Multimodal fusion of video signals for remote evaluation of emotional/cognitive processing

Ferrer Ferrer, Mar

Curs 2019-2020

Director(s): Federico M. Sukno, Alex Pereda

GRAU EN ENGINYERIA DE SISTEMES AUDIOVIDUALS

upf. Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Treball de Fi de Grau

Multimodal fusion of video signals for remote evaluation of emotional/cognitive processing

Mar Ferrer Ferrer

BACHELOR THESIS UPF / YEAR 2019-2020

THESIS DIRECTOR(S):

Federico M. Sukno; Departament CMT
Alex Pereda; Data Science and Big Data Analytics Unit, Eurecat

Acknowledgements

First, I would like to thank my supervisors Federico Sukno and Alex Pereda for their help and guidance throughout this project.

I am especially grateful to my parents, Carlos and Cati, for their unconditional love and support throughout the years.

Finally, I would also like to thank Mireia and Adrià, for their genuine help, advice and infinite moral support.

Abstract

Remote photoplethysmography consists in measuring the blood pulse of a subject given a video signal of, for example, his/her face. The progress made in this area has been very beneficial for human emotion study from the physiological perspective, since the traditional measures such as the electrocardiogram (ECG) are very intrusive. However, extracting pulse signals from video is very challenging due to the noise caused by motion and illumination

This project consists on two main parts: the first one is the pulse extraction from video: how to detect the region of interest (ROI), how to filter out the noise produced by the illumination changes, and finally, how to filter and detrend the signal succesfully. The second part consists on the emotion estimation given the pulse signal obtained before. Those emotions are quantified in the Valence-Arousal bidimensional space, which tries to represent as faithfully as possible the human emotions. Finally a support-vector machine (SVM) model has been trained to predict those emotions.

Resum

La fotopletismografía remota consisteix en mesurar el pols sanguini mitjançant un senyal de vídeo del subjecte en qüestió. L'estudi de les emocions des de la perspectiva dels senyals fisiològics s'ha beneficiat enormement dels avenços en aquest camp, ja que podria deixar enrere tècniques intrusives com l'electrocardiograma. Tot i això, extraure el pols del senyal de vídeo és una tasca complicada degut a la poca intensitat d'aquest i al soroll causat pels canvis d'il·luminació i els moviments del subjecte.

Aquest projecte consta de dues parts diferenciades: la primera és l'extracció del pols sanguini a partir de vídeos de la cara dels subjectes. En aquesta part es presentarà

com detectar la regió d'interès, com eliminar el soroll produït per les variacions de lumíniques i es farà un estudi extensiu dels mètodes de filtrat i de *detrending* del senyal. La segona part consisteix en l'estimació de les emocions dels subjectes a partir del pols extret. Aquestes emocions estan quantificades amb mesures de València i Excitació les quals intenten representar, en un espai bidimensional, les emocions humanes. Finalment, s'ha entrenat un model de *support-vector machine* per a poder predir aquestes emocions.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

In the last decade, there has been a constant and progressive change towards the wireless world, where cables are an option, rather than a necessity. This comes from the dependency that we, as humans, have developed towards many devices, which has driven us to find a way to simplify their usage. But in some fields, such as the medical one, it is very complicated to pivot to wireless equipment, as there is not any room for errors. Even the extremely common activity of heart (HR) monitoring is still being carried out with several cables attached to the human body. Despite that, non-contact HR measurement from facial videos has attracted high interest due to its convenience and many applications. Its main advantage is that eases people's life when they have to be constantly monitored, by relieving them from the restrictions that an electrocardiogram (ECG) presents, as it is measured by attaching adhesive patches to the skin. These sensors are not convenient for long-term wear and wearing them can become uncomfortable over time. In addition, these devices can damage the fragile skin of premature newborns or elderly people. For these populations especially, a non-contact means of detecting pulse could be very beneficial.

Besides, commercial cameras such as webcams, surveillance cameras or mobile phone

cameras can be found almost everywhere, so taking as example the current situation (COVID19 pandemic), lots of patients could still do their regular checks via video call. Moreover, it opens the door to the remote analysis of human behaviour from the physiological perspective. For example, Picard et al.[4] proved that physiological signals are more pertinent than other modalities when detecting genuine emotions, since they are originated in the peripheral nervous system, which means that they cannot be hidden.

At first, the main motivation for exploring this area from the non-invasive point of view was that many people might feel uncomfortable when monitored, so the results obtained with the acquired data might not translate well to "real" situations. Nevertheless, a new motivation has been fuelling this work, and it is due to the current health crisis, which has let us see the huge necessity of an immediate application of remote methods to estimate physiological signals such as HR. In particular, remote emotion recognition would be useful in fields such as teaching, for online classes, and telemedicine, among others.

In this bachelor thesis we attempt to perform an extensive study on different HR estimation methods using video signals, and analyse which ones give the best result when it comes to emotion estimation.

1.1 Emotion detection from the physiological perspective

Emotion can be perceived from many different body signals, for example, the first that might come to mind is the analysis of facial expressions [5] [6] [7], which is the most common method due to its intuitiveness (we do it all the time), and also, there is a lot of literature and it is quite developed. Furthermore, the tone and pitch of the voice are another very straightforward indicator of a certain feeling [8] [9]. Also, the body posture can indicate comfort or discomfort, the predisposition to do something or if the subject is nervous, among others [10].

Having that said, why should emotions be studied from the physiological perspective if there are plenty available methods that seem more intuitive and straightforward? In fact, all the measures above have one thing in common, they can be distorted by the subject, so if he/she is talented enough, will be able to deceive the interlocutor [11]. Also, for people with difficulties expressing their feelings verbally or physically, like autistic people, the above measures might not be of any use [12][3].

1.2 Previous work

Human feelings are studied from the point of view of several physiological measures (see Appendix A). Koeltstra *et al.* (2012) [13] studied face video, electroencephalography (EEG) and peripheral physiological signals: blood volume pulse (BVP), galvanic skin response (GSR), respiratory volume (RESP), skin temperature (SKT), electrooculogram (EOC) and electromyogram (EMG). The DEAP database [14] was

used, which consists of a series of physiological measurements, recorded from several subjects that where induced to different emotional states by watching video clips fragments (see Section. 2.1). Emotions were classified in terms of valence, arousal and dominance. Numerically, on valence and arousal obtained from peripheral measures, they achieved 57% and 62.7% of accuracy respectively. This work constitutes a reference benchmark for emotion classification based on physiological measurements.

From the same year, Soleyemani *et al.* (2012)[15] classified feelings also in the valence-arousal space, but using the MAHNOB-HCI database (that follows the same idea than the DEAP database), which will be used later in this project. For all signals (ECG, EEG, Respiration Rate, SKT, GSR and eye gaze), they reached an accuracy of 45.5% for valence and a 46.2% for arousal (except for eye gaze). However, the best percentages were obtained using a combination of signals: EEG and eye gaze, with a 76.1% for valence and a 67.7% for arousal.

More recently, Santamaria *et al.* (2018) [16] approached the valence arousal problem using Neural Networks, using Convolutional Neural Networks. Fernandez *et al.* (2016) proposed a combination of physiological signals with facial gestures [17]. Furthermore, in 2019, Chacon *et al.* [2] have addressed the remote emotional processing using video signals, and compared it with the traditional methods like HRV, ECG or EEG, using the DEAP dataset. They have also tried with several Neural Network architectures for the different signals, Recurrent Neural Networks (RNN) were used specifically for the signal extracted from the video, obtaining an accuracy of 57% for valence and a 58% for arousal. The obtained accuracies above correspond to three class classification experiments, so in order to compare our results with the benchmark we will reproduce the same method. Nevertheless, we will also present a nine

class classification, since the ground truth annotations given by the users go from 1 to 9 in both valence and arousal. Furthermore, all of those projects have one thing in common, the valence and arousal tags in the used data have not been annotated in real-time, in fact, they are given at the end of each stimulus by the subjects, which reduces the reliability of the labels. In this project we have performed an intensive tagging in order to deal with this problem.

1.3 Project structure

This project covers several differentiated parts: in Chapter 2 we present the dataset that has been used and how it has been used. Furthermore, there is a brief introduction into emotion characterisation and the model of valence and arousal, which will be mentioned along this document. At the end of this chapter, we also discuss the effect of the recording camera and which ones will give better results.

Then, in Chapter 3, we cover the heart rate extraction from the video. There is an extensive explanation on that matter: from how to better extract it to how to post-process the signal.

Afterwards, in Chapter 4, the training process will be discussed. It includes the necessary pre-processing and an explanation of the used techniques. Furthermore, since we performed a valence-arousal annotation refinement in order to improve the quality of the predictions, it will also be explained in this Chapter.

Finally, Chapter 5 and 6 presents the final results and the conclusions we can ex-

tract from them, respectively.

Chapter 2

MATERIALS AND METHODS

2.1 Database

Since we are performing an evaluation of the different emotion states a human can experience, we need database with such annotations. In this case, the MAHNOB-HCI *multimodal tagging* database had the desired information. It consists of a dataset with physiological tags such as EEG, ECG, eye gaze, etc; and also emotional tags: valence, arousal and emotion which are auto-reported by the subject at the end of each stimulus. Each one of the stimulus consisted of film fragments, specially chosen to induce various emotional states.
Also, the MAHNOB-HCI database is recorded simulating real life conditions: head and body motions, facial expressions and illumination changes, which will challenge and test the soundness of the proposed technique.

Furthermore, there is another database that will be mentioned along this document, as it has been used in many state-of-the-art publications: DEAP. It consists of EEG and peripheral physiological signals annotations plus valence, arousal, like/dislike,

dominance and familiarity. Those last tags are also self-reported, but in this case, the stimulus belong to music videos [14].

2.1.1 Use of the MAHNOB-HCI database

Particularly in this project, we used the ECG measurements as ground truth. Out of a total of 40 subjects, we picked 9 to perform our experiments which correspond to the IDs 1, 2, 3, 4, 5, 6, 8, 9 and 30. Unfortunately, subject 9, had a corrupted ground truth, so we had to remove all the estimations carried out using this subject's footage. Each of this subject underwent around twenty experiments where all of this experiments were recorded separately, so we work with approximately 160 videos. As said above, a small subset of subjects was used, instead of the whole dataset and this is due to the fact that we performed a manual and time consuming data augmentation task, which consisted on annotating more specifically those 160 videos, in terms of valence and arousal.

2.2 Emotion characterisation

There are several studies that have tried to distcretise the feelings we can experience in order to make them more tangible. Most theorists endorse the view that emotions comprise three components: subjective experience (*e.g.*, feeling angry), the expressive component (*e.g.*, severe frown), and the physiological component (*e.g.*, sympathetic nervous system (SNS) activation) [18] [19]. So, in general, positive and negative emotions include behavioural components of approach and withdrawal, respectively (Frijda, 1986)[1].

Eckman's model of emotions is based on the universal expression of emotion based

just on six words: Happiness, Sadness, Surprise, Fear, Anger and Disgust [20], and he suggests that these feelings are innate. There is also Plutchik's model [21] which works with eight fundamental emotions: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger and Anticipation. But sometimes, it is difficult to map a feeling to a word. The proponents of the dimensional view claim that emotions are fundamentally similar in most respects, differing only in terms of one or more dimensions. In this project we have chosen a model that is based on a two-dimensional evaluation: the valence-arousal one developed by Russell [22].

- **Valence:** The valence axis represents the range of affect between pleasure and displeasure, in other words: the degree "good"-ness and the "bad"-ness of a certain event. In Figure 2.1, valence dimension refers to the hedonic quality or pleasantness of an affective experience and ranges from unpleasant to pleasant.

- **Arousal:** It measures the level of excitement or alertness. It has also been conceptually defined as "a drive state or a nonspecific energiser of behaviour, something that describes the intensity of an experience but not its quality" (Duffy, 1962; Mandler, 1992).

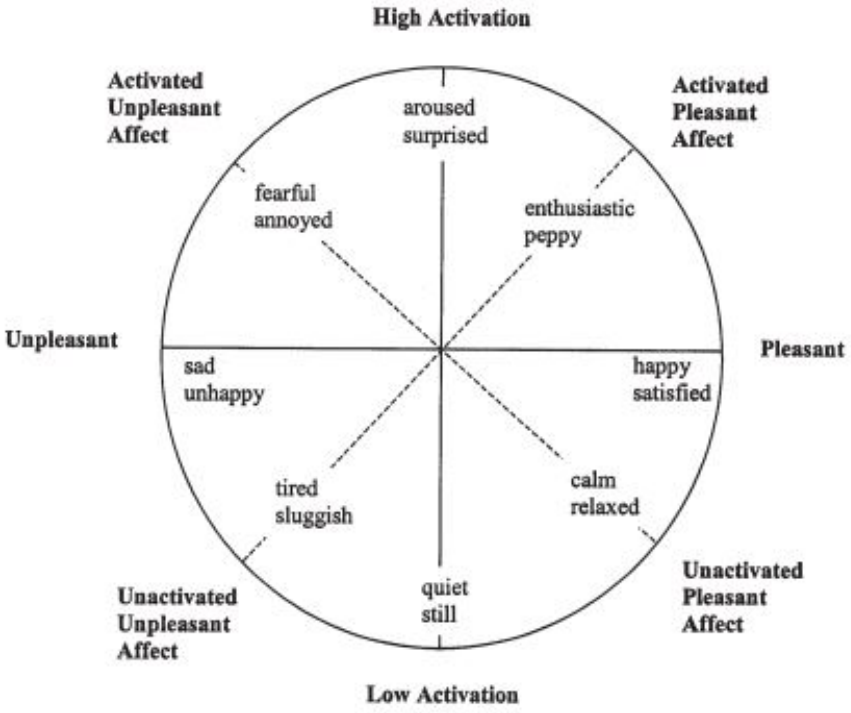


Figure 2.1: Two-dimensional structure of affect wheel [1].

2.3 The impact of the camera

Following a previous bachelor thesis, from Guillem Garcia, our first tests were performed using a dataset recorded with a Basler ace acA640-120uc camera, which is able to store footage in uncompressed formats. Using these data, extracting the pulse was quite straightforward and very accurate results where obtained without any kind of motion or illumination correction. Figure 2.2 shows the estimated pulse using uncompressed footage, where to estimate it, we have just applied a band-pass filter between [0.7,4] Hz.

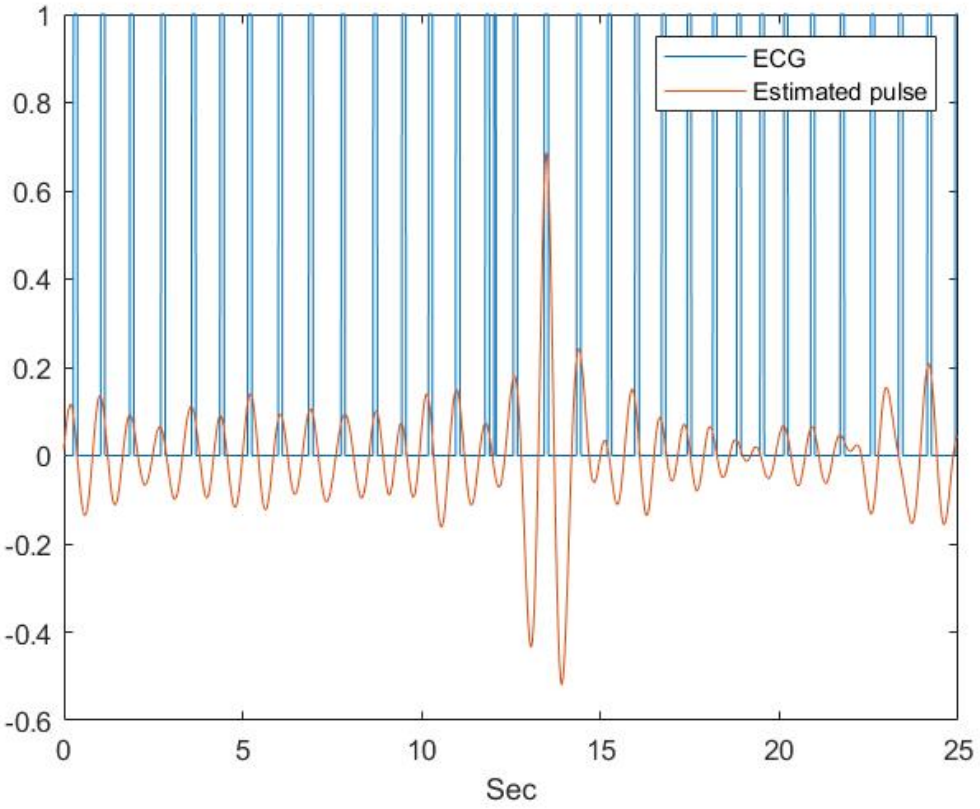


Figure 2.2: ECG ground truth (blue) and estimated pulse (orange).

Nevertheless, this database could not be used because it lacks of the relevant annotations: valence and arousal. Having said that, it is important to note that all the results shown in this thesis would probably be outperformed if the used videos were not compressed.

As mentioned in Section 2.1, we have used the MAHNOB-HCI database, which videos were recorded with a regular three-colour channel camera with a frame rate of 61 fr/sec. It has been proven that the pulse can also be extracted from the colour variations of current cameras [23][24]. Despite the very good results obtained using a non-compression camera, this approach is more realistic since the results we obtain

can be easily reproduced using almost any current camera, making it accessible to everybody and easy to install without many budget adjustments.

Chapter 3

HR ESTIMATION

Traditional heart rate (HR) measurement methods rely on optical or electronic sensors which might be quite uncomfortable and can disturb and contaminate the sample when taking emotional annotations, since the set-up and the devices create an "artificial" environment. That is why researchers are focusing more on non-intrusive methods, as the subject will have an experience "closer to real-life" and the emotions will be evoked more naturally. For this reason, one of the main contributions of this thesis is a detailed explanation on how to estimate HR from facial images.

3.1 Estimate HR from colour information

Heart rate can be detected without contact through photo-plethysmograpy (PPG), which is based on the principle that blood absorbs light more than surrounding tissue, so variations in blood volume affect transmission or reflectance correspondingly. With this principle, blood variations under the skin can be detected remotely using a camera.

Verkruysse *et al.* [24] found that, even though the pulse could be detected from red, green, and blue channels of colour video of exposed skin, it had a predominant presence in the green channel. The cause is the hemoglobin, which is a protein found in in the red blood cells, responsible of the oxygen transportation. It happens that the hemoglobin has absortion peaks for green and yellow light wavelengths, making the green chanel, the most suitable for blood flow estimation.

Even though other studies have been conducted using the green channel [25], Poh *et al.* found that the signal can be extracted with less noise using blind source separation (BSS), in particular, Independent Component Analysis (ICA) [23].

3.2 Independent Component Analysis

We need BSS in every situation where data consist of multiple time series but there is no prior knowledge of the signals that make up the mixture. Also, the BSS assumes that the mixtures are non-Gaussian. ICA is a special case of the BSS problem.

ICA is a multivariate statistical method that seeks to uncover hidden variables in high-dimensional data. In this problem we know the number of components we are looking for, since they belong to each colour channel (three components) or the colour channels plus the gray values (four components), which have to be mutually independent.

In order to apply the ICA algorithm, the samples X must be centered and sphered.

Centering means that the components of X have mean zero, and after sphering them, they are uncorrelated with unit variances. In its most general form, the ICA model assumes that X is generated by

$$X = f(S) + e \tag{3.1}$$

where S is an unobservable m-vector and e represents measurement noise and any other variability that cannot be attributed to the sources. The main goal is to invert f and estimate S.

The simplest approach is to assume that f is linear: $f(S) = AS$, where A is a *mixing* matrix, and that e is zero. So, having that in mind, equation (3.1) can be rewritten:

$$X = AS \tag{3.2}$$

For a given matrix A, there exists an *unmixing* matrix such that $S = WX$ and the sources can be recovered. Since, in practice, A is unknown and our aim is to estimate the *unmixing* matrix just with the observed information X, we will have an estimate of the *unmixing* matrix: $\hat{W}$.

Finally, the source component vector S will be approximated by:

$$\hat{S} = \hat{W}X \tag{3.3}$$

In our case, the pulse variation is the S component, and X is the video information captured by the camera.

Furthermore, in this thesis, we have used the JADE implementation of the ICA algorithm [26]. JADE stands for Joint Approximate Diagonalization of Eigenmatrices. This version separates observed mixed signals into latent source signals by exploiting fourth order moments. The fourth order moments are a measure of non-Gaussianity used as a proxy for defining independence between the source signals.

Morover, since the intuition is: "the more samples adding more information to the mixture, the better the sources will be separated", two methods will be compared in Chapter 5. One experiment will be performed using just the RGB information from the camera, and the other will be computed by adding a grayscale recording captured simultaneously with the RGB recording.

3.3 Extracting the colour information

Even though the cardiac pulse can be recovered from ordinary cameras, the MAHNOB-HCI simulates a realistic human-computer interaction situation, which hinders the task quite a bit, as the change of colour caused by the pulse is very weak in comparison with other factors like illumination changes or head movements.
The proposed framework (Fig. 3.1) will reduce noise that those illumination changes or head motions might introduce in the perceived signal.

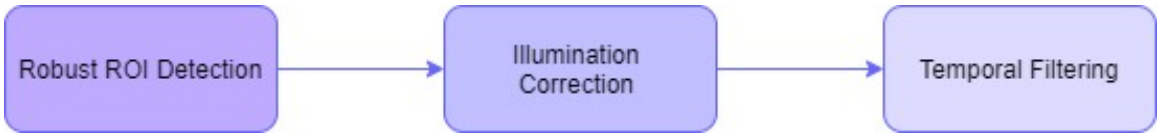


Figure 3.1: Proposed framework

3.3.1 ROIs selection

ROI stands for Region Of Interest and it means the parts of the face that will provide useful information. Also, in this case there are, at least, two ROIs or patches, since features inside the bounding box delimiting the mouth and the noise have to be avoided.

In order to extract the heart rate oscillations, the colour variation in those ROIs along the video has to be computed. Therefore, there are several approaches to consider: in the first one, for each frame, there is just one value for each colour channel. Otherwise, there will be as many colour triplets as patches. Let us compare both options:

A. One colour triplet

The colour oscillations along the video are expressed such that:

$$\mathbf{r} = [r_1, r_2, ..., r_n]$$

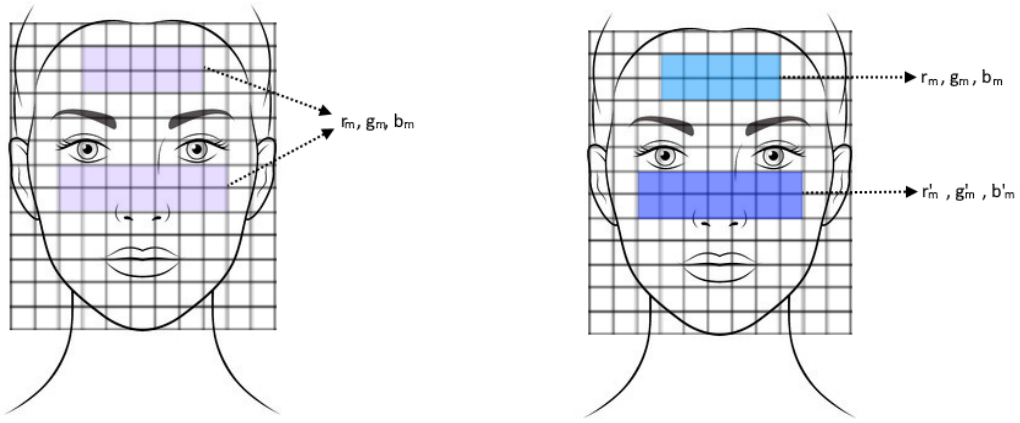
$$\mathbf{g} = [g_1, g_2, ..., g_n]$$

$$\mathbf{b} = [b_1, b_2, ..., b_n]$$

where n is the frame number. For each frame, there is just one value of r, g and b since they are computed by calculating the average per colour channel of all the pixels in the ROIs, as in Figure 3.2a.

B. Multiple colour triplets

In this case, there are as many triplets of colour values as ROIs, in order to see the different inputs of each ROI and study which is the optimal to extract the pulse. However, this method has a main drawback: the number of averaged pixels is smaller, so the present noise is not as mitigated as in the previous method. This problem is accentuated when the ROI is extracted from regions such as the forehead or the chin, since they are specially small (see Figure 3.2b).



(a) Averaging all the coloured pixels in each ROI.

(b) Averaging each ROI separately and obtaining several triplets.

Figure 3.2: Methods A and B for colour estimation, respectively.

Since the pulse is a very weak signal, the noise must be minimised at all costs, so the method B will be discarded. It could be useful for a future work, where non-compression cameras are used.

Furthermore, the ROIs depend on the features of each individual, *i.e*, if he has a beard, even if it is short, the pulse will not be well detected using that part of the face. The same happens with any obstacle between the lens of the camera and the

skin, such as glasses. So, in order to extract the optimal colour patches from the subjects' faces, each one of them has been classified into one of the following groups, depending on what patch combination is optimal for each face.

1. Forehead and centre of the face (nose and cheeks).

2. Chin and centre of the face (nose and cheeks).

In conclusion, for this dataset, all the pixels in the different ROIs are averaged at every frame in order to extract the mean *rgb* values. Furthermore, there are two ROIs per face, and their position will depend on the group that best adjusts to the subject's features (Figure 3.3).

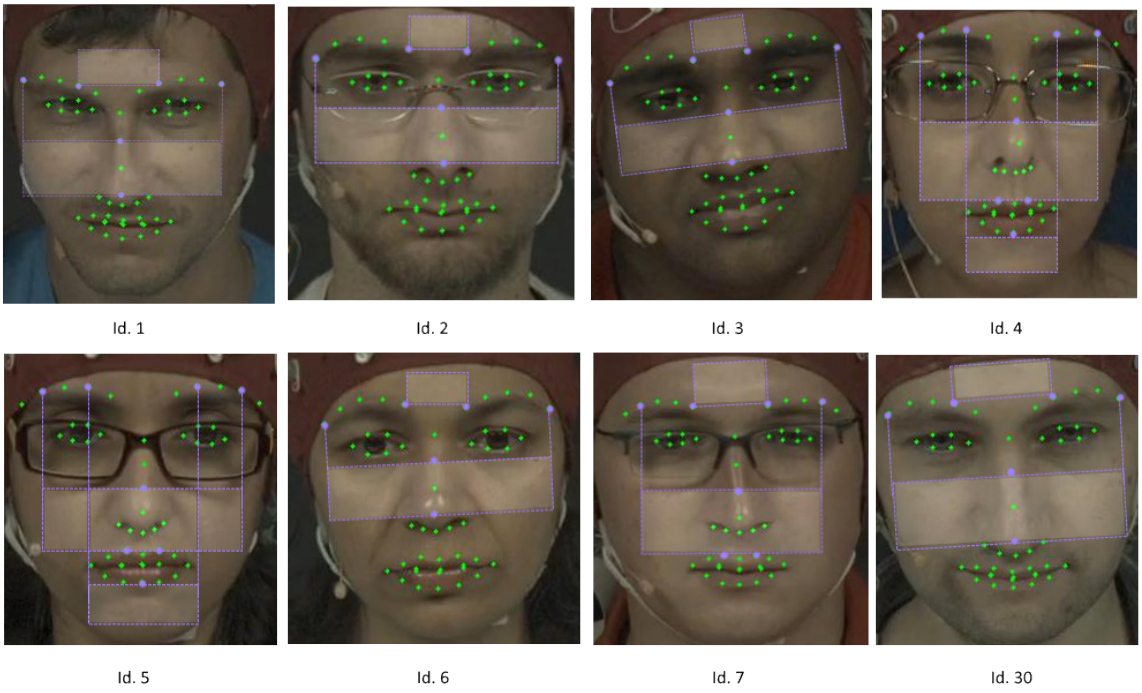


Figure 3.3: Landmarks and ROIs for each of the subjects. Lilac landmarks are the ones that have been used for each subject.

3.3.2 Robust ROI Detection

Furthermore, head movements and facial expressions have to be taken into account, as they can introduce a lot of noise. Hence, instead of using a regular Viola Jones detector, we chose a landmarks method [27] which has a quite robust implementation, making it ideal for facial expressions or some head motions.
Once this is implemented, the ROI's limits are defined with landmarks, so the ROI is adaptable to almost every frame conditions. Figure 3.4 shows how the landmarks are distributed and how the ROIs are extracted in three different frames, which present some kind of head movement and different facial expressions.

Finally, the raw rgb signal per frame is computed with the mean value of each of this three colours, as it is stated in the previous section. Also, those three equations in Section 3.3.1 can be summed up like $c_i = [c_1, c_2, ..., c_n]$, where $c_i = r, g, b$ and n is the frame number.

Figure 3.4: Landmarks and ROIs in different frames. Lilac landmarks are the ones that have been used for this subject.

3.3.3 Illumination Correction

Once motion interference is corrected, illumination interferences have to be addressed. Therefore, the variations on the obtained signal are caracterised by two additive factors: the pulse variations and the illumination changes [25].

$$c_i = c_{hr_i} + l \tag{3.4}$$

c_{hr_i} denotes the colour value caused by the blood pressure and l represents the variations caused by illumination oscillations.

Despite the value of l cannot be computed directly, in the majority of human-computer interaction (HCI) situations (*e.g.* watching a movie), the background is also affected by the same lighting changes that affect the face, and the MAHNOB-HCI videos are not an exception. In fact, since the background is very dark, those illumination variations can be detected very straightforward.

Let us denote the background mean colour variations as $c_{bg_i} = r_{bg}, g_{bg}, b_{bg}$. If they are computed at every frame, we can say that they are proportional to l [25].

$$l \approx h c_{bg_i} \tag{3.5}$$

However, in our case, the background illumination is just affected by the illumination that the computer screen radiates, since there is no other light source, so $l \equiv c_{bg_i}$.

In conclusion, taking equation (3.4),

$$c_{hr_i} \equiv c_i - c_{bg_i} \tag{3.6}$$

In practice, the mean colour of the background is extracted with ease, given that the colours of the background are very similar for every subject, since the recordings were performed in the same room under the same light conditions, with the screen as the only light source. That is why we can set a certain interval of colour values that correspond to the background. However, calculating the *rgb* values for each pixel of the image, in order to see if they fall into the background colour interval is very costly. That is why the image is divided into superpixels [28], where the mean *rgb* is computed in each one of them.

Once we know which superpixels belong to the background we also know the ones that belong to the foreground; therefore a foreground extraction is performed (Figure 3.6). Now, it is easy to compute the mean colour value of *rgb* of the background in order to apply Equation (3.6).

Figure 3.5: Original frame

Figure 3.6: Foreground extraction

3.3.4 Temporal Filtering

After extracting the colour signal, and subtracting the oscillations created by illumination changes, ICA is performed in order to find the raw pulse, but sadly, this is just fulfilled in an ideal situation. However, since the videos are quite noisy, further processing must be carried out.

So, the final step to obtain our noise-less signal is to perform a temporal filtering, which, in this case, consists of two steps.

A. Detrending

This first filter is a detrending method that is based on a smoothing of the signal [29], and it is mostly applied for heart rate variability analysis. It is very simple to use since the frequency response can be adjusted to different signals by a single parameter: the smoothing factor λ.

In this project, we provide an exhaustive study of the behaviour of the signal depending on the smoothing parameter, since we need to select the optimal one for HR extraction using video signals. As it is stated in [29], the higher the λ, the more restrictive is the filter.

Since the dataset is quite diverse, an optimal λ must be found. We conducted an experiment with several λ and then, computed the error behaviour for each one of its

values along the videos. The chosen smoothing coefficients were:

$$\lambda_i = 10^{-3} \cdot [0.125, 0.5, 0.67, 1, 1.25, 1.67, 2.5, 5, 10]$$

where coefficients larger than $10 \cdot 10^{-3}$ filtered excessively the signal, that is why $\lambda = 10 \cdot 10^{-3}$ is considered as maximum detrending. Meanwhile, coefficients smaller or equal than $0.125 \cdot 10^{-3}$ did not affect the original signal, that is why when we apply $\lambda = 0.125 \cdot 10^{-3}$ we consider it as a no-detrending approach. Those value delimit the range of the smoothing factor and, in order to explain the in-between values it is more intuitive to work with the inverse of λ:

$$\frac{1}{\lambda_i} = [8000, 2000, 1500, 1000, 800, 600, 400, 200, 100]$$

Those nine values were empirically chosen once we observed that the signal changed more significantly when using smaller values of $\frac{1}{\lambda}$ ($200 \leq \frac{1}{\lambda} \geq 1000$), so the step there is smaller (200). Furthermore, the bigger the values, the less significant the changes they produced with a small step, that is why we chose a 500 step from 1000 to 2000.

After choosing all the possible λ_i, the data is separated in training set (80%) and in test set (20%). Such separation will be maintained until the end of the project in order to have a reliable test set, and the study of the λ factor will be conducted just with the training set.

The results are displayed in Figure 3.7, where the mean and the standard deviation of each one of the nine coefficients are also shown. Even though they are quite similar, $\lambda_4 = 10^{-3}$ will be chosen, since it has the lowest mean error and does not have

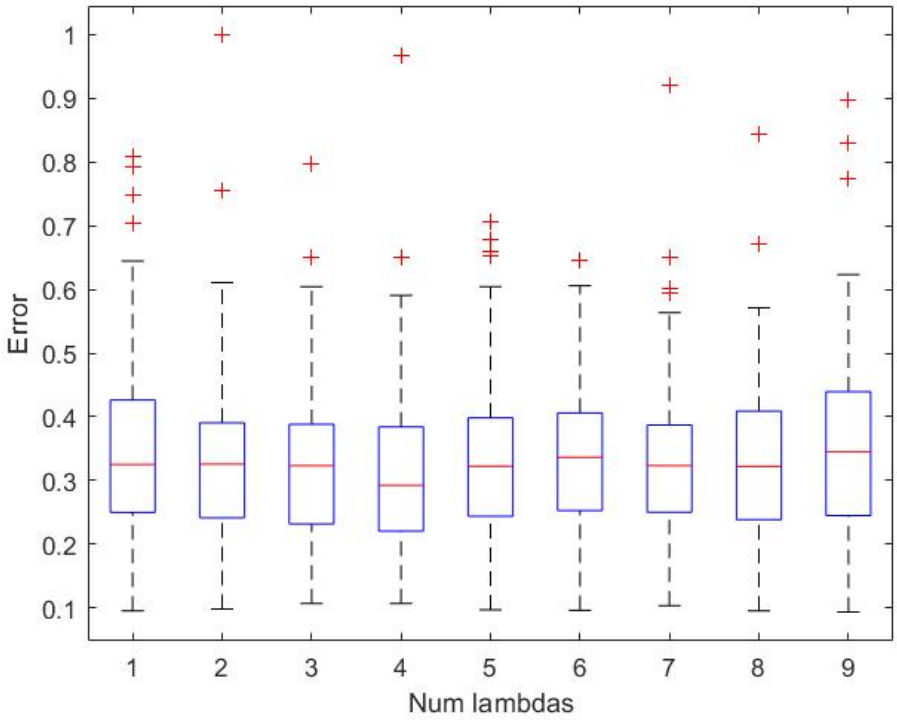


Figure 3.7: Boxplot of the error behaviour of the estimated signal, depending on each of the nine possible smoothing coefficients.

many outliers. How we computed the corresponding error is explained in Section 3.5.

B. Band-pass filter

Once the detrending is performed, the signal still has oscillations that do not correspond to the cardiac pulse. That is why frequencies are reduced to the range of interest, which is from 42 beat-per-minute (bpm) to 240 bpm. In frequency, it corresponds to a band-pass filter between [0.7 4] Hz.

3.4 Calculate the HR

Once the colour variations are extracted from the ROIs and the illumination changes are rectified, ICA is performed. Furthermore, the signal is detrended and finally filtered. This means we now have the "cleanest" possible data, so the HR can be computed from the obtained pulse signal.

Since the HR varies over time, the first intuition might be to calculate the time difference between each pair of peaks (each of the peaks represents a heart beat). This is the inter beat interval (IBI), but since the signal is not perfect, this technique might introduce error. Then, instead of focusing in just two peaks at a time, a sliding window will be used, and in our case, we used a trial and error process, where we experimented with different window sizes: 10 seconds, 7 seconds and 5 seconds respectively. The one that gave the best results was $wind_s = 7$ seconds. Now, for each step of the window, the following formula is applied:

$$HR_{bpm} = 60 \cdot fr \frac{\#peaks - 1}{loc(peak_n) - loc(peak_1)} \tag{3.7}$$

where $loc(peak_1)$ is the location on the x axis of first peak in the window and $loc(peak_n)$ is the location of the last. Also, in order to transform the result into beats-per-minute (bpm), we multiply by the frame rate $fr = 61$ and by 60 (seconds). And even though it follows the same idea as the IBI, it is much more robust since it is not that susceptible to erroneous peaks.

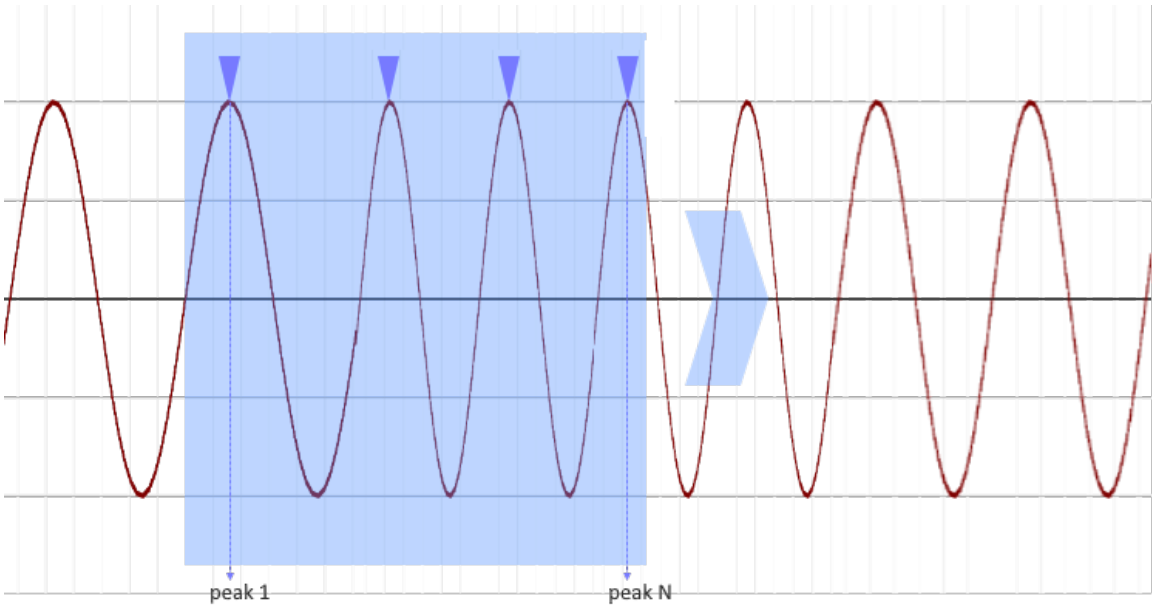


Figure 3.8: Illustration of the sliding window.

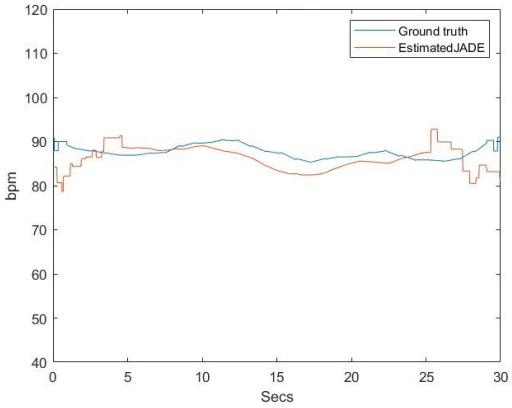
3.5 Error computation

Once the final HR curve is found, its differences from the ground truth have to be calculated. So, the first step is to transform the ground truth into the desired form, that means, transforming the ECG signal into a HR signal. To do so, there are two steps:

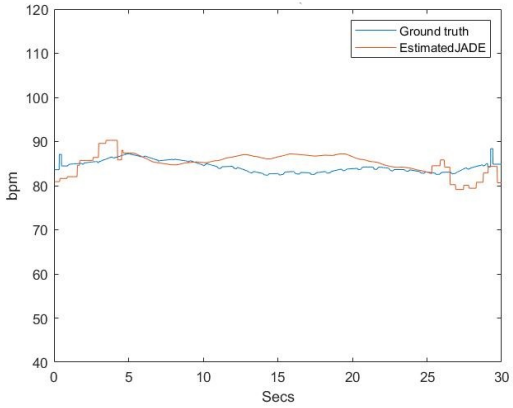
- Filter the ECG signal with a low-pass filter at 0.5 Hz to remove minor frequency changes [3].
- Following the same procedure as in Section 3.4, for each time window (of seven seconds, in this case), equation (3.7) is going to be applied.

Finally, the error between the estimated signal and its ground truth, is simply calculated by computing the area in between those curves. Also, in order to make the area

value more understandable, all errors have been normalised between 0 and 1. Figure 3.9 and Figure 3.10 indicate the behaviour of the ground truth and estimated HR along 30 seconds, and it shows that the most problematic segments are the beginning and the end parts.



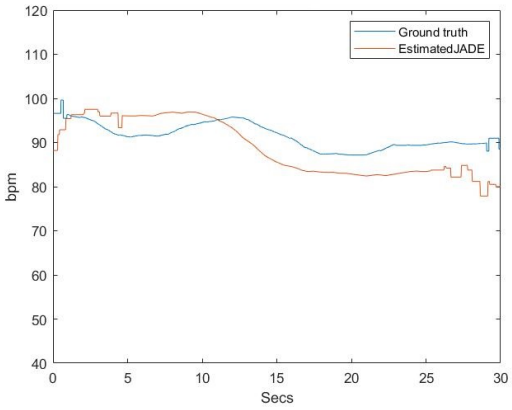
(a) Experiment 142, $\lambda = 10^{-3}$, $err_n = 0.19$



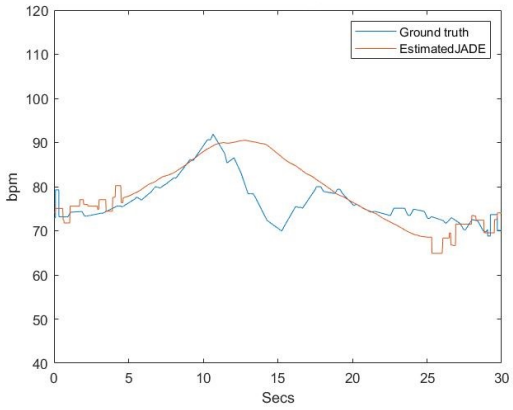
(b) Experiment 4, $\lambda = 10^{-3} \cdot 1.25$, $err_n = 0.15$

,

Figure 3.9: Estimated HR (orange) vs ground truth (blue)



(a) Experiment 160, $\lambda = 10^{-3} \cdot 0.125$, $err_n = 0.29$



(b) Experiment 284, $\lambda = 10^{-3} \cdot 0.125$, $err_n = 0.27$

Figure 3.10: Estimated HR (orange) vs ground truth (blue)

Chapter 4

TRAINING

In order to train a successful model that can predict the valence and arousal states we have to follow the steps that are presented in this chapter, such as data augmentation and windowing.

4.1 Data augmentation

The stimuli videos used to create the MAHNOB dataset belong to a series of film fragments specially chosen in order to induce certain emotions. Despite that, each of these fragments can also be subdivided since they consist of a succession of actions that might boost certain emotions momentarily (e.g. a shooting, a joke, a sudden scream...). That is why we believe the global valence and arousal annotations performed by the users in the MAHNOB database could be improved to provide more information, since in each of these fragments it is very unlikely that the user finds itself experiencing just one emotional state.

It is very important to note that when we talk about "stimuli" we mean the clip

that was shown to each subject. But when we say "video" we are referring to the actual clip of the subject's facing the screen, while he/she is watching the stimuli fragment. Furthermore, each subject is shown twenty stimuli and each one of the reactions is recorded in a different video.

In order to obtain results that are closer to the emotional reality of the subjects, the valence and arousal annotations have to be more specific. So, for each one of the videos, we started a data augmentation task that consisted on two steps.

- **Division of the video:** It consists on watching the stimuli clip at the same time as the corresponding video and dividing it into "emotionally different" segments. This is done by paying attention to the user's expressions and also by analysing the different stages of the stimuli clip itself. Nevertheless, normally every video is divided in the same segments for every user, as the majority of the stimuli have differentiated stages.

- **Annotate:** A new valence and a new arousal value are assigned for each of the subsegments, and to do so, the global valence and arousal that the subject originally provided are greatly taken into account, as they are our only connection to the user's opinion. Hence, the new values I chose depend on the physical reaction of the user, but they are set around the numbers they provided.

All the annotations have been performed manually in Excel sheets, with one sheet for each stimuli; e.g, Figure 4.1 and Figure 4.2 display the time segments an corresponding tags, for stimuli 58.avi and 107.avi, respectively. The numbers in *Init* and *End* columns are expressed in seconds and represent the starting and finishing second of the segment for each video indicated in *Vid*. Also, V and A stand for valence and arousal respectively.

Once all the annotations were finished, we went from 166 global valence and arousal tags, respectively, to 479 valence tags and 479 arousal tags (after removing subject's 9 annotations).

Vid	Init1	End1	Init2	End2	Init3	Fin3	Init4	Fin4	V1	A1	V2	A2	V3	A3	V4	A4
6	0	13	14	25	26	54	55	58	7	5	9	8	7	5	8	8
156	0	13	14	25	26	54	55	58	4	2	5	2	5	2	6	3
292	0	13	14	25	26	54	55	58	5	1	5	1	4	1	5	1
418	0	13	14	25	26	54	55	58	8	4	7	4	8	5	9	5
554	0	13	14	25	26	54	55	58	5	1	5	1	5	1	6	1
688	0	13	14	25	26	54	55	58	5	3	5	3	6	4	6	4
946	0	13	14	25	26	54	55	58	6	3	5	3	6	3	8	3
3782	0	13	14	25	26	54	55	58	6	2	7	2	6	2	7	2

Figure 4.1: Subsegments and valence-arousal tags for stimuli 58.avi

Vid	Init1	End1	Init2	End2	V1	A1	V2	A2
22	0	12	13	34	4	7	4	8
170	0	12	13	34	4	4	5	5
294	0	12	13	34	4	3	4	5
428	0	12	13	34	2	6	2	7
538	0	12	13	34	2	6	1	8
672	0	12	13	34	1	6	1	9
940	0	12	13	34	2	5	2	9
3798	0	12	13	34	3	4	3	6

Figure 4.2: Subsegments and valence-arousal tags for stimuli 107.avi

4.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning model that can solve classification, regression and outliers detection tasks, among others. Furthermore, it solves linear and also non-linear classification problems, depending on the used kernel. More specifically, the SVM constructs one or several hyperplanes in the

corresponding dimensional space. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalisation error of the classifier [30].

4.2.1 SVM in emotion classification

Many different machine learning algorithms have been successfully used in the emotion classification task: k-Nearest Neighbours (k-NN), Bayesian Network and Regression Tree (RT) and SVM, for instance. Despite that, in this project we have focused on the use of SVM as it is the most popular in the matter, since, Rani *et al.* did a comparative study of several classification methods [31] and showed that SVM was the most accurate, outperforming the Bayesian Network and k-NN. Furthermore, despite the rise of the Deep Neural Networks (DNN), the SVM has one clear advantage: it needs far less data in order to give a good accuracy, which is a decisive factor since in the field of emotion detection using physiological signals there is a lack of precise emotional ground truth.

Wiem et al. [3] conducted a study to compare the performance of several kernels in the emotion classification task, using an ECG. They compared Linear, Polynomial, Sigmoid and Gaussian kernels in a three class classification of valence and arousal. The best scoring kernels where the Linear for valence and the Polynomial for arousal. But since their differences were minimum, in this project we have used the Linear kernel due to its simplicity and good performance. Particularly for the implementation, we used the LibSVM library under the MATLAB platform [32], which uses the

One vs One approach.

4.3 Preprocess the signal

Once the pulse signal has been extracted from the colour and grayscale information, and has been transformed into HR units, there are still some steps remaining before being able to train the SVM model properly.

4.3.1 Windowing

Looking at Figure 4.1 and Figure 4.2, one might notice that, as expected, the segments do not have the same duration. This is a problem when it comes to training the SVM model, since it needs a feature matrix of size $M \times W$ and a label vector $\mathbf{l}$ of size $M \times 1$.

The solution is to window the signal with a window of size w_s, so, for example, taking a feature of N samples such that:

$$\mathbf{f} = [f_1,\ f_2, ..., f_N] \tag{4.1}$$

where $\mathbf{f}$ has a size of size $1 \times N$, the following expression is going to be applied:

$$M' = \left\lfloor \frac{N}{w_s} \right\rfloor \tag{4.2}$$

where M' represents the number of new feature vectors of size w_s after the windowing. So, the new feature matrix $\mathbf{f}$' has a size of $M' \times w_s$.

$$\mathbf{f}' = \begin{bmatrix} f_{11} & \cdots & f_{1w} \\ \vdots & & \vdots \\ f_{M'1} & \cdots & f_{M'w} \end{bmatrix} \tag{4.3}$$

Furthermore, since the number of vectors is increased, the valence and arousal annotations have to be taken into account and have to match the new dimensions. So the new label vector $\mathbf{l}$' has a size of $M' \times 1$.

Moreover, even though resampling the signal into the desired size might seem like a good alternative to windowing, this is not advised, as the training has to be carried out with a real-time consistence, *i.e.*, if 10 seconds of a signal are resampled to the same number of samples as a 5 second signal, it will be very difficult to generalize, and in consequence, to learn, for the SVM algorithm.

4.3.2 Grouping the valence and arousal

In previous works the classification task is reduced to a two or three class classification [3][33][2], so in order to be able to compare, we will follow a three class division (Table. 4.1).

Categorisation		**Rating**
Valence	**Arousal**	**"r" value**
Unpleasant	Calm	$1 \leq r \leq 3$
Neutral	Medium	$4 \leq r \leq 6$
Pleasant	Excited	$7 \leq r \leq 9$

Table 4.1: Three defined classes given the valence-arousal annotations

4.3.3 Normalising the data

Not all humans have the same cardiac pulse range and the neutral stage can be manifested various ranges of bpm. This difference of pulse range between humans hinders the valence-arousal prediction, e.g., 90 bpm can mean excitement for one subject, but meanwhile, it might be the rest sage pulse for another one. So, in order to reduce the difference between users that solely depends on each ones features and is independent from the stimuli, we are going to normalise the HR on each segment by mapping all the values in the feature, to an interval between 0 and 1.

Even though this step is totally optional, we have observed that has improved our results.

Chapter 5

RESULTS

In this chapter we will show the obtained results regarding the valence and arousal predictions and also compare it with other studies. In particular, the classification accuracy results are shown in Section 5.1, and the overall discussion can be found in Section 5.2.

5.1 Classification accuracy

This Section presents the results of numerical accuracy for valence and arousal given a 3-class classification and a 9-class classification. The obtained percentages correspond to the test set; bear in mind that we split the dataset into 80% for training and 20% for test.

5.1.1 3-class Classification

As seen in Table 4.1 in the previous Section, both valence-arousal tags can be distributed into 3 different classes each; in particular, valence can be classified as unpleasant / neutral / pleasant, while arousal can be associated to calm / medium /

excited tags. Moreover, as we discussed in Section 4.3.3, we also want to perform a small ablation study depending on the number of ICA channels and the potential effect of normalising data *a priori*. Therefore, 4 tables can be seen in this Subsection:

- Table 5.1 shows the valence-arousal results when using 3 ICA channels plus normalisation.

- Table 5.2 shows the valence-arousal results when using 4 ICA channels plus normalisation.

- Table 5.3 displays the valence-arousal results when using 3 ICA channels without normalisation.

- Table 5.4 displays the valence-arousal results when using 4 ICA channels without normalisation.

In all the above-mentioned Tables, the classification accuracy of valence-arousal is displayed together with the corresponding window-size (from 2 seconds to 20).

Window	Valence	Arousal
20s	40.23%	36.56%
18s	42.6%	39.44%
16s	50.09%	40.12%
14s	51.29%	40.83%
12s	53.6%	42.12%
10s	61.60%	48.00%
8s	58.28%	46.63%
6s	68.4%	39.28%
4s	67.7%	36.45%
2s	49%	18.28%

Table 5.1: Valence-Arousal accuracy using a 3-channel ICA with norm. data (3-class).

Window	Valence	Arousal
20s	41.72%	38.87%
18s	45.37%	40.55%
16s	49.25%	39.08%
14s	56.20%	36.95%
12s	51.36%	40.52%
10s	51.20%	46.40%
8s	60.74%	41.1%
6s	69.54%	34.65%
4s	65.6%	30.23%
2s	49.07%	18.42%

Table 5.2: Valence-Arousal accuracy using a 4-channel ICA with norm. data (3-class).

Window	Valence	Arousal
20s	40.19%	38.44%
18s	43.29%	39.19%
16s	42.05%	35.95%
14s	41.24%	36.71%
12s	46.40%	39.17%
10s	44.9%	44.8%
8s	51.53%	42.58%
6s	48.24%	33.8%
4s	56.48%	35.54%
2s	45.33%	28.68%

Table 5.3: Valence-Arousal accuracy using a 3-channel ICA without norm. data (3-class).

Window	Valence	Arousal
20s	41.19%	38.88%
18s	46.14%	37.07%
16s	44.38%	37.88%
14s	50.18%	30.39%
12s	49.12%	45.33%
10s	49.6%	45.6%
8s	57.66%	40.65%
6s	50.43%	45.31%
4s	51.45%	37.55%
2s	49.97%	36.18%

Table 5.4: Valence-Arousal accuracy using a 4-channel ICA without norm. data (3-class).

5.1.2 9-class Classification

In order to compare our results with other State-of-the-Art methods, the 3-class classifier was trained, which is the one being used in the related literature. However, we also wanted to display the accuracy obtained by building a 9-class SVM, since both valence and arousal were originally tagged within a 1 to 9 interval. For valence it goes from negative to positive (1-9) where 5 means neutral. Similarly, for arousal it goes from not excited to maximum excitement (1-9)

This section follows the same structure as the previous one: first we display the results tables obtained from normalising the data (Table 5.5 and Table 5.6), and afterwards, the final accuracy calculated from non-normalised data (Table 5.7 and

Table 5.8).

Window	**Valence**	**Arousal**
20s	11.27%	11.27%
18s	16.23%	12.11%
16s	17.98%	11.80%
14s	18.55%	6.18%
12s	19.63%	10.43%
10s	16.00%	11.20%
8s	19.63%	10.43%
6s	18.86%	11.40%
4s	16.39%	9.23%
2s	13.60%	5.52%

Table 5.5: Valence-Arousal accuracy using a 3-channel ICA with norm. data (9 class).

Window	Valence	Arousal
20s	10.63%	10.64%
18s	16.67%	7.40%
16s	15.05%	8.55%
14s	13.92%	8.86%
12s	20.61%	13.40%
10s	19.20%	12.00%
8s	15.95%	9.20%
6s	12.5%	8.44%
4s	13.37%	8.02%
2s	13.81%	5.52%

Table 5.6: Valence-Arousal accuracy using a 4-channel ICA with norm. data (9 class).

Window	Valence	Arousal
20s	14.68%	11.27%
18s	16.56%	11.66
16s	16.25%	11.70%
14s	15.01%	11.39%
12s	25.07%	11.34
10s	18.40%	9.06%
8s	13.00%	8.40%
6s	14.92%	10.08%
4s	13.99%	11.04%
2s	13.00%	8.40%

Table 5.7: Valence-Arousal accuracy using a 3-channel ICA without norm. data (9 class).

Window	Valence	Arousal
20s	12.76%	11.27%
18s	14.81%	9.26%
16s	15.92%	8.59%
14s	17.72%	7.6%
12s	16.49%	15.43%
10s	16.20%	15.20%
8s	15.78%	8.15%
6s	16.30%	10.52%
4s	15.67%	9.28%
2s	12.89%	8.41%

Table 5.8: Valence-Arousal accuracy using a 4-channel ICA without norm. data (9 class).

5.2 Discussion

On the one hand, results show that in order to estimate the valence and arousal for each subject, the 4-channel ICA gives better results, in general, than the 3-channel ICA; this matches our intuition which states that, the more channels, the better ICA separates each source. Nevertheless, since the improvements are not very substantial we can say that the contributions of the gray channel do not have a huge impact on the final result.

On the other hand, the effect of normalisation has to be studied carefully. When attempting to classify within 3 classes, the normalisation step provides us with meaningful insights that produce better results. Since not all humans have the same

average heart rate, by normalising, we make sure to take only into account the heart rate differences with respect to a baseline value. On the contrary, when training the 9-class model (1-9 valence, 1-9 arousal), this normalisation step does not ensure better percentages, since it is really difficult to distinguish the differences between neighbouring rating values. Take for instance a real-life example: while it is pretty easy to detect the difference between joy-fear expressions, it is quite tough to distinguish close levels of valence/arousal; for example, if we compare 2 frames of the same person with an arousal level of 2 and 3 respectively, we probably would not be able to spot any changes. For this reason, adding this normalisation step does not add any value to our classifier (the overall results drop a little bit), since neighbouring classes are really close in the feature-space.

5.2.1 Impact of the window

Even though our intuition was that the bigger windows would produce the best results, this has not been fulfilled, and the explanation is quite straightforward: the smaller the window, the higher the number of features (4.2). Thus, the consequent increase in the training data compensates the shrinking size of the window.

Furthermore, in every results table, the percentages for the biggest and the smallest window are the lowest ones. The explanation follows the same reasoning as before, for a window size of 20 seconds, the number of features is $\sim$ 250, which leads to an underfitted model. Whereas for a window of size 2 seconds, the opposite effect happens, since there are $\sim$ 3000 features. Furthermore, apart from the overfitting, a window of 2 seconds does not have enough information, so the SVM classifier cannot generalise.

In conclusion, when choosing a window length, we have to take into account the compromise between the number of features and the significance of each one of those.

5.3 Accuracy comparisons

Table 5.9 shows the comparison between our results and other related literature. We see that this project outperforms the others in terms of valence prediction. Nevertheless, the accuracy in the arousal metric is not as high as in the other projects, which might be due to the fact that we are not using as many videos and as many footage time; as a matter of fact Wiem *et al.* [3] used around 350 more videos than us. Nevertheless, our videos had a fine-grained tagging that led to a solid performance.

Classification Model	**Valence**	**Arousal**
RNN (Signal from video), Chacon *et al.*	57.00%	58.00%
RNN (Pulse), Chacon *et al.*	59.60%	58.50%
ECG, Wiem *et al.*	52.12%	51.4%
SVM (HR)	69.54%	48.00%

Table 5.9: Valence-Arousal accuracy comparisons [2] [3]

Chapter 6

CONCLUSIONS

In this project we have been able to predict the valence and the arousal states of a group of subjects given a video signal of their faces. First, the mean r, g, b values from each frame were extracted in order to compute a 3-channel ICA. Moreover, we added a fourth channel corresponding to the grayscale video in order to improve the BSS performance. It was shown that, even though the gray channel, combined with the r, g, b, can increase the prediction accuracy, it does not guarantee it.

Furthermore, we built a SVM model that successfully classifies valence and arousal into three classes: unpleasant, neutral and pleasant for valence and calm, medium and excited for arousal, where the best results have been obtained in the valence metric. Nevertheless, and despite the success in the 3-class classification, we also performed a 9-class classification, where the labels lay in the 1 to 9 possible values of valence or arousal that the user can experience in each video segment; and the results were not as satisfactory since the obtained accuracy is quite small. Hence, even though the valence accuracy is above the random classification percentage, the values are still too low to use in a real situation.

In conclusion, we have seen that accurate and fine-grained annotations can substantially contribute to the emotion estimation task. Moreover, we show that a meticulous processing of the video signal cannot be skipped and also ensures good prediction of the valence and arousal in humans. This project is yet another prove that HR estimation via regular camera is possible and that tasks such as emotion recognition can be carried out with this method, obtaining a State-of-the-Art accuracy.

6.1 Future work

Future research on the implication of the duration of the stimulus on the arousal prediction might confirm the hypothesis we have stated in this project (Section 5.3). That is, since previous works that used more video footage with less annotations obtained higher accuracy in that matter, the arousal will be better estimated with bigger windows.

Also, increasing the annotations dataset would increase the range of experiments we can perform regarding the search of the optimal window size. Once this window size was known, we could implement a real-time valence-arousal detector with an answer delay equal to the window size.

Appendix A

Physiological signals

The physiological signals are:

1. **Electroencephalogram (EEG):** Neurophysiological exploration, based on the registration of the bioelectric activity.

2. **Electrocardiogram (ECG):** Using electrodes attached to the skin it measures the voltage versus the time of the electrical activity of the heart.

3. **Heart rate variability (HRV):** Variation in the time intervals between each beat.

4. **Galvanic skin response (GSR):** Measure of the continuous variations in the electrical characteristics of the skin, i.e. for instance the conductance, caused by the variation of the human body sweating.

5. **Electromyogram (EMG):** Registers graphically the electrical activity of the muscles.

6. **Skin temperature (SKT):** Measures the temperature of the skin.

7. **Blood volume pulse (BVP):** Signal originated when a volume of blood moves across a tissue.

8. **Respiratory volume (RESP):** It refers to the volume of gas in the lungs at a given time during the respiratory cycle.

Bibliography

[1] N. Ravaja, "Contributions of psychophysiology to media research: Review and recommendations," *Media Psychology*, vol. 6, no. 2, pp. 193–235, 2004.

[2] L. A. B. Chacon, A. Fedoskin, E. Shcheglakova, S. Neamsup, and A. Rashed, "Emotion analysis using heart rate data," in *International Conference on Database and Expert Systems Applications*, pp. 147–154, Springer, 2019.

[3] M. B. H. Wiem and Z. Lachiri, "Emotion classification in arousal valence model using mahnob-hci database," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 8, no. 3, 2017.

[4] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[5] P. Suja, S. Tripathi, *et al.*, "Real-time emotion recognition from facial images using raspberry pi ii," in *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 666–670, IEEE, 2016.

[6] N. Chanthaphan, K. Uchimura, T. Satonaka, and T. Makioka, "Facial emotion recognition based on facial motion stream generated by kinect," in *2015 11th*

International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 117–124, IEEE, 2015.

[7] C. Turan, K.-M. Lam, and X. He, "Facial expression recognition with emotion-based feature fusion," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–6, IEEE, 2015.

[8] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[9] F. Chenchah and Z. Lachiri, "Acoustic emotion recognition using linear and non-linear cepstral coefficients," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 11, pp. 135–138, 2015.

[10] N. Dael, M. Mortillaro, and K. R. Scherer, "Emotion expression in body action and posture.," *Emotion*, vol. 12, no. 5, p. 1085, 2012.

[11] S. Wioleta, "Using physiological signals for emotion recognition," in *2013 6th International Conference on Human System Interactions (HSI)*, pp. 556–561, IEEE, 2013.

[12] K. Smitha and A. Vinod, "Hardware efficient fpga implementation of emotion recognizer for autistic children," in *2013 IEEE International Conference on Electronics, Computing and Communication Technologies*, pp. 1–4, IEEE, 2013.

[13] R. A. L. Koelstra, *Affective and Implicit Tagging using Facial Expressions and Electroencephalography.* PhD thesis, Queen Mary University of London, 2012.

[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physio-

logical signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[15] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 211–223, 2011.

[16] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2018.

[17] A. Fernández-Caballero, A. Martínez-Rodrigo, J. M. Pastor, J. C. Castillo, E. Lozano-Monasor, M. T. López, R. Zangróniz, J. M. Latorre, and A. Fernández-Sotos, "Smart environment architecture for emotion detection and regulation," *Journal of biomedical informatics*, vol. 64, pp. 55–73, 2016.

[18] R. E. Kleck, R. C. Vaughan, J. Cartwright-Smith, K. B. Vaughan, C. Z. Colby, and J. T. Lanzetta, "Effects of being observed on expressive, subjective, and physiological responses to painful stimuli.," *Journal of Personality and Social Psychology*, vol. 34, no. 6, p. 1211, 1976.

[19] K. R. Scherer and P. Ekman, *Approaches to emotion*. Psychology Press, 2014.

[20] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[21] R. Plutchik, *The emotions*. University Press of America, 1991.

[22] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[23] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[24] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[25] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4264–4271, 2014.

[26] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," in *IEE proceedings F (radar and signal processing)*, vol. 140, pp. 362–370, IET, 1993.

[27] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 532–539, 2013.

[28] K. S. A. L. P. F. S. S. Radhakrishna Achanta, Appu Shaji, "Lslic superpixels compared to state-of-the-art superpixel methods.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274–2282, 2012.

[29] M. P. Tarvainen, P. O. Ranta-Aho, and P. A. Karjalainen, "An advanced detrending method with application to hrv analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 172–175, 2002.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[31] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis and Applications*, vol. 9, no. 1, pp. 58–69, 2006.

[32] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[33] A. Subramaniam and R. Kanjirappuzha, "Spectral reflectance based heart rate measurement from facial video," 01 2019.