(6+k) \times (6+k)}$.

A.2 Appearance model

The appearance model parameterizes what the face looks like in an image. Such a model that relates the data x to the world state w can be either (1) generative, i.e., it models the contingency of the data on the world state $p(x \mid w)$, or (2) discriminative, i.e., it models the contingency of the world state on the data $p(w \mid x)$. The first type of model is able to generate a valid face appearance from a given valid set of parameters, whereas the second type is able to estimate how valid a given face appearance is. In both cases, the model enables the measurement of a misalignment error that a landmark is experiencing at a position in the image. In particular, constrained local models (CLM) [32, 121] is a family of face-alignment approaches that locally and discriminatively model the face appearance using local image patches centered on each face landmark. Compared to a holistic appearance model, e.g., such as the one used by active appearance models (AAM) [31], a local appearance model has the downside of being harder to build and optimize but has the advantages of generalizing better, of being faster, and thus of having stronger performance.

More specifically, a local appearance model consist of n independent local detectors, called local experts, each associated to one face landmark. The ith local expert discriminatively produces, from the local image patch centered on $\mathbf{x}_i$, a probability map (called a response map) that holds the probability of correct alignment, $\pi_{\mathbf{y}_i}$, that the ith landmark is experiencing at each of the positions, $\mathbf{y}_i \in \mathbf{\Psi}_i$, within the grid-like neighborhood of $\mathbf{x}_i$, $\mathbf{\Psi}_i$. The production of such response map can be efficiently performed using a convolution, the parameters (i.e., weights) of which can be learned from a set of training local image patches that were either aligned or misaligned with the annotated landmark locations. However, due to their small local support and the large variation in the local appearance around their landmark, these simple local experts suffer from the problem of ambiguity. Indeed, their response map may be multimodal, where the maximum of the response may not always correspond to the correct landmark location. However, this ambiguity problem is naturally addressed during the fitting procedure, where the n local response maps are holistically used to align the PDM's face shape on the target image.

A.3 Regularized landmarks mean-shift fitting procedure

Regularized landmarks mean-shift (RLMS) [121] is an iterative algorithm to fit the deformable shape model to the location likelihood grids (i.e., response maps) produced by the local appearance models. The RLMS algorithm works in a way similar to that of the Expectation-Maximization (EM) [38] algorithm. By considering the true landmark locations as hidden, latent variables, the RLMS algorithms alternates between performing an expectation (E) step, and a maximization (M) step.

Expectation step

During the E-step, using the current estimate for the face shape parameters $\mathbf{p}$ and for each $x_i = P_{wp,i}\left(\mathbf{p}\right)$ independently, we evaluate the posterior probabilities over candidate positions $\mathbf{y}_i$ as

$$w_{\mathbf{y}_i} = p\left(\mathbf{y}_i | l_i = 1, \mathbf{x}_i, \mathcal{I}\right) = \frac{\pi_{\mathbf{y}_i} \mathcal{N}\left(\mathbf{x}_i; \mathbf{y}_i, \rho \mathbf{I}\right)}{\sum_{\mathbf{z}_i \in \mathbf{\Psi}_i} \pi_{\mathbf{z}_i} \mathcal{N}\left(\mathbf{x}_i; \mathbf{z}_i, \rho \mathbf{I}\right)}, \tag{A.5}$$

where $w_{\mathbf{y}_i}$ is the posterior probability of the candidate $\mathbf{y}_i$, i.e., the posterior probability that the candidate position $\mathbf{y}_i$ is where the ith landmark is truly located ($l_i = 1$) in the image $\mathcal{I}$; $\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}\right)$ is the evaluation at $\mathbf{x}$ of the multivariate Gaussian function with a location $\boldsymbol{\mu}$ and a covariance matrix $\mathbf{\Sigma}$; and ρ is the noise variance on landmark location, usually computed as the arithmetic average of the eigenvalues in the subspace orthogonal to $\mathbf{\Phi}$ [121]. Note that the sum over all candidates equals one, i.e., $\sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} w_{\mathbf{y}_i} = 1$.

Maximization step

During the M-step, we find a new set of parameters, $\mathbf{p}^{(t+1)}$, that maximizes the posterior found in the E-step, i.e., we minimize $\mathcal{Q}\left(\mathbf{p}\right)$:

$$\begin{aligned} \mathbf{p}^{(t+1)} &= \arg\min_{\mathbf{p}} \mathcal{Q}\left(\mathbf{p}\right) \\ &= \arg\min_{\mathbf{p}} \|\mathbf{p}\|^2_{\tilde{\mathbf{\Lambda}}^{-1}} + C \sum_{i=1}^{n} \sum_{\mathbf{y}_i \in \mathbf{\Psi}_i} w_{\mathbf{y}_i} \left\|P_{wp,i}\left(\mathbf{p}\right) - \mathbf{y}_i\right\|^2, \end{aligned} \tag{A.6}$$

where C is a weighting coefficient, $P_{wp,i}\left(\mathbf{p}\right) = \mathbf{x}_i$ is the 2D location of the PDM's ith landmark, and $w_{\mathbf{y}_i}$ is constant (i.e., as computed in the E-step). This least-squares problem can be solved with the Gauss-Newton algorithm by finding the additive update, $\Delta\mathbf{p}$, to the current estimate of shape parameters, $\mathbf{p}^{(t)}$, that satisfies

$$\left.\frac{\partial \mathcal{Q}\left(\mathbf{p}+\Delta\mathbf{p}\right)}{\partial \Delta\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}^{(t)}} = 0, \tag{A.7}$$

where $\mathcal{Q}\left(\mathbf{p}+\Delta\mathbf{p}\right)$ is linearized about the point $\mathbf{p}=\mathbf{p}^{(t)}$ using Taylor series expansion:

$$\mathcal{Q}\left(\mathbf{p}^{(t)}+\Delta\mathbf{p}\right) = \|\mathbf{p}^{(t)}+\Delta\mathbf{p}\|^2_{\tilde{\mathbf{\Lambda}}^{-1}} + C\sum_{i=1}^{n}\sum_{\mathbf{y}_i\in\mathbf{\Psi}_i} w_{\mathbf{y}_i}\left\|\left(P_{wp,i}\left(\mathbf{p}^{(t)}\right)+\mathbf{J}_i\Delta\mathbf{p}\right)-\mathbf{y}_i\right\|^2, \tag{A.8}$$

where $\mathbf{J}_i \in \mathbb{R}^{2\times(6+k)}$ is the Jacobian matrix of the projected position of the ith PDM landmark *w.r.t.* the PDM parameters. The solution to A.7 is therefore given by

$$\Delta\mathbf{p} = \left(\tilde{\mathbf{\Lambda}}^{-1}+C\mathbf{J}^T\mathbf{J}\right)^{-1}\left(-\tilde{\mathbf{\Lambda}}^{-1}\mathbf{p}^{(t)}+C\mathbf{J}^T\mathbf{v}\right), \tag{A.9}$$

where $\mathbf{J}\in\mathbb{R}^{2n\times(6+k)}$ is the Jacobian matrix, $\mathbf{v}=[\mathbf{v}_1;\ldots;\mathbf{v}_n]\in\mathbb{R}^{2n}$ is the concatenation of the mean-shift vectors from each landmark, $\mathbf{v}_i$, given by

$$\mathbf{v}_i = \left(\sum_{\mathbf{y}_i\in\mathbf{\Psi}_i} w_{\mathbf{y}_i}\mathbf{y}_i\right) - P_{wp,i}\left(\mathbf{p}^{(t)}\right), \quad \forall i\in\{1,\ldots,n\}. \tag{A.10}$$

Notice that the term in parenthesis of Equation A.10 is reminiscent of the well-known mean-shift algorithm [52] which is used to iteratively locate the mode(s) of a density function (i.e., the response map in the RLMS algorithm). Therefore, the M-step produces the new estimate of the shape parameters

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)}+\Delta\mathbf{p}. \tag{A.11}$$

We redirect the interested reader to the PhD thesis of Tadas Baltrušaitis [9] for further details on CLM, in particular on implementation details and on how to compute the Jacobian matrix of the PDM, $\mathbf{J}$.

A.4 Adjustments for the baseline system

For the "face landmarks" module of our baseline system, we made several adjustments to the classic formulation of CLM and RLMS presented above. One reason is that, in addition to the image $\mathcal{I}$, our baseline system should be able to process the depth map $\mathcal{D}$ which is perfectly aligned with $\mathcal{I}$. Details of these adjustments follow. The equations in this section are original.

- We use a multimodal appearance model because the local appearance around a landmark can vary drastically, e.g., when the eye is closed vs open, or when the mouth is closed vs open. More specifically, we compute, during the E-step, the multimodal response maps with probability values given by

  $$w_{\mathbf{y}_i} = \max\left(w^1_{\mathbf{y}_i},\ldots,w^{M_i}_{\mathbf{y}_i}\right), \tag{A.12}$$

  where M_i is the number of modes for the ith landmark, and $w^m_{\mathbf{y}_i}$ is the mth-mode's posterior probability of the candidate position $\mathbf{y}_i$ in $\mathcal{I}$. We use $M_i=2$ for the 12 eye landmarks and the 20 mouth landmarks, and $M_i=1$ for the other 36 face landmarks. We detail the training procedure of the multimodal appearance model in the next section.

- For the purpose of fully exploiting the 3D information contained in $\mathcal{D}$, we reduce the projection errors by using a full perspective projection rather than a weak perspective projection. As a consequence, the re-parametrization of the rigid shape parameters (i.e., 3D rotation and 3D translation) to incorporate the projection is no longer feasible given that the “scaling factor” becomes function of the landmark and its distance from the camera (in the z axis). Therefore, the 3D locations of the landmarks are given by
$$\mathbf{X}_i = W_i\left(\mathbf{p}\right) = \mathbf{R}\left(\bar{\mathbf{X}}_i + \mathbf{\Phi}_i\mathbf{q}\right) + \mathbf{T}, \tag{A.13}$$
where $\mathbf{X}_i$ is the 3D location of the PDM's ith landmark in the camera coordinates, $W_i : \mathbb{R}^{6+k} \to \mathbb{R}^3$ maps the shape parameters $\mathbf{p}$ to $\mathbf{X}_i$, $\mathbf{R} \in \mathbb{R}^{3\times3}$ is the 3D rotation matrix, and $\mathbf{T} \in \mathbb{R}^3$ is the 3D translation vector. The 2D locations of the landmarks are given by
$$\mathbf{x}_i = P_f\left(\mathbf{X}_i = \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix}\right) = \frac{f}{Z_i}\begin{bmatrix} X_i \\ Y_i \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix}, \tag{A.14}$$
where $P_f : \mathbb{R}^3 \to \mathbb{R}^2$ denotes the full perspective projection, f is the focal length of the camera, and c_x and c_y are the x- and y-coordinates, respectively, of the principal point of the camera. Note that weak perspective projection assumes that each individual Z_i can be replaced by a constant depth, Z_*, because the depth of the object/face is assumed to be small/negligible compared to its distance from the camera, thereby leading to the definition of a scaling factor $s = f/Z_*$ and a 2D translation vector $\mathbf{t} = [t_x; t_y] = [sT_x + c_x; sT_y + c_y]$.

- We incorporate $\mathcal{D}$ into the fitting procedure by adding a likelihood term that minimizes the depth residual between (1) the depth values of $\mathcal{D}$ and (2) the depth values of the PDM's landmarks. The misalignment error function becomes
$$\mathcal{Q}(\mathbf{p}) = \underbrace{C_1\left\|\mathbf{p}\right\|^2_{\tilde{\mathbf{\Lambda}}^{-1}}}_{\text{prior}} + \underbrace{\sum_{i=1}^{n}\sum_{\mathbf{y}_i\in\mathbf{\Psi}_i} w_{\mathbf{y}_i}\left\|\mathbf{x}_i - \mathbf{y}_i\right\|^2}_{\text{likelihood from image } \mathcal{I}} + C_2 \underbrace{\sum_{i=1}^{n}\left\|\mathcal{D}\left(\mathbf{x}_i\right) - Z_i\right\|^2}_{\text{likelihood from depth map } \mathcal{D}}, \tag{A.15}$$
where $\mathcal{D}\left(\mathbf{x}\right)$ is the depth value of $\mathcal{D}$ at the 2D location $\mathbf{x}$, and C_1 and C_2 are weighting coefficients empirically set to 20 and 1e−4, respectively..

- For the purpose of increasing the robustness to occlusions, e.g., occluding hands or self-occlusions, we weigh each landmark in $\mathcal{Q}\left(\mathbf{p}\right)$ using the depth residual. During the E-step, we compute the weights according to the Welsch weighting function [64]:
$$\gamma_i = \exp\left(-\left(\frac{r_{d,i}}{\tau}\right)^2\right), \tag{A.16}$$
where γ_i is the depth-disparity weight of the ith landmark, $r_{d,i} = \mathcal{D}\left(\mathbf{x}_i\right) - Z_i$ is the depth residual of the ith landmark, and τ is a scaling factor that is empirically set to 20. The weight γ_i decreases towards zero when $r_{d,i}$ grows large, i.e., when the ith landmark is most probably occluded. Note that we set $\gamma_i = 1$ for the two iterations after initializing $\mathbf{p}$ to let the PDM coarsely fit the video frame content. The misalignment error function becomes
$$\mathcal{Q}(\mathbf{p}) = C_1\|\mathbf{p}\|^2_{\tilde{\mathbf{\Lambda}}^{-1}} + \sum_{i=1}^{n}\gamma_i\left[\sum_{\mathbf{y}_i\in\mathbf{\Psi}_i} w_{\mathbf{y}_i}\left\|\mathbf{x}_i - \mathbf{y}_i\right\|^2 + C_2\left\|r_{d,i}\right\|^2\right]. \tag{A.17}$$

Using the Gauss-Newton algorithm, the solution to this re-weighted least-squares problem is

$$\Delta \mathbf{p} = \left(C_1 \tilde{\boldsymbol{\Lambda}}^{-1} + \mathbf{J}^T \tilde{\boldsymbol{\Gamma}} \mathbf{J} + C_2 \mathbf{J}_{r_d}^T \boldsymbol{\Gamma} \mathbf{J}_{r_d} \right)^{-1} \left(-C_1 \tilde{\boldsymbol{\Lambda}}^{-1} \mathbf{p}^{(t)} + \mathbf{J}^T \tilde{\boldsymbol{\Gamma}} \mathbf{v} - C_2 \mathbf{J}_{r_d}^T \boldsymbol{\Gamma} \mathbf{r}_d \right), \tag{A.18}$$

where $\mathbf{r}_d = [r_{d,1}; \ldots; r_{d,n}]$ is the concatenation of the depth residuals from each landmark; $\boldsymbol{\Gamma} = \text{diag}\{[\gamma_1; \ldots; \gamma_n]\} \in \mathbb{R}^{n \times n}$ and $\tilde{\boldsymbol{\Gamma}} = \text{diag}\{[\gamma_1; \gamma_1; \gamma_2; \gamma_2; \ldots; \gamma_n; \gamma_n]\} \in \mathbb{R}^{2n \times 2n}$ are diagonal matrices of the depth-disparity weights; and $\mathbf{J}_{r_d} = \mathbf{J}_{\mathcal{D}} - \mathbf{J}_Z \in \mathbb{R}^{n \times (6+k)}$ is the Jacobian matrix of the depth residuals.

A.5 Training for the baseline system

For the “face landmarks” module of our baseline system, we trained 14 subject-specific modules, i.e., subject-specific deformable shape models and subject-specific multimodal appearance models. The equations and notations in this section are original.

Training of the deformable shape model

We trained the subject-specific shape models in four steps using the annotated 3D face shapes.

1. We translationally aligned the 720 annotated face shapes by centering each of them on zero, i.e., the mean of the X-, Y-, and Z-coordinates of their 68 landmarks are set to zero.

2. We doubled the number of samples by flipping every zero-centered face shapes about the X-axis of the camera.

3. We rotationally aligned the 1440 zero-centered face shapes using Procrustes Analysis [33], which iteratively minimizes the sum of distances of each shape to the mean shape.

4. We trained 14 subject-specific shape model using PCA with $k = 25$ non-rigid deformation modes. More specifically, we find a linear subspace shape model for subject j by applying PCA on (1) the aligned face shapes of subject j, augmented with (2) the aligned face shapes of the other 13 subjects that had their subject-specific neutral face shape (i.e., the mean of their aligned face shapes) replaced by the neutral face shape of subject j.

Training of the multimodal appearance model

We trained the subject-specific local experts using only their respective images annotated with the 2D face landmarks. More specifically, for each landmark i of subject j, we trained the M_i modes of the local expert in four steps using $\{(\mathcal{I}^k, \mathbf{x}_i^k) : k \in [1, l_j]\}$ with k the index of the annotated frame and l_j the number of annotated frames for subject j. Let $P(\mathcal{I}; \mathbf{x}) : \mathbb{R}^{H \times W}; \mathbb{R}^2 \to \mathbb{R}^{121}$ be the function that extracts a local patch (with size of 11×11 pixels, squeezed into a vector of size 121) of the image $\mathcal{I}$ (with size of $H \times W$ pixels) centered on the 2D location $\mathbf{x}$.

1. If $M_i > 1$, we clustered the l_j patches around the ith landmark, $\{P(\mathcal{I}^k; \mathbf{x}_i^k) : k \in [1, l_j]\}$, by fitting a mixture of M_i Gaussians using the EM algorithm. In such a

manner, we effectively arranged the l_j samples into M_i clusters, $\{(\mathcal{I}^k, \mathbf{x}_i^k) : k \in [1, l_j^m], m \in [1, M_i]\}$ with the sum of l_j^m being l_j. We applied the next steps on each cluster independently.

2. We trained a discriminative Support Vector Machine (SVM) model on the aggregation of three types of patches, for a total number of $6l_j^m$ patch samples: the perfectly-located patches, $\{P(\mathcal{I}^k; \mathbf{x}_i^k) : k \in [1, l_j^m]\}$; the nearby-located patches, $\{P(\mathcal{I}^k; \mathbf{x}_i^k + \mathcal{U}^k(1, 4)) : k \in [1, 2l_j^m]\}$ where $\mathcal{U}(a, b)$ is a random 2D perturbation uniformly distributed between a and b pixels; and badly-located patches, $\{P(\mathcal{I}^k; \mathbf{x}_i^k + \mathcal{U}^k(5, 11)) : k \in [1, 3l_j^m]\}$. We labeled the perfectly-located patches as “positive”, and the nearby- and badly-located ones as “negative”. We only retained the trained weights, $\mathbf{w}_{i,j,m} \in \mathbb{R}^{121}$, for the next steps.

3. We trained a logistic regression model $\sigma_{i,j,m} : \mathbb{R} \to [0, 1]$ defined as
$$\sigma_{i,j,m}(t) = \left(1 + \exp\left(-\left(a_{i,j,m}t + b_{i,j,m}\right)\right)\right)^{-1}, \tag{A.19}$$
where $t = \mathbf{w}_{i,j,m}^T P(\mathcal{I}; \mathbf{x})$ relates to the projection of the local patch (from image $\mathcal{I}$ centered on $\mathbf{x}$) on the direction orthogonal to the SVM's hyperplane. To do so, we associated to each patch sample of the second step, $P(\mathcal{I}^k; \mathbf{x}^k)$ with $k \in [1, 6l_j^m]$, a target value, z^k, given by
$$z^k = \exp\left(-\left(d^k\right)^2/2\right), \tag{A.20}$$
where $d^k = \|\mathbf{x}_i^k - \mathbf{x}^k\ \|^2$ is the distance between the annotated landmark location, $\mathbf{x}_i^k$, and the center of the sample patch, $\mathbf{x}^k$.

4. For the purpose of streamlining the inference, we (1) re-arranged the SVM weights onto a 2D grid, $\mathbf{w}_{i,j,m} \in \mathbb{R}^{121} \mapsto \mathbf{w}'_{i,j,m} \in \mathbb{R}^{11\times 11}$, (2) combined $a_{i,j,m}$ with $\mathbf{w'}_{i,j,m}$, and (3) redefined the patch extraction function, $P(\mathcal{I}; \mathbf{x}) \mapsto P'(\mathcal{I}; \mathbf{x}) : \mathbb{R}^{H\times W}; \mathbb{R}^2 \to \mathbb{R}^{11\times 11}$. The correct alignment probability $\pi_{\mathbf{x}}^{i,j,m} \in [0, 1]$ at location $\mathbf{x}$ in image $\mathcal{I}$ for the ith landmark, subject j, and cluster m is thus given by
$$\pi_{\mathbf{x}}^{i,j,m} = \left(1 + \exp\left(-\left(a_{i,j,m}\mathbf{w}'_{i,j,m}\right) \circledast P'(\mathcal{I}; \mathbf{x}) - b_{i,j,m}\right)\right)^{-1}, \tag{A.21}$$
where $\circledast$ is the 2D convolution operator. Note that, if $P'(\mathcal{I}; \mathbf{x})$ produces a local patch of size $\mathbb{R}^{w\times w}$, we can straightforwardly obtain from equation (A.21) the response map of size $(w - 11 + 1) \times (w - 11 + 1)$ candidate locations. We set w to 21 to produce response maps of size 11×11 candidate locations.

A.6 Comparison with “modern” face alignment techniques

Modern face alignment techniques rely more and more on machine learning models, and this for modeling the shape model and the appearance model as well as for learning the fitting procedure, thereby achieving greater performance than CLM. Indeed, instead of iteratively computing additive updates of the shape parameters with the Gauss-Newton algorithm, modern face alignment techniques iteratively estimate the additive update either (1) of the landmark positions [23, 81, 114] or (2) of the shape parameters [149] that are learned via cascaded regression models. The replacement of the Gauss-Newton algorithm by machine learning algorithms has multiple advantages: (1) the processing is generally faster, because inference does not require the computation of an inverse matrix as in Equation A.9; and (2) the appearance model and the optimization procedure are jointly contained in the regression model, the optimization of which leads to faster convergence and better performance, but requires a greater amount of data. Note that, when directly estimating the additive update of the landmark positions, the shape model, i.e., the inter-landmarks relationship, is contained in the regression model.

Appendix B

Support Vector Machine and Regression

B.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) [19, 35] is a discriminative binary classifier, i.e., a classification machine learning model, formally defined by an hyperplane separating training instances of two classes by an optimal margin. An SVM is trained in a supervised manner from a set of l training points, $\{(\mathbf{x}_i, y_i) : i \in [1, l]\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ is a vector of n features, and $y_i \in \{1, -1\}$ is the ground-truth class. Under given the regularization hyper-parameter $C > 0$, training an SVM classifier means solving the following primal optimization problem:

$$\begin{aligned} \underset{\mathbf{w},b,\boldsymbol{\xi}}{\text{minimize}} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i \left(\mathbf{w}^T \phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \quad i = 1, \ldots, l, \end{aligned} \tag{B.1}$$

where $\mathbf{w}$ is the weight vector of the hyperplane, b is the bias of the hyperplane, $\phi(\mathbf{x}_i)$ is a function that maps $\mathbf{x}_i$ into a high-dimensional space, C is a weighting coefficient, and ξ_i is the slack variable of the ith training point. However, because of the possible high dimensionality of the weight vector $\mathbf{w}$, we usually solve the following dual problem:

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{minimize}} \quad & \tfrac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{y}^T \boldsymbol{\alpha} = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, l, \end{aligned} \tag{B.2}$$

where $\mathbf{e} = [1, \ldots, 1]^T]$ is the vector of all ones, Q is an l by l positive semidefinite matrix with elements $Q_{ij} \equiv y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is the kernel function which quantifies the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Note that the mapping function, $\phi(\mathbf{x})$, does not need to be explicitly defined, given that only the kernel function, $K(\mathbf{x}_i, \mathbf{x}_j)$, appears in the dual problem. Popular kernel functions include

- the linear kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$;
- the polynomial kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \left(\gamma \mathbf{x}_i^T \mathbf{x}_j + r\right)^d$ with hyper-parameters r, d, and $\gamma > 0$;
- the radial basis function (RBF) kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ with hyper-parameter $\gamma > 0$.

After solving problem B.2 and by using the primal-dual relationship, the optimal $\mathbf{w}$ satisfies

$$\mathbf{w} = \sum_{i=1}^{l} y_i \alpha_i \phi(\mathbf{x}_i), \tag{B.3}$$

and the decision function is

$$y(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{w}^T \phi(\mathbf{x}) + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right). \tag{B.4}$$

B.2 Support Vector Regression (SVR)

Support Vector Regression (SVR) [135] is an extension of SVMs for regression problems. An SVR is trained in supervised manner from a set of l training points, $\{(\mathbf{x}_i, z_i) : i \in [1, l]\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ is a vector of n features, and $z_i \in \mathbb{R}$ is the ground-truth target. Under given hyper-parameters $C > 0$ and $\epsilon > 0$, training an SVR means solving the following primal problem:

$$\begin{array}{ll} \underset{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*}{\text{minimize}} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \sum\limits_{i=1}^{l} \xi_i + C \sum\limits_{i=1}^{l} \xi_i^* \\ \text{subject to} & \mathbf{w}^T \phi(\mathbf{x}_i) + b - z_i \leq \epsilon + \xi_i, \\ & z_i - \mathbf{w}^T \phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, i = 1, \ldots, l. \end{array} \tag{B.5}$$

The dual problem is

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*}{\text{minimize}} & \frac{1}{2}\left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\right)^T Q \left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\right) + \epsilon \sum\limits_{i=1}^{l} \left(\alpha_i + \alpha_i^*\right) + \sum\limits_{i=1}^{l} z_i(\alpha_i - \alpha_i^*) \\ \text{subject to} & \mathbf{e}^T \left(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\right) = 0, \\ & 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, l, \end{array} \tag{B.6}$$

where $Q_{i,j} \equiv K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$.

After solving problem B.6, the approximate function is

$$z(\mathbf{x}) = \sum_{i=1}^{l} \left(-\alpha_i + \alpha_i^*\right) K(\mathbf{x}_i, \mathbf{x}) + b. \tag{B.7}$$

If one uses the linear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, the approximate function becomes

$$z(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b, \tag{B.8}$$

where $\mathbf{w}$ is given by

$$\mathbf{w} = \sum_{i=1}^{l} \left(-\alpha_i + \alpha_i^*\right) \mathbf{x}_i. \tag{B.9}$$

Appendix C

Convolutional neural networks

Convolutional neural networks (CNNs) are non-linear, composite machine learning models that are commonly used for efficiently processing structured data such as videos, images, and time series. Applications of CNNs are diverse and include: image classification [63], landmarks alignment [22], object detection [113], object segmentation [89], semantic background subtraction [21], face recognition [130], video frame interpolation [102], text-to-speech generation [134], and natural language processing [79]. In Chapters 5 and 6, our applications of CNNs are eyelids distance regression and drowsiness characterization.

C.1 Model architecture

A CNN model is a composition of functions, but this model is preferably formulated as a graph of layers, which is mathematically equivalent but more convenient to work with and design around. In the context of the layer-based formulation, there exist a wide range of layer types and a wide range of model architectures that combine such layers. In the simplest case, the model architecture is a cascade of layers with the four following types of layer:

1. the convolutional layer, which performs a spatial and/or temporal convolution operation;

2. the non-linear activation layer, which performs an element-wise non-linear operation, e.g., the rectified linear unit (ReLU) $\max(0, x)$;

3. the pooling layer, which performs a downsampling operation along spatial and/or temporal dimensions, e.g., average pooling;

4. the fully-connected layer, which performs a matrix product operation.

Each of these layer types has a specific purpose. The convolutional layer efficiently extracts local patterns from the data. The non-linear activation layer evidently introduces non-linearity into the model. The pooling layer reduces the data size, which enables the convolutional layer to cover, in a faster way, larger spatial and/or temporal patterns in data. The fully-connected layer efficiently maps a vector of features to another, which is mainly used nowadays for formatting the output vector. When adequately combined together, these layers make up a powerful non-linear model able to fit well the intricate statistical properties of structured data.

This simple architecture is also the architecture that "VGGNet" [127] uses, with convolutions over 3×3 pixel windows and pooling operations over 2×2 pixel windows. More complex architectures exist, such as "GoogleLeNet" [129], "Xception" [27], and "ResNet" [63].

C.2 Model training

Typically, a CNN model is trained in a iterative and supervised manner from a set of training data. More specifically, training a CNN consists in the repetition of the following two steps. In the first step, one computes, via the backpropagation algorithm, the gradient of the loss function, i.e., the partial derivative of the loss function *w.r.t.* each of the parameters (of the CNN model). This gradient is generally evaluated for a mini-batch (i.e., a small subset) of training data. Doing so speeds up the model convergence considerably since the mini-batch gradient approximates relatively well the full gradient (i.e., evaluated for all of the training data) due to the actual correlation between training data. In the second step, one updates the parameters via gradient descent, i.e., by subtracting a fraction (*a.k.a.* the learning rate) of the mini-batch gradient from the current estimate of the parameters.

In practice, this two-step training procedure has shown to produce models that generalize well to previously unseen data (thus not used for training), which is the fundamental goal of machine learning. However, such high-dimensional optimization has many pitfalls that may prevent good generalization. Indeed, good generalization may require a large amount of training data, especially when the task to learn is complex, e.g., the classification of non-rigid objects in an image. If the number of training of data is not large enough, a classic approach for artificially increasing this number is to “augment”, e.g., randomly rotate, crop, and scale, the training data. Furthermore, good generalization is greatly dependent on the initialization of the parameters due to vanishing/exploding gradient, especially when the model architecture is composed of many layers. Therefore, a good initialization strategy is often required. However, this sensitivity to initialization can be mitigated by using a residual architecture (e.g., “ResNet”) and/or normalization layers [74].

Bibliography

[1] C. Ahlström, A. Anund, C. Fors, and T. Åkerstedt. The effect of daylight versus darkness on driver sleepiness: a driving simulator study. *Journal of Sleep Research*, 27(3):1–9, June 2018. 7

[2] T. Åkerstedt. Consensus statement: Fatigue and accidents in transport operations. *Journal of Sleep Research*, 9(4):395, December 2000. 7

[3] T. Åkerstedt, C. Bassetti, F. Cirignotta, D. García-Borreguero, M. Gonçalves, J. Horne, D. Léger, M. Partinen, T. Penzel, P. Philip, and J. Verster. Sleepiness at the wheel: white paper. Technical report, ASFA and INSV, Paris, France, September 2013. 6, 7, 8, 9

[4] T. Åkerstedt and M. Gillberg. Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, 52(1-2):29–37, June 1990. 2, 5, 8, 11, 13, 16, 23, 30

[5] American Association Automobile Foundation for Traffic Safety. 2017 traffic safety culture index. Technical report, AAA Foundation for Traffic Safety, Washington, DC, USA, March 2018. 2

[6] A. Anund, G. Kecklund, B. Peters, and T. Åkerstedt. Driver sleepiness and individual differences in preferences for countermeasures. *Journal of Sleep Research*, 17(1):16–22, March 2008. 9

[7] A. Anund, G. Kecklund, B. Peters, A. Forsman, L. Arne, and T. Åkerstedt. Driver impairment at night and its relation to physiological sleepiness. *Scandinavian Journal of Work, Environment & Health*, 34(2):142–150, April 2008. 5, 7, 11, 14, 33

[8] K. Armstrong, P. Obst, T. Banks, and S. Smith. Managing driver fatigue: education or motivation? *Road & Transport Research*, 19(3):14–20, September 2010. 9

[9] T. Baltrušaitis. *Automatic facial expression analysis.* PhD thesis, University of Cambridge, March 2014. 101, 104

[10] S. Banks, J. Dorrian, M. Basner, and D. Dinges. Sleep deprivation. In *Principles and Practice of Sleep Medicine*, chapter 5, pages 49–55. Elsevier, Philadelphia, PA, USA, sixth edition, December 2017. 1, 3, 5, 6, 7, 8, 12, 23

[11] J. Baranski, R. Pigeau, and R. Angus. On the ability to self-monitor cognitive performance during sleep deprivation: a calibration study. *Journal of Sleep Research*, 3(1):36–44, March 1994. 5, 8, 12, 13

[12] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, September 2006. 19, 24

[13] M. Basner and D. Dinges. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep*, 34(5):581–591, May 2011. 3, 12, 23, 29, 52, 67, 86

[14] M. Basner, D. Mollicone, and D. Dinges. Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta Astronautica*, 69(11):949–959, 2011. 3, 12, 23

[15] M. Basner, H. Rao, N. Goel, and D. Dinges. Sleep deprivation and neurobehavioral dynamics. *Current Opinion in Neurobiology*, 23(5):854–863, October 2013. 5, 7

[16] National Transportation Safety Board. NTSB 2016 most wanted transportation safety improvements: Reduce fatigue-related accidents, 2016. 1

[17] National Transportation Safety Board. NTSB 2017–2018 most wanted transportation safety improvements: Reduce fatigue-related accidents, 2017. 1, 2

[18] M. Bonnet and S. Moore. The threshold of sleep: perception of sleep as a function of time asleep and auditory threshold. *Sleep*, 5(3):267–276, 1982. 9

[19] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, New York, NY, USA, 1992. ACM. 109

[20] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 25(11):120, 122–125, November 2000. 46, 62, 77

[21] M. Braham, S. Piérard, and M. Van Droogenbroeck. Semantic background subtraction. In *IEEE International Conference on Image Processing (ICIP)*, pages 4552–4556, Beijing, China, September 2017. 111

[22] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision (ICCV)*, pages 1021–1030, Venice, Italy, October 2017. 77, 111

[23] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, April 2014. 77, 107

[24] R. Carpenter. Oculomotor procrastination. In *Eye movements: Cognition and Visual Perception*, chapter 4, pages 237–246. Lawrence Erlbaum, Hillsdale, NJ, USA, January 1981. 12, 66, 76

[25] M. Carskadon and W. Dement. Sleep tendency: an objective measure of sleep loss. *Sleep Research*, 6:200, 1977. 6

[26] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 52, 72, 88

[27] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. 111

[28] The Presidential Commission. *Report of the Presidential Commission on the Space Shuttle Challenger Accident – Appendix G: human factor analysis*, volume II. Government Printing Office, Washington, DC, USA, January 1986. 2

[29] U.S. Nuclear Regulatory Commission. Investigation into the March 28, 1979 Three Mile Island accident by Office of Inspection and Enforcement (Investigative Report No. 50-320/79-10). Technical Report NUREG–0600, U.S. NRC, Washington, DC, USA, July 1979. 2

[30] U.S. Nuclear Regulatory Commission. Report on the accident at the Chernobyl nuclear power station. Technical Report NUREG–1250, U.S. NRC, Washington, DC, USA, January 1987. 2

[31] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001. 77, 102

[32] T. Cootes and C. Taylor. Active shape models — 'smart snakes'. In *British Machine Vision Conference (BMVC)*, pages 266–275, Leeds, England, September 1992. 102

[33] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, Manchester, UK, March 2004. 106

[34] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 20, 24

[35] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. 109

[36] R. Dahlhaus. Locally stationary processes. *CoRR*, abs/1109.4174, September 2011. 76

[37] D. Dawson and K. Reid. Fatigue, alcohol and performance impairment. *Nature*, 388(6639):235, July 1997. 1, 7

[38] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977. 103

[39] Association des sociétés françaises d'autoroutes. Key figures. Technical report, ASFA, Paris, France, 2018. 1

[40] D. Dinges, M. Mallis, G. Maislin, and J. Powell. Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. Technical Report DOT HS 808 762, NHTSA, Washington, DC, USA, April 1998. 2, 12, 13, 14, 23

[41] D. Dinges, M. Mallis, G. Maislin, and J. Powell. PERCLOS, a valid psychophysiological measure of alertness as assessed by psychomotor vigilance. Technical Report FHWA-MCRT-98-006, FHWA, Washington, DC, USA, October 1998. 2, 13, 14, 23

[42] D. Dinges, F. Pack, K. Williams, K. Gillen, J. Powell, G. Ott, C. Aptowicz, and A. Pack. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep*, 20(4):267–277, April 1997. 3, 7, 12, 23

[43] H. Van Dongen, G. Maislin, J. Mullington, and D. Dinges. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep*, 26(2):117–126, March 2003. 7, 8, 27, 32

[44] J. Dorrian, N. Rogers, and D. Dinges. Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In *Sleep Deprivation: Clinical Issues, Pharmacology, and Sleep Loss Effects*, chapter 4, pages 39–70. Marcel Dekker, January 2005. 5, 12, 29, 31, 38, 52, 67, 86

[45] P. Ebrahim, A. Abdellaoui, W. Stolzmann, and B. Yang. Eyelid-based driver state classification under simulated and real driving conditions. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3190–3196, San Diego, CA, USA, October 2014. 17, 20, 21, 22, 24, 25, 32

[46] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, CA, USA, 1978. 19

[47] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, (9):1871–1874, August 2008. 52, 72

[48] S. Fisher, R. Foster, and S. Peirson. The circadian control of sleep. In *Circadian Clocks*, volume 217 of *Handbook of Experimental Pharmacology*, pages 157–183. Springer Berlin Heidelberg, Berlin, Germany, March 2013. 10

[49] National Center for Statistics and Analysis. Traffic safety facts crash stats: Drowsy driving. Technical Report DOT HS 811 449, NHTSA, Washington, DC, USA, March 2011. 1

[50] P. Forsman, B. Vila, R. Short, C. Mott, and H. Van Dongen. Efficient driver drowsiness detection at moderate levels of drowsiness. *Accident Analysis & Prevention*, 50(Supplement C):341–350, 2013. 13

[51] C. François, T. Hoyoux, T. Langohr, J. Wertz, and J. Verly. Tests of a new drowsiness characterization and monitoring system based on ocular parameters. *International Journal of Environmental Research and Public Health*, 13(2):174, January 2016. 18, 21, 24, 25

[52] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, January 1975. 104

[53] I. García, S. Bronte, L. M. Bergasa, J. Almazán, and J. Yebes. Vision-based drowsiness detector for real driving conditions. In *IEEE Intelligent Vehicles Symposium*, pages 618–623, Alcala de Henares, Spain, June 2012. 18, 19, 21, 22, 24, 25

[54] M. Gillberg, G. Kecklund, and T. Åkerstedt. Relations between performance and subjective ratings of sleepiness during a night awake. *Sleep*, 17(3):236–241, April 1994. 5, 8, 16, 30

[55] M. Gillberg, G. Kecklund, and T. Åkerstedt. Sleepiness and performance of professional drivers in a truck simulator — comparisons between day and night driving. *Journal of Sleep Research*, 5(1):12–15, March 1996. 5, 8, 11, 34

[56] H. Godthelp, P. Milgram, and G. Blaauw. The development of a time-related measure to describe driving strategy. *Human Factors*, 26(3):257–268, June 1984. 13

[57] N. Goel, H. Rao, J. Durmer, and D. Dinges. Neurocognitive consequences of sleep deprivation. *Seminars in Neurology*, 29(04):320–339, September 2009. 1, 7

[58] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 1–8, Amsterdam, The Netherlands, September 2008. 79

[59] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, May 2010. 65

[60] S. Gupta and S. Mittal. Yawning and its physiological significance. *International Journal of Applied and Basic Medical Research*, 3(1):11–15, January 2013. 14

[61] D. Hallvig, A. Anund, C. Fors, G. Kecklund, J. Karlsson, M. Wahde, and T. Åkerstedt. Sleepy driving on the real road and in the simulator—a comparison. *Accident Analysis & Prevention*, 50:44–50, January 2013. 31, 32

[62] Y. Harrison and J. Horne. The impact of sleep deprivation on decision making: A review. *Journal of Experimental Psychology: Applied*, 6(3):236–249, September 2000. 5, 7, 13

[63] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. 111

[64] R. He, B. Hu, X. Yuan, and L. Wang. M-estimators and half-quadratic minimization. In *Robust Recognition via Information Theoretic Learning*, chapter 2, pages 3–11. Springer, Cham, Switzerland, July 2014. 105

[65] J. Herbert. Report of the committee on methods of clinical examination in electroencephalography: 1957. *Electroencephalography and Clinical Neurophysiology*, 10(2):370–375, May 1958. 30

[66] J. Higgins, J. Michael, R. Austin, T. Åkerstedt, H. Van Dongen, N. Watson, C. Czeisler, A. Pack, and M. Rosekind. Asleep at the wheel—the road to addressing drowsy driving. *Sleep*, 40(2):1–9, 2017. 1

[67] E. Hoddes, V. Zarcone, H. Smythe, R. Phillips, and W. Dement. Quantification of sleepiness: A new approach. *Psychophysiology*, 10(4):431–436, July 1973. 16

[68] J. Horne and S. Baulk. Awareness of sleepiness when driving. *Psychophysiology*, 41(1):161–165, January 2004. 8

[69] J. Horne and L. Reyner. Driver sleepiness. *Journal of Sleep Research*, 4(s2):23–29, December 1995. 6, 9, 33

[70] J. Horne and L. Reyner. Counteracting driver sleepiness: Effects of napping, caffeine, and placebo. *Psychophysiology*, 33(3):306–309, May 1996. 9, 74

[71] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger. Densely connected convolutional networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Honolulu, HI, USA, July 2017. 64

[72] X.-P. Huynh, S.-M. Park, and Y.-G. Kim. Detection of driver drowsiness using 3D deep neural network and semi-supervised gradient boosting machine. In *Asian Conference on Computer Vision Workshops (ACCV Workshops)*, volume 10118 of *Lecture Notes in Computer Science*, pages 134–145, Cham, Switzerland, March 2017. Springer. 18, 20, 21, 24, 25

[73] Belgian Road Safety Institute. Sleepy at the wheel: analysis of the extent and characteristics of sleepiness among Belgian car drivers. Technical Report 2015-R-06-EN, BRSI, Brussels, Belgium, July 2015. 2

[74] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, Lille, France, July 2015. 63, 78, 112

[75] M. Johns. A new perspective on sleepiness. *Sleep and Biological Rhythms*, 8(3):170–179, July 2010. 6, 10, 17

[76] M. Johns, A. Tucker, R. Chapman, K. Crowley, and N. Michael. Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. *Somnologie - Schlafforschung und Schlafmedizin*, 11(4):234–242, December 2007. 21, 24

[77] L. Johnson, C. Freeman, C. Spinweber, and S. Gomez. Subjective and objective measures of sleepiness: Effect of benzodiazepine and caffeine on their relationship. *Psychophysiology*, 28(1):65–71, January 1991. 8

[78] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa. Validation of the karolinska sleepiness scale against performance and EEG variables. *Clinical Neurophysiology*, 117(7):1574–1581, 2006. 8, 16

[79] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *Annual Meeting of the Association for Computational Linguistics*, pages 655–665, Baltimore, MD, USA, June 2014. Association for Computational Linguistics. 111

[80] A. Kasiński, A. Florek, and A. Schmidt. The PUT face database. *Image Processing & Communication*, 13(3):59–64, January 2008. 79

[81] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, Columbus, OH, USA, June 2014. 24, 62, 77, 83, 107

[82] D. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, January 2009. 62, 83

[83] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, December 2014. 68, 83

[84] A. Kircher, M. Uddman, and J. Sandin. Vehicle control and drowsiness. Technical report, VTI, Linköping, Sweden, May 2002. 12, 13, 23

[85] W. Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2–3):169–195, April 1999. 37

[86] J. Leach and R. Carpenter. Saccadic choice with asynchronous targets: evidence for independent randomisation. *Vision Research*, 41(25):3437–3445, November 2001. 76

[87] Y. Liang, W. Horrey, M. Howard, M. Lee, C. Anderson, M. Shreeve, C. O'Brien, and C. Czeisler. Prediction of drowsiness events in night shift workers during morning driving. *Accident Analysis & Prevention*, online first, November 2017. 17, 18, 21, 24, 25, 32

[88] H.-O. Lisper, H. Laurell, and J. van Loon. Relation between time to falling asleep behind the wheel on a closed track and changes in subsidiary reaction time during prolonged driving on a motorway. *Ergonomics*, 29(3):445–453, 1986. PMID: 3698972. 1, 8, 12, 14, 15, 23

[89] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, Boston, Massachusetts, USA, June 2015. 111

[90] O. Lowenstein, R. Feinberg, and I. Loewenfeld. Pupillary movements during acute and chronic fatigue: A new test for the objective evaluation of tiredness. *Investigative Ophthalmology & Visual Science*, 2(2):138–157, April 1963. 14

[91] Q. Massoz, T. Langohr, C. François, and J. Verly. The ULg multimodality drowsiness database (called DROZY) and examples of use. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–7, Lake Placid, NY, USA, March 2016. 4, 30, 42, 44

[92] Q. Massoz and J. Verly. Vision-based system for monitoring vehicle operator responsiveness from face images. In *International Conference on Managing Fatigue*, pages 1–3, San Diego, CA, USA, March 2017. 4, 75

[93] Q. Massoz, J. Verly, and M. Van Droogenbroeck. Multi-timescale drowsiness characterization based on a video of a driver's face. *Sensors*, 18(9):1–17, August 2018. 4, 31, 42, 61

[94] H. Matsuo and A. Khiat. Prediction of drowsy driving by monitoring driver's behavior. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 3390–3393, November 2012. 15, 18, 19, 21, 24, 25

[95] L. Michael, S. Passmann, and R. Becker. Electrodermal lability as an indicator for subjective sleepiness during total sleep deprivation. *Journal of Sleep Research*, 21(4):470–478, August 2012. 8, 12, 16

[96] M. Mitler, K. Gujavarty, and C. Browman. Maintenance of wakefulness test: a polysomnographic technique for evaluating treatment efficacy in patients with excessive somnolence. *Electroencephalography and Clinical Neurophysiology*, 53(6):658–661, June 1982. 6

[97] H. Moller, L. Kayumov, E. Bulmash, J. Nhan, and C. Shapiro. Simulator performance, microsleep episodes, and subjective sleepiness: normative data using convergent methodologies to assess driver drowsiness. *Journal of Psychosomatic Research*, 61(3):335–342, September 2006. 8

[98] T. Monk. A visual analogue scale technique to measure global vigor and affect. *Psychiatry Research*, 27(1):89–99, January 1989. 16

[99] A. Muzet, T. Pébayle, J. Langrognet, and S. Otmani. AWAKE pilot study no. 2: Testing steering grip sensor measures. Technical Report IST-2000- 28062, Center for European Policy Analysis (CEPA), 2003. 11

[100] National Highway Traffic Safety Administration. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Technical Report DOT HS 810 594, NHTSA, Washington, DC, USA, April 2006. 1

[101] National Highway Traffic Safety Administration. Asleep at the wheel: A national compendium of efforts to eliminate drowsy driving. Technical Report DOT HS 812 352, NHTSA, Washington, DC, USA, March 2017. 1

[102] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive separable convolution. In *International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017. 111

[103] J. Nishiyama, K. Tanida, M. Kusumi, and Y. Hirata. The pupil as a possible premonitor of drowsiness. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1586–1589, Lyon, France, August 2007. 14

[104] R. Nopsuwanchai, Y. Noguchi, M. Ohsuga, Y. Kamakura, and Y. Inoue. Driver-independent assessment of arousal states from video sequences based on the classification of eyeblink patterns. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 917–924, Beijing, China, October 2008. 10, 15, 18, 20, 21, 24, 25

[105] M. Ohayon, R. Priest, J. Zulley, S. Smirne, and T. Paiva. Prevalence of narcolepsy symptomatology and diagnosis in the European general population. *Neurology*, 58(12):1826–1833, June 2002. 7

[106] J. Owens, T. Dingus, F. Guo, Y. Fang, M. Perez, J. McClafferty, and B. Tefft. Prevalence of drowsy driving crashes: Estimates from a large-scale naturalistic driving study. Technical report, AAA Foundation for Traffic Safety, Washington, D.C., USA, 2018. 1

[107] P. Philip, P. Sagaspe, J. Taillard, N. Moore, C. Guilleminault, M. Sanchez-Ortuno, T. Åkerstedt, and B. Bioulac. Fatigue, sleep restriction, and performance in automobile drivers: a controlled study in a natural environment. *Sleep*, 26(3):277–280, May 2003. 8, 31

[108] P. Philip, P. Sagaspe, J. Taillard, C. Valtat, N. Moore, T. Åkerstedt, A. Charles, and B. Bioulac. Fatigue, sleepiness, and performance in simulated versus real driving conditions. *Sleep*, 28(12):1511–1516, December 2005. 7, 31

[109] A. Picot, S. Charbonnier, and A. Caplier. Monitoring drowsiness on-line using a single encephalographic channel. In *Recent Advances in Biomedical Engineering*, In-Tech, pages 145–164. Carlo Alexandre Barros de Mello, December 2009. 11

[110] A. Picot, S. Charbonnier, and A. Caplier. On-line detection of drowsiness using brain and visual information. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 42(3):764–775, May 2012. 11

[111] N. Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):136–143, February 2008. 7

[112] A. Rechtschaffen and A. Kales. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. National Institutes of Health Publication, Los Angeles, CA, USA, 1968. 29

[113] J. Redmon, S. Divvala, R. Gorshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2016. 111

[114] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1692, Columbus, OH, USA, June 2014. 77, 107

[115] L. Reyner and J. Horne. Evaluation of "in-car" countermeasures to sleepiness: Cold air and radio. *Sleep*, 21(1):46–50, February 1998. 9

[116] L. Reyner and J. Horne. Falling asleep whilst driving: Are drivers aware of prior sleepiness? *International Journal of Legal Medicine*, 111(3):120–123, February 1998. 8, 9

[117] T. Roehrs, M. Carskadon, W. Dement, and T. Roth. Daytime sleepiness and alertness. In *Principles and Practice of Sleep Medicine*, chapter 4, pages 39–48. Elsevier, Philadelphia, PA, USA, sixth edition, December 2017. 5, 6, 7, 8, 32

[118] T. Roth. Insomnia: Definition, prevalence, etiology, and consequences. *Journal of Clinical Sleep Medicine*, 3(5 Suppl):S7–S10, August 2007. 7

[119] D. Royal. National survey of distracted and drowsy driving attitudes and behavior: 2002. Technical Report DOT HS 809 566, NHTSA, Washington, DC, USA, April 2013. 2

[120] T. Rupp, N. Wesensten, and T. Balkin. Trait-like vulnerability to total and partial sleep loss. *Sleep*, 35(8):1163–1172, August 2012. 7

[121] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, January 2011. 24, 30, 77, 101, 102, 103

[122] R. Schleicher, N. Galley, S. Briest, and L. Galley. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010, June 2008. 2, 8, 11, 13, 14, 23

[123] E. Schmidt, M. Schrauf, M. Simon, M. Fritzsche, A. Buchner, and W. Kincses. Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving. *Accident Analysis & Prevention*, 41(5):1087–1093, 2009. 8

[124] H. Schulz, S. Volk, and A. Yassouridis. Measuring tiredness by symptoms. *Sleep Research*, 20:515, 1991. 16

[125] T.-H. Shih and C.-T. Hsu. MSTN: Multistage spatial-temporal network for driver drowsiness detection. In *Asian Conference on Computer Vision Workshops (ACCV Workshops)*, volume 10118 of *Lecture Notes in Computer Science*, pages 146–153, Cham, Switzerland, March 2017. Springer. 18, 20, 21, 24, 25

[126] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 63

[127] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, pages 1–14, San Diego, CA, USA, May 2015. 20, 24, 78, 111

[128] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, January 2014. 68, 80

[129] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA, June 2016. 111

[130] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, June 2014. 111

[131] B. Tefft. Prevalence of motor vehicle crashes involving drowsy drivers, United States, 2009–2013. Technical report, AAA Foundation for Traffic Safety, Washington, DC, USA, November 2014. 1

[132] M. Thorpy and B. Michel. *Sleepiness: Causes, consequences and treatment*. Cambridge University Press, January 2011. 1

[133] T. Tieleman and G. Hinton. Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012. 65, 80

[134] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. 111

[135] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998. 110

[136] W. Verwey and D. Zaidel. Predicting drowsiness accidents from personal attributes, eye blinks and ongoing driving behaviour. *Personality and Individual Differences*, 28(1):123–142, January 2000. 2, 13, 23

[137] J. Vicente, P. Laguna, A. Bartra, and R. Bailón. Drowsiness detection using heart rate variability. *Medical & Biological Engineering & Computing*, 54(6):927–937, June 2016. 11

[138] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, Kauai, HI, USA, December 2001. 46, 62, 77, 87

[139] E. Vural, M. Bartlett, G. Littlewort, M. Cetin, A. Ercil, and J. Movellan. Discrimination of moderate and acute drowsiness based on spontaneous facial expressions. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 3874–3877, August 2010. 15, 17, 19, 21, 24, 25

[140] E. Vural, M. Çetin, A. Erçil, G. Littlewort, M. Bartlett, and J. Movellan. Machine learning systems for detecting driver drowsiness. In *In-Vehicle Corpus and Signal Processing for Driver Behavior*, chapter 8, pages 97–110. Springer, Boston, MA, USA, 2009. 14, 15, 17, 19, 21, 24, 25

[141] X. Wang and C. Xu. Driver drowsiness detection based on non-intrusive metrics considering individual specifics. *Accident Analysis & Prevention*, 95(Part B):350–357, October 2016. 17, 19, 21, 24, 25

[142] N. Watson, M. Badr, G. Belenky, D. Bliwise, O. Buxton, D. Buysse, D. Dinges, J. Gangwisch, M. Grandner, C. Kushida, R. Malhotra, J. Martin, S. Patel, S. Quan, and E. Tasali. Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *Sleep*, 38(6):843–844, June 2015. 1, 6, 7

[143] C.-H. Weng, Y.-H. Lai, and S.-H. Lai. Driver drowsiness detection via a hierarchical temporal deep belief network. In *Asian Conference on Computer Vision Workshops (ACCV Workshops)*, volume 10118 of *Lecture Notes in Computer Science*, pages 117–133, Cham, Switzerland, March 2017. Springer. 18, 19, 21, 24, 25

[144] W. Wierwille and L. Ellsworth. Evaluation of driver drowsiness by trained raters. *Accident Analysis & Prevention*, 26(5):571–581, October 1994. 10, 13, 15, 23

[145] W. Wierwille, L. Ellsworth, S. Wreggit, R. Fairbanks, and C. Kirn. Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness. Technical Report DOT HS 808 247, NHTSA, Washington, DC, USA, December 1994. 13, 14

[146] B. Wilhelm, H. Wilhelm, H. Lüdtke, P. Streicher, and M. Adler. Pupillographic assessment of sleepiness in sleep-deprived healthy subjects. *Sleep*, 21(3):258–265, June 1998. 14

[147] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, chapter 11, pages 355–396. CRC Press, January 1999. 19, 24

[148] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, June 2013. 19, 24

[149] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, Las Vegas, NV, USA, June 2016. 77, 107