UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Análisis de la actividad ocular y la dilatación pupilar durante estimulación emocional en una muestra de sujetos sanos

Trabajo Fin de Máster

Máster Universitario en Ingeniería Biomédica

AUTOR/A: Profili , Alessandro

Tutor/a: Naranjo Ornedo, Valeriana

Cotutor/a externo: MAZA PINO, ANNY MICHELLE

Director/a Experimental: LLORENS RODRIGUEZ, ROBERTO

CURSO ACADÉMICO: 2021/2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIERÍA INDUSTRIAL VALENCIA

BIOMEDICAL ENGINEERING MASTER THESIS

ANALYSIS OF VISUAL ACTIVITY AND PUPIL DILATION DURING EMOTIONAL STIMULATION IN A SAMPLE OF HEALTHY SUBJECTS

AUTHOR: Alessandro Profili

SUPERVISOR: Valeriana Naranjo Ornedo

SUPERVISOR: Anny Michelle Maza Pino

Roberto Llorens Rodríguez

Academic year: 2021-2022

RESUMEN

El reconocimiento de emociones es un campo de investigación de gran interés para áreas como la sanidad y la psicología. En las últimas décadas se han desarrollado muchos modelos y algoritmos de reconocimiento de emociones para identificar y comprender las emociones humanas. El reconocimiento de las emociones puede intentarse mediante el análisis de las respuestas fisiológicas a determinados estímulos emocionales. Entre otras respuestas fisiológicas, la actividad oculomotora y la dilatación de la pupila han sido investigadas repetidamente en la literatura, debido a su bajo coste y naturaleza no invasiva. Este trabajo investiga, en una muestra de 15 participantes, las diferencias en estas respuestas ante estímulos audiovisuales emocionales personalizados, consistentes en vídeos pregrabados en los que aparecen conocidos o desconocidos destacando características positivas y planes de futuro de las personas estudiadas o de otros individuos, respectivamente. Para ello, se han utilizado algoritmos de aprendizaje automático, en primer lugar, para diferenciar entre periodos de estimulación y no estimulación; en segundo lugar, para diferenciar entre estímulos protagonizados por conocidos o desconocidos; y, por último, para diferenciar entre estímulos de placer y excitación específicos autodeclarados. Nuestros resultados mostraron una precisión superior al 90% a la hora de diferenciar los periodos de estimulación y no estimulación, y una precisión en torno al 80% a la hora de diferenciar los vídeos de conocidos y extraños. La precisión a la hora de identificar estímulos con características específicas fue superior al 65% y comparable a la reportada por estudios anteriores. A la luz de estos resultados, se necesitan estudios adicionales que incluyan más participantes y una gama más amplia de emociones.

Palabras Clave: emociones, reconocimiento de emociones, machine learning, clasificación, simulación audiovisual, seguimiento ocular, eventos oculomotores, dilatación de la pupila

RESUM

El reconeixement d'emocions és un camp d'investigació de gran interés per a àrees com la sanitat i la psicologia. En les últimes dècades s'han desenrotllat molts models i algoritmes de reconeixement d'emocions per a identificar i comprendre les emocions humanes. El reconeixement de les emocions pot intentar-se per mitjà de l'anàlisi de les respostes fisiològiques a determinats estímuls emocionals. Entre altres respostes fisiològiques, l'activitat oculomotora i la dilatació de la pupil·la han sigut investigades repetidament en la literatura, a causa del seu baix cost i naturalesa no invasiva. Este treball investiga, en una mostra de 15 participants, les diferències en estes respostes davant d'estímuls audiovisuals emocionals personalitzats, consistents en vídeos pregravats en els que apareixen coneguts o desconeguts destacant característiques positives i plans de futur de les persones estudiades o d'altres individus, respectivament. Per a això, s'han utilitzat algoritmes d'aprenentatge automàtic, en primer lloc, per a diferenciar entre períodes d'estimulació i no estimulació; en segon lloc, per a diferenciar entre estímuls protagonitzats per coneguts o desconeguts; i, finalment, per a diferenciar entre estímuls de plaer i excitació específics autodeclarats. Els nostres resultats van mostrar una precisió superior al 90% a l'hora de diferenciar els períodes d'estimulació i no estimulació, i una precisió entorn del 80% a l'hora de diferenciar els vídeos de coneguts i estranys. La precisió a l'hora d'identificar estímuls amb característiques específiques va ser superior al 65% i comparable a la reportada per estudis anteriors. A la llum d'estos resultats, es necessiten estudis addicionals que incloguen més participants i una gamma més àmplia d'emocions.

Paraules clau: emocions, reconeixement d’emocions, machine learning, classificació, simulació audiovisual, seguiment ocular, esdeveniments oculomotors, dilatació de la pupila

ABSTRACT

Emotion recognition is a research field of major interest to areas such as health care and psychology. Many emotion models and emotion recognition algorithms have been developed in the last decades in order to identify and understand human emotions. The recognition of emotions can be attempted through the analysis of physiological responses to specific emotional stimuli. Among other physiological responses, oculomotor activity and pupil dilation have been repeatedly investigated in the literature, due to its low cost and non-invasive nature. This work investigates, in a sample of 15 participants, differences in these responses to customized emotional audio-visual stimuli, consisting of pre-recorded videos featuring either acquaintances or strangers highlighting positive characteristics and future plans of either the persons under study or other individuals, respectively. To this aim, machine learning algorithms have been used, first, to differentiate between stimulation and non-stimulation periods; second, to differentiate between stimuli featuring acquaintances or strangers; and, finally, to differentiate between stimuli of specific self-reported pleasure and excitation. Our findings showed an accuracy higher than 90% at differentiating stimulation and non-stimulation periods, and an accuracy around 80% at differentiating videos from acquaintances and strangers. The accuracy at identifying stimuli with specific characteristics was higher than 65% and comparable to that reported by previous studies. In light of these results additional studies are needed including more participants and a wider range of emotions.

Keywords: emotions, emotion recognition, machine learning, classification, audio-visual stimulation, eye-tracking, oculomotor events, pupil dilation

INDEX

CONTENT

DISSERTATION

1. INTRODUCTION

1.1. PROJECT PRESENTATION

In the last decades, a large increase in the interest of automatic recognition of emotions has been observed. Emotion recognition has become popular among affective computing researchers since a robust emotion recognition algorithm could provide many useful applications in the fields of neuromarketing, entertainment, computer gaming, psychology, education etc. as stated by J. Z. Lim in [1]. Moreover, as stated by W. Zheng et al. in [2], many mental diseases are reported to be relevant to emotions, such as depression, autism, attention deficit hyperactivity disorder, and game addiction. Therefore, automatic recognition of emotions may contribute also to healthcare, e.g. R. Calvo et al. in [3] state that it can be used to diagnose and treat diseases such as post-traumatic stress disorder. Additionally, A. M. Kaysi et al. in [4] state that it can contribution to the treatment of depression. In order to implement algorithms of automatic emotion recognition, it is usually required the acquisition of one or more physiological signals coming from the subject such as electroencephalography, electrocardiography, eye-tracking measures etc. In the last decade, eye-tracking has become more popular in the area of cognitive science and affective information processing and it has been exploited in few attempts also in emotion recognition algorithms alone or in combination with other physiological signals [5].

The main objective of this study is to investigate physiological responses related to audio-visual emotional stimulation in healthy subjects. In particular, eye-tracking data will be used to implement an emotion recognition algorithm exploiting a machine learning approach as well as an algorithm able to discern between the various stimuli. These input stimuli will be of two different types: videos showing an acquaintance of the subject under study (relative, partner, etc.) and videos showing An unknown person to the subject. All the videos, either from referred or unknown people, will be followed by resting periods and evaluation sections in which self-reported measures will be used to assess the level of emotional involvement of the subject during the audio-visual stimulation. The stimulus will be displayed through a virtual reality headset. Once the eye-tracking data will be acquired, the implemented algorithm should be able to recognize which type of videos the subject has watched (known person, unknown person or rest) and to recognize the level of emotional involvement of the subject.

This project has been carried out in the Neurorehabilitation and Brain Research Group (NRHB) of the Instituto de Investigación e Innovación en Bioingenieria (i3B) of the Universidad Politécnica de Valencia (UPV).

1.2. EMOTIONS

Emotions play a major role in the life of every human being. People feel emotions during everyday tasks, during interpersonal communication, when making decisions, learning, or during cognitive activities. Although there is not a universal accepted definition of emotion due to the subjective

nature of it, generally it refers to a mental state that arises spontaneously rather than through conscious effort, which is often accompanied by physical and physiological changes involving organs and tissues such as brain, heart, skin, blood flow, muscle, etc [5]. The internal experience of emotion is highly personal and often confusing, so throughout the centuries researchers have tried to get a useful understanding of emotions by defining them in terms of what their adaptive function might be and by categorising them inside comprehensive models. Thus, there is more and more interest in finding a robust way to individually assess what kind of emotion a subject is perceiving while achieving a specific task. Due to the advances in technology, a large increase in the number of studies developing methods of automatic recognition of emotion has been observed.

1.2.1. CLASSIFICATION AND MODELS

In most of the attempts of categorising emotions, it is given as a basic hypothesis the fact that it is possible to define a set of separate discrete emotions. In particular, most of the emotion categorizations are based on the idea that all emotions can be built as a combination of a set of "basic emotions". In his seminal studies, Paul Ekman defined a set of basic emotions [6]. Ekman suggests that there are three meanings of the term "basic". First, it distinguishes those emotions that differ one another in important ways. For example, all negative emotions such as fear, anger, disgust, sadness and contempt, differ in their appraisal, antecedent events, probable behavioural response and physiology. The second meaning of the adjective "basic" is to indicate instead the view that emotions evolved for their adaptive value in dealing with fundamental life tasks. This meaning is based on the presumption that emotions are designed to deal with inter-organismic encounters, between people or between people and other animals, even though emotions can occur when we are not in presence of other people, e.g. we can have emotional reactions to thunder, music and auto-erotic activity among others. Finally, the term "basic" can be also used to describe elements that can combine to form more complex or compound emotions. At first, Ekman selects a set of six basic emotions which are anger, disgust, fear, happiness, sadness and surprise. Later in his career, in light of other cross-cultural studies, he theorises that other universal emotions may exist beyond these six such as amusement, awe, contempt, desire, embarrassment and others.

Another important researcher in the field of emotions who deals with the categorization and modelling of emotion is Robert Plutchik. First, he theorises the relationship between cognition, emotion and evolution [7]. He states that, from the point of view of evolution, cognition developed to predict the future more effectively, in fact the human brain has evolved as an adaptation to changing and difficult environment. Thus, if emotion is a chain of events, cognition is generally near the beginning of the chain. Emotions are not simply linear events, but rather are feedback processes whose function is to restore the individual to a state of equilibrium when unexpected or unusual events create disequilibrium. Overall, emotion is a kind of homeostatic negative feedback system in which behaviour meditates progress toward equilibrium. The block scheme of the feedback system theorized by Plutchik is shown in Figure 1.

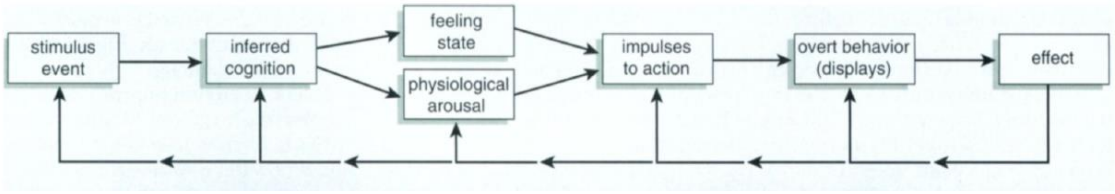

Figure 1. Feedback loops in emotion show how sensory information is evaluated and translated into action or some other outcome that normalizes the relationship between the individual and the triggering event. [2]

Another important contribution given by Plutchik to emotion theory is the so-called “wheel of emotions”. It is a circumplex model where the primary emotions can be conceptualised in a fashion analogous to a colour wheel placing similar emotions close together and opposites 180 degrees apart, like complementary colours. Other emotions are mixtures of the primary emotions, just as some colours are primary and others made by mixing the primary colours. Such "circumplex" modelling can be used as an analytical tool in understanding personality as well. The circumplex model can be extended into a third dimension, representing the intensity of emotions, so that the so-called structural model of emotions is shaped like a cone as shown in Figure 2.

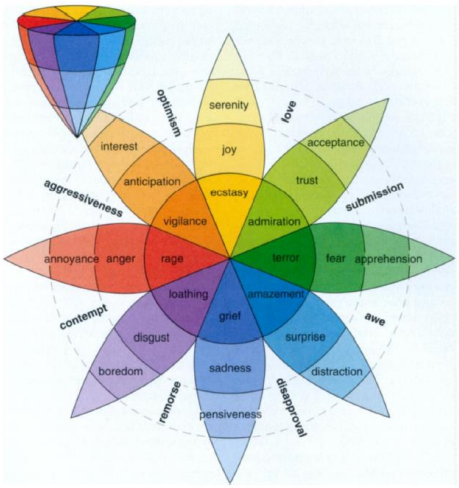

Figure 2 Author's three-dimensional circumplex model describes the relations among emotion concepts, which are analogous to the colours on a colour wheel. The cone's vertical dimension rep resents intensity, and the circle represents degrees of similarity among the emotions. The eight sectors are designed to indicate that there are eight primary emotion dimensions defined by the theory arranged as four pairs of opposites. In the exploded model the emotions in the blank spaces are the primary dyads emotions that are mixtures of two of the primary emotions. [2]

Another circumplex model of emotions was developed by James Russel [8]. As depicted in Figure 3, this model suggests that emotions can be distributed in a two-dimensional circular space containing arousal and valence dimensions. Arousal, the state of being physiologically alert, awake and attentive, represents one of the dimensions of the space (vertical axis). Valence, on the contrary, refers to the intrinsic attractiveness (positive valence) or averseness (negative valence) of an event or situation, and it is placed in the horizontal axis of the plane forming the second dimension of the space.

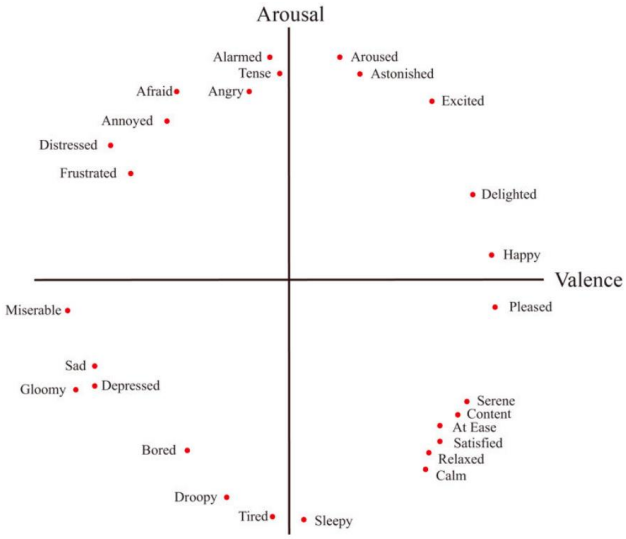

Figure 3 Circumplex model of emotions with emotions placed in the plane

Finally, it is possible to add a third dimension to this model, known as the dominance. It indexes the interactive relationship that exists between the perceiver and the perceived, with high dominance associated with the one having maximum control in the situation. Through this model, emotional states can be represented at any level of valence, arousal and dominance, or at a neutral value of one or all of these factors. Due to its versatility, the circumplex model of emotions is the most used one in recent studies concerning emotions, and it is the one that will be used in this study as well.

1.2.2. ASSESSMENT OF EMOTIONS

The self-assessment of emotions is a fundamental step that needs to be carried out in most of the studies where the inner emotional state of the subjects must be defined. For that reason, psychologists and researchers have developed a series of different scales and questionnaires for the self-assessment of internal feeling states. In this chapter two assessment techniques are going to be analysed: the Semantic Differential scale and the Self-Assessment Manikin (SAM). These scales and in particular the SAM are the most widely used scales whose main purpose is the measurement of instantaneous human reactions to standardised visual, verbal, auditory, or physical stimuli as stated by M. Murdoch in [9].

1.2.2.1. SELF-REPORTED MEASURES

The Semantic Differential scale devised by Mehrabian and Russel [10] is a widely used instrument for assessing the 3-dimensional structure of objects, events and situations. It consists of a set of 18 bipolar adjective pairs that are each rated along a 9-point scale. Factors analyses of the resulting 18 ratings generate a score on the dimensions of pleasure, arousal and dominance. Although this

method is informative, there are a number of difficulties associated with it. First, it is cumbersome to measure 18 different ratings for each stimulus presented in an experimental session. Indeed, there is a heavy investment of time and effort, and results in a relatively large database that requires statistical expertise for resolution. Moreover, the reliance on verbal rating systems makes it difficult to utilise this methodology in populations which are not linguistically sophisticated (e.g., children, aphasics, etc.).

To overcome these limitations, Lang [11] developed a picture-oriented questionnaire called the Self-Assessment Manikin (SAM) to directly assess the valence, arousal and dominance associated with the emotional response to a stimulus. The SAM consists of a nonverbal, graphic depiction of various points along each of the three major directions. As shown in Figure 4, SAM ranges from an unhappy figure to a smiling, happy figure when representing the pleasure dimension, and ranges from a relaxed, sleepy figure to an excited, wide-eyed figure for the arousal dimension. The dominance dimension represents changes in control with changes in the size of SAM: a large figure indicates maximum control in the situation. SAM is a useful instrument when determining the subjective experience of emotion associated with processing most stimuli and can be employed with a variety of subject populations. It allows to rapid assess what appear to be the fundamental dimensions in the organisation of human emotional experiences. Moreover, Lang et al. [12] also generated a database indicating that physiological signals such as cardiac signals, electrodermal response and facial displays of emotion, vary with differences in affect as indexed by the SAM dimensions of valence and arousal. In summary, SAM is a simple and efficient technique for evaluating natural emotions in three dimensions and can be used as a reliable ground truth estimate for emotion recognition systems as stated by M. Grimm in [13].

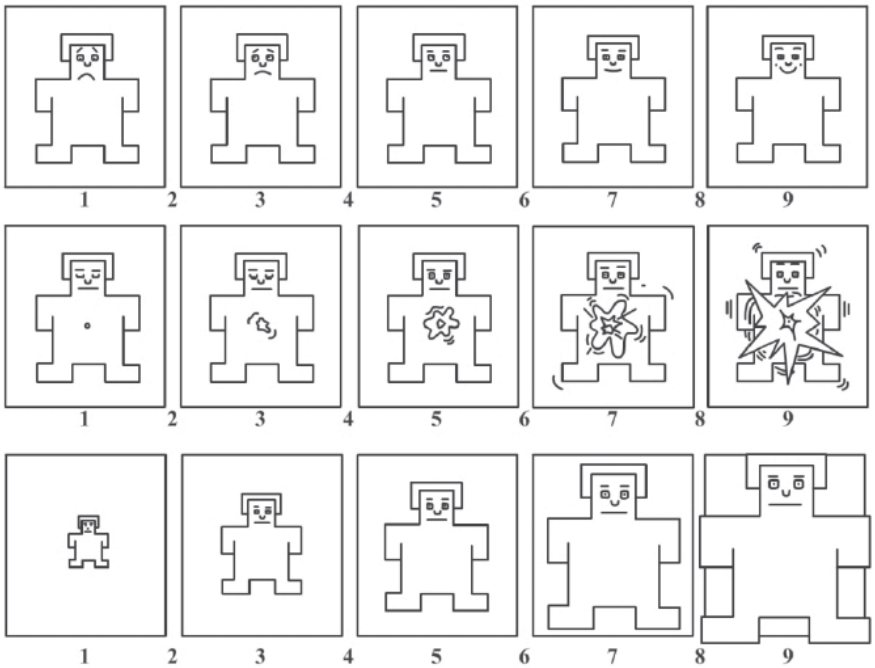

Figure 4 The Self-Assessment Maniskin (SAM) used to rate the affective dimensions of valence (top panel), arousal (middle panel) and dominance (bottom panel) [5]

In [12] Lang et al. also compare the two scales and find a near perfect agreement in ratings of pleasure and arousal for a set of pictures using the semantic differential scale and SAM, proving that SAM is a robust alternative to the semantic differential scale. They also state that pleasure and arousal dimensions are primary, and they typically account for most of the variance in emotional judgements, including when the semantic differential is the measuring instrument. For what concerns the dominance, this factor has accounted for the least variance in affective judgements and is the most variable in terms of semantic label across investigation. Since this rating is inherently relational, dominance judgments will clearly need to specify which member of the interaction is being judged: the subject or the object. This suggests that the semantic differential method may produce confusion regarding which element of the interaction is being rated for dominance. When using SAM, on the other hand, the subject can reliably rate his or her own reaction to the pictured object and so it may elicit more consistent judgements.

1.2.3. PHYSIOLOGICAL MEASURES

Emotion recognition methods can be classified into two major categories. A first category is based in the use of human physical signals such as facial expressions, speech, gesture, posture etc. The main advantage of these signals lies in the fact that they are easy to collect and that they have been studied for years. The main disadvantage is that the reliability of those can not be guaranteed, as it's relatively easy for people to control the physical signals like facial expression or speech to hide their real emotions especially during social communications. A second category can be defined that is based in the use of physiological signals, being the electroencephalogram, temperature, electrocardiogram, galvanic skin response, respiration, electromyogram and eye-tracking, the most relevant for emotion recognition, as stated by L. Shu et al. in [14]. Moreover, they also state that all these signals are subjected to specific changes when the subject needs to face certain situations and they are in response to the central nervous system and the peripheral nervous system. In particular, the peripheral nervous system is divided into the somatic nervous system and the autonomic nervous system and, together with the central nervous system, the latter affects the physiological signals while the subject is emotionally stimulated. Throughout the research made in this field, it was found that it was relatively difficult to precisely reflect emotional changes by using a single physiological signal, while if multiple physiological signals are used it is possible to get to more accurate results as it is possible to check in the works presented and analysed in chapter 1.3.4.

Regarding the methodology of data analysis, emotion recognition studies can be mainly divided into two categories. One is using the traditional machine learning methods, such as the Support Vector Machine (SVM) models, the Linear Discriminant Analysis (LDA) model and the k Nearest Neighbour model (k-NN), which require carefully designed hand-crafted features and feature optimization. The other is using the deep learning methods, which can learn the inherent principle of the data and extract features automatically. Signal pre-processing, which is included both in traditional methods and deep learning methods, is adopted to eliminate the noise effects caused by the crosstalk, measuring instruments, electromagnetic interferences, etc. For the traditional machine learning methods, it is necessary to explore which characteristics of the original signals contain the information related to the perceived emotion and select from them the most important

features to enhance the recognition model performance. After feature optimization and fusion, classifiers which are capable of classifying the selected features are used. Unlike the traditional methods, deep learning methods no longer require manual features, which eliminate challenging feature engineering stages of the traditional methods even though they require a very large amount of data in order to give acceptable results as well as they can be computationally expensive due to complex data models. In Figure 5, the whole emotion recognition framework is shown.

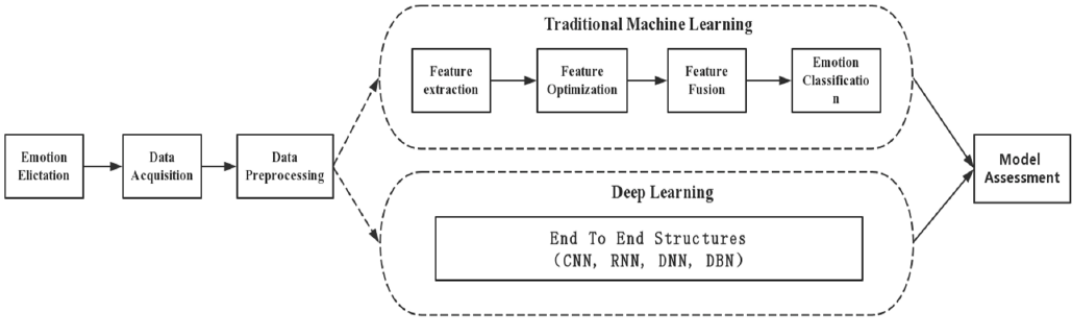

Figure 5 Pipeline to follow in emotion recognition algorithms using physiological signals. In the block diagram there is the distinction between traditional machine learning algorithms and deep learning algorithms. [7]

As it happens in the physiological measures explained before, the measures coming from eye-tracking can also be exploited to study the mechanisms that underlie behaviour and emotions as stated by M. K. Eckstein et al. [15], and so they can be subjected to the before described pipeline for emotion recognition algorithms.

1.3. EYE-TRACKING

As stated by J. Z. Lim et al. in [5] eye-tracking is a technological solution used to determine the point of gaze as well as the pupil diameter of a subject given a particular visual or audio-visual stimulus. Thus, it is a versatile and comfortable technology which is currently applied to many areas including cognitive science, medical research, and human-computer interaction. In particular, in the last two decades it has gained more and more importance in the brain imaging research as a way to study the mechanisms that underlie behaviour and emotional responses to certain kinds of stimuli as stated by W. Zheng in [16].

As stated by B. Mahanama et al in [17], oculomotor studies have provided a contribution in studying cognitive development and plasticity and it has been proven that eye-tracking measures can complement both behavioural and brain measures. Moreover, the possibility of using eye-tracking in this field provides a series of advantages if compared to other measurement systems such as EEG or fMRI. First of all, as stated by M. K. Eckstein in [15] due to the high sampling frequency used by modern eye-trackers which can range from 25 to 2000 Hz, it achieves a high temporal resolution up to the sub-milliseconds. Furthermore, eye-trackers are extremely comfortable and versatile, they can be used while seated comfortably at the table or, if they are

wireless, the data can be acquired out of the laboratory permitting the inclusion of a larger and more diverse population. Moreover, eye-tracking is an extremely versatile technology due to its comfort, as it does not need the usage of conducting gel or electrodes such as in the case of the electroencephalography or electrocardiography, and the subject does not need to stay still such as in the case of fMRI (functional magnetic resonance imaging). For all these reasons, less time is needed for the experiment to be set, which is particularly relevant for experimentation with individuals with specific pathologies or children, as time-efficiency is more relevant.

1.3.1. PHYSIOLOGICAL BASIS

As stated by M. K. Eckstein in [15], the most commonly used ocular measure in cognitive and affective studies is that of eye gaze, since it provides a moment-by-moment assessment of thought processes in a wide variety of context. The authors also state that there are other two measures that can be used due to their link with the mechanisms underlying human behaviour: pupil dilation and spontaneous blink rate. As shown in Figure 6, pupil dilation and constriction are caused by two antagonistic muscles which together with other muscles are responsible for eye-movements: the *dilator pupillae* and the *sphincter pupillae*.

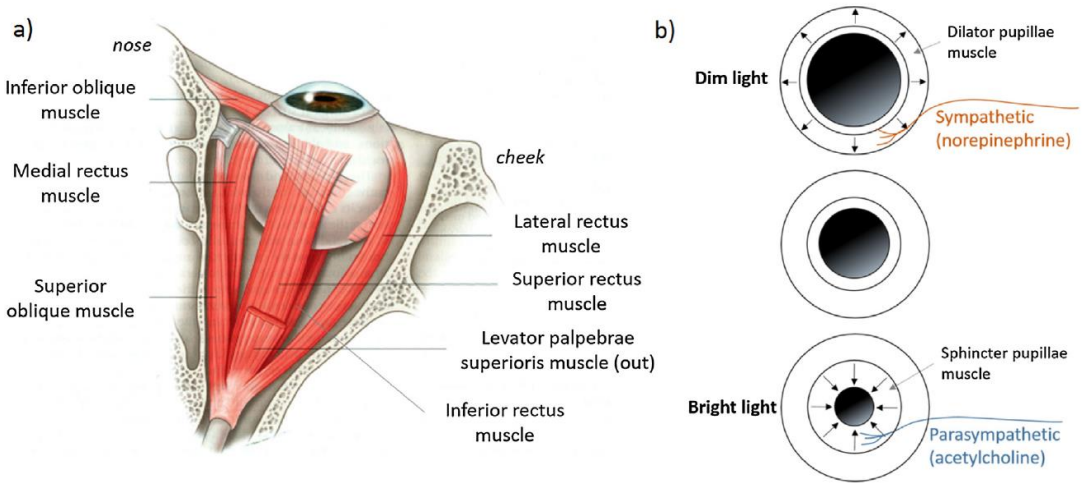

Figure 6 Eye muscles responsible for eye movements and pupil dilation and contraction. a) Superior view of the eye. b)Top: The dilator pupillae muscle. Bottom: The sphincter pupillae muscle. [14]

The constricting sphincter muscle not only receives input from brain systems involved in the pupillary light reflex, but, together with the dilator pupillae, receives inputs from brain systems involved in cognitive and autonomic functions [15]. In particular, pupil dilation is controlled by the Autonomic Nervous System (ANS). In particular, it is modulated by the noradrenergic locus coeruleus, a small nucleus in the brainstem that plays a central role in the regulation of physiological arousal. The main function of locus coeruleus is the releasement of norepinephrine, a neuromodulator which is essential for normal brain development and whose projections to the locus coeruleus's cortical target structures are already present at birth. The relationship between

the pupillary system and the locus coeruleus - norepinephrine activity has been established through numerous anatomical and physiological studies in both adult humans and animals. In fact, it has been demonstrated that there is a strong temporal coupling between locus coeruleus firings and pupil diameter in monkeys [18].

Regarding blink nature, the muscles that are engaged in the opening and closing of the eyelids are the *levator palpebrae* superioris and *orbicularis oculi* muscles. Although the precise neural circuitry that controls blink rate still requires further investigation, it has been hypothesised that dopamine modulates the frequency of spontaneous blinks indirectly [19]. Thanks to this link between dopamine activity and blink rate, we are able to indirectly assess cognitive control. As a matter of fact, our ability to control impulses, maintain long-term goals, and flexibly adapt to changing rules from the environment are all supported by dopamine. During goal-oriented behaviour, dopamine aids in the maintenance of abstract goals in higher levels of the cognitive control hierarchy, while also allowing flexibility in updating lower-level rules guiding attainment of subgoals. Most of the research in this domain has associated baseline blink rate with cognitive control, often showing a positive linear relationship [17].

1.3.2. EYE-TRACKING MODALITIES

There are different types of eye-tracking technology, and the main settings are the desktop eye-tracking, the mobile eye-tracking and the eye-tracking in virtual reality [5]. Desktop eye-tracking can be implemented only if the desktop which is being used comes with an eye-tracker. High-end desktop eye-trackers' main components are cameras and projectors which use infrared technology and they are exploited for extracting information about what is attracting the user's attention [5]. A cheaper alternative to this kind of desktops are low-cost cameras which can be placed close to the desktop. The main disadvantages provided by these cameras are, first, their accuracy is more limited than the infrared eye-trackers and, second, they do not work properly in low-light environments. For what concerns the mobile eye-tracking, it is typically mounted onto a lightweight pair of glasses. The mobile eye-tracking (MET) data acquired by the glasses are obtained in the natural environment of the subject and are then usually sent to an application of a mobile phone. Nevertheless, the main drawbacks of MET are that they must be used in highly controlled environments and they are typically rather expensive [5]. Finally, eye-tracking technology is increasingly being incorporated into many virtual reality headsets. Thanks to the virtual reality headset, it is possible to recreate any type of virtual environment while continuously acquiring information about what the subject is most interested in and what is its pupil response to the stimulus. This latter modality is the one used in this project [5]. Figure 7shows an example of eye-tracking device currently in commerce per each one of the three modalities. In particular, these devices are the Gazepoint GP3 eye-tracker (desktop eye-tracker), the Tobii Pro Glasses 2 eye-tracker (MET) and the HTC VIVE Pro eye (virtual reality headset).

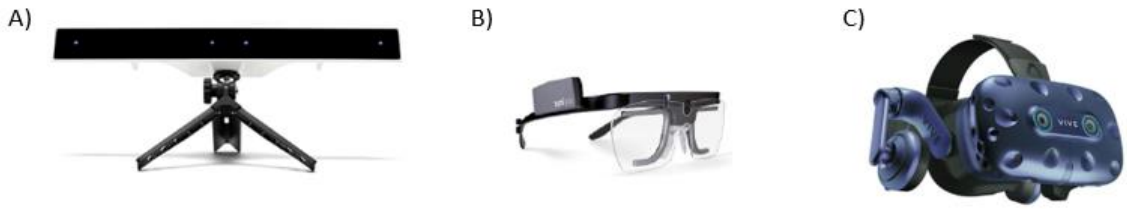

Figure 7 A) Gazepoint GP3 eye-tracker, B) Tobii Pro Glasses 2 eye-tracker, C) HTC VIVE Pro eye.

1.3.3. EYE-TRACKING MEASURES

All the measures that can be extracted from eye-tracking data presented in this chapter have been grouped by Bhanuka Mahanama et al. [17]. In total, there are five distinct types of eye movements, two of them are called gaze-stabilising movements while the other three movements are called gaze-orienting movements. This chapter will be focused on gaze-orienting movements, as they are the ones that lead the fovea toward objects of interest and therefore they have the potential for wide adoption in several applications and research [17]. The oculomotor events responsible for gaze-orienting movements are: fixations and saccades, smooth pursuit, fixational eye movements, blinks and ocular vergence. Nevertheless, only fixations, saccades and blinks are going to be analysed in this chapter since they are the measures that are later exploited in this study. Finally, another measure which can be obtained through eye-tracking data is the pupil diameter.

1.3.3.1. FIXATIONS AND SACCADES

Eye movement can be interpreted as a sequence of fixations and saccades. A fixation is a period where the visual gaze remains at a particular location. On the other hand, a saccade is a rapid eye movement between two consecutive fixations. Humans perform 3–5 saccades per second, but this rate varies with current perceptual and cognitive demands [17].

Once fixations and saccades have been recognised, several metrics relevant to oculomotor behaviour can be derived from them. For what concerns fixations it is possible to derive count and duration of the fixations. Fixation count is the number of fixations identified within a given time period, while fixation duration indicates a time period in which the eyes stay still in one position. Generally, the distribution of fixation durations is non-normal and left-skewed with typical median durations of 200-250 ms and mean durations of 300-350 ms as stated by S. Negi et al. in [20], longer or shorter fixations can be possible depending on the level of cognitive processing of the subject [17]. The average fixation duration is often used as a baseline to compare with fixation duration data at different levels. Regarding the measures of saccade, we can extract amplitude, direction, velocity, latency and rate. The amplitude of a saccade is the distance travelled by a saccade during

an eye movement and it can be measured either by visual degrees (angular distance) or pixels. It is usually approximated via the Euclidean distance between fixation points. The direction of a saccade can be represented as either an absolute value, a relative value, or a discretized value. Absolute saccade direction is calculated using the coordinates of consecutive fixation while relative saccade direction is calculated using the difference of absolute saccade direction of two consecutive saccades. Saccade velocity is calculated by taking the first derivative of time series of gaze position data and usually the main interest lies in the average saccadic velocity and in the peak saccadic velocity. Finally, saccadic latency is the duration between the onset of a stimulus and the initiation of the saccade while saccadic rate is the number of saccadic eye movements per unit of time.

1.3.3.2. BLINKS

A blink is essentially the closing and reopening of the eyelids. A blink is considered voluntary blink if it originates from a voluntary action, while it is called spontaneous or reflexive if it is non-voluntary. In particular, reflexive blinks are generated if they are a reaction to an external stimulus as a form of protection, while spontaneous blink is a rapid and automatic closing and opening of the eyelids essential for tear film spreading over the ocular surface. Furthermore, spontaneous blinks may also indicate the state of an individual. For example C. Ranti et al. in [21] demonstrate that spontaneous blink rate patterns can be used to measure changes in individual and group engagement that unfold over relatively short and long timescale (from one to sixty seconds).

Once the blinks have been identified, it is possible to extract different measures of interest. First of all blink rate can be determined, which is usually measured in blink per minute. Blink rate is not only subjected to factors such as lighting, temperature, wind, age and sex, but it is also influenced by the degree of concentration of the subject [17]. E.g. in the study conducted by A. Maffei et al. [22] it is shown that blink rate can be used as an index of attention and emotion during film clips since blink modulation is related with the motivational relevance and biological significance of the stimuli. Secondly, it is possible to compute the blink amplitude, the measure of the downward distance of upper eyelid travelled by in the event of a blink.

1.3.3.3. PUPIL DIAMETER

The pupil diameter can be measured through modern eye-trackers and the measure captures both the tonic and the phasic components of the pupil dilation [17]. The tonic component is the effect of the pupil diameter changes caused by slow contractions, while the phasic component is the effect of quick or transient contractions [17]. The human pupil diameter can range from a minimum of 2 to a maximum of 8 mm [23] and changes in pupil diameter can be due to the pupillary light reflex, the dynamic changes in pupil size induced by changes in the environmental luminance as stated by C. Daluwatte in [24], or due to the autonomous nervous system activity [17]. From the pupil diameter measure, a series of time and frequency measurements can be extracted such as

the mean value of the PD and the power contained by the power spectral density of PD in specific frequency bands.

1.3.4. EYE-TRACKING AND EMOTION

As stated by G. L. Lohse et al. in [25], due to the link present between eye-tracking measures and the autonomous nervous system, in the last decades many studies have been developed in the area of cognitive science and affective information processing in which eye-tracking data was exploited. Nevertheless, a relatively small number of studies in which only eye-tracking data was used for emotion recognition has been achieved.

In order to corroborate the link between the ANS and the pupillary system, Timo Partala et al. [26] have presented a study in which they subjected thirty participants to positive and negative highly arousal sounds and emotionally neutral sounds. The results showed that the pupil size was significantly larger during both emotionally negative and positive stimuli than during neutral stimuli. Moreover, several studies have been proposed to exploit pupillary behaviour in emotion recognition studies. For instance Areej Babiker et al. [27] have proposed a study in which they identified pupil dilation differences due to individual's positive and negative emotional states. As a result, they noticed a significant increase in pupil dilation during negative emotions compared to positive ones as well as a steeper, higher, more sustained and longer dilation in high arousal negative stimuli. A. Alhargan et al. in [28] propose a study in which, as in the forementioned work of Areej Babiker et al., the pupil dilation signal is used for an emotion recognition algorithm. The participants of this study are subjected to different sets of affective games while eye-tracking data is obtained. The defined classes of emotions are relative to the level of arousal (low, neutral, high arousal) and to the level of valence (negative, neutral, positive) of the games while the machine learning model used was a SVM model. As a result, A. Alhargan et al. get an average accuracy of the model equal to 70.0% for the arousal classification and 56.1% for the valence classification. Another study carried out by A. Alhargan et al. in [29] consists in the implementation of an emotion recognition algorithm given as input to the subject a set of video games. The defined classes are the same as the one defined in [28] as well as the trained machine learning model, but in this case they decided to use not only features extracted from the pupil dilation signals, but also to use features extracted from the oculomotor events achieving an average accuracy of the 71.4% for the arousal level classification and an accuracy of the 58.7% for the valence level classification. A worth-mentioning study carried out by P. Tarnowski is explained in [30]. In this study an emotion recognition algorithm which exploits only eye-tracking data is presented. In particular, in this study the Valence-Arousal plane of emotions is divided in three different sections representing the three categories used as output from the machine learning models adopted. On the contrary, as input to these models they decide to use a set of eleven features extracted from eye-tracking data obtained after the stimulation of the subject with movie clips. The exploited features were both concerning oculomotor events and pupil dilation. As a result, P. Tarnowski et al. obtained a maximum classification accuracy using a SVM model of the 80%. From now on a series of emotion recognition studies in which more than one physiological signals are going to be cited, yet only the results

relative to the use of solely eye-tracking data are going to be discussed. In [31] L. Soleymani present an emotion recognition algorithm trained using eye-tracking data of subjects stimulated by a set of video clips whose purpose was the elicitation of different kinds of emotional responses. The input features extracted are got from the analysis of the pupil dilation signal and the trained machine learning model is a SVM. As a result, an accuracy of the 71.1% is obtained in the case of the arousal level classification (calm, medium aroused, activated), while an accuracy of the 66.6% is obtained in the case of the valence level classification (unpleasant, neutral, pleasant). Another study where the SVM model is used is proposed by W. L. Zheng in [16]. In this study a series of emotional audio-visual clips are used to elicit positive, negative or neutral emotions. Also in this case, only the pupil dilation features are extracted and exploited and an average accuracy of the 45.78% and of the 58.90% is obtained exploiting different set of features. Finally, the work of Y. L. Lu in [32] can be cited. In this study a set of eye-tracking features extracted from the pupil dilation signals and from the oculomotor events is given as an input to a SVM model. The labels proposed as output to the machine learning model are relative to the kind of emotion elicited (positive, neutral, negative), and the obtained results show an average accuracy of the 77.80%.

Several studies have investigated the most suitable machine learning models for emotion recognition. J.Z. Lim et al. in [5] present a comprehensive review on emotion recognition using eye-tracking technology, and in particular they state that there are many classifiers that can be used in emotion recognition algorithms such as Naïve Bayes classifiers, KNN, decision trees, SVM and neural networks. Nevertheless, they also state that most of the studies available in literature choose the SVM as classifier and in most of these cases an accuracy higher than 80% is not achieved. Moreover, they also state that, even if different SVM have been implemented (linear, non-linear, polynomial etc.), usually the authors do not specify the type of SVM chosen. L. Shu et al. also affirm in [14] in their comprehensive review on physiological signal based emotion recognition that, even if all the fore-mentioned models can be exploited, the SVM is the most widely used machine learning model used in emotion recognition algorithms, even in the studies in which a multimodal approach is used. The works by M. Soleymani et al. in [31] and W. Zheng et al. in [16] are good examples. In these studies, the emotion recognition algorithms use, respectively, SVM and linear SVM. On the contrary, P. Tarnowski et al. in [33] opt for the use three different machine learning models: SVM, LDA and k-NN. Also in this case the highest accuracy was achieved by the SVM model.

1.4. MOTIVATIONS

As it has been largely discussed in the previous chapters, the implementation of an algorithm for the automatic recognition of emotions could provide a huge contribution to different research fields and applications. Up to now, most of the studies in this field exploit a multimodal approach in which more than one physiological signal is exploited to recognise what emotion the subject has perceived as a response to specific stimuli ([5], [32]). The current eye-tracking devices are usually wearable and comfortable, so the implementation of an emotion recognition algorithm which has as an input solely the eye-tracking data could provide the possibility to avoid uncomfortable experimental settings as well as the possibility to make the subject respond to the stimulus in the most natural way possible. Moreover, the use of a virtual reality headset enables the administration

of realistic and dynamic stimuli of any kind, making the subject dive into a completely new environment. Most of the studies present in literature, in which only eye-tracking data is used, provide a lower accuracy result with respect to the modality fusion attempts in which more physiological signals are exploited [32]. Thus, this study represents one of the few attempts to recognise emotions using only eye-tracking data obtained through a virtual reality headset. Additionally, even if this study is developed on healthy subjects, hypothesising as a result a robust algorithm with high level of accuracy, it could give a contribution to new experimental studies whose purpose is the neurorehabilitation of patients such as the project "*ACTIVA: Alteraciones de la Consciencia: protocolos de Intervención y Valoración Activa"* by *Conselleria de Educación, Investigación, Cultura y Deporte of Generalitat Valenciana* (SEJI/2019/017). In particular, the ACTIVA project is a study whose final purpose is the investigation of physiological responses given as input to patients with disorders of consciousness emotional audio-visual stimuli. Together with other physiological signals (electroencephalography, electrocardiography, fNIRS), eye-tracking data is acquired, and the analysis of it can give a contribution to the treatment of diseases related to disorders of consciousness.

1.5. HYPOTHESIS AND OBJECTIVES

1.5.1. GENERAL HYPOTHESIS

The main hypothesis of this project is that it is possible to recognise and identify different emotional responses through the analysis of eye-tracking data. In particular, it is hypothesised that the ocular response of the subject can be exploited in machine learning algorithms for the classification of the delivered emotional audio-visual stimuli such as a video of a known and loved person, a video of a completely unknown person, and a neutral resting phase. Moreover, it is hypothesised that ocular data will be also able to distinguish the kind of emotion the subject has felt in relation to the level of arousal (or excitation) and valence (or pleasure) perceived.

1.5.2. SPECIFIC HYPOTHESIS

In order to establish if the general hypothesis is truly achieved, it is necessary the formulation of specific hypothesis whose achievement is at the base to draw out general conclusions.

For what concerns the visual stimuli used during the study:

- The view of a video of a referred person answering questions about the participant will provide a strong emotional response.
- The view of a video of a completely unknown person will produce to the subject a neutral emotional response different from the one produced by the video of a referred person.

For what concerns the self-assessment of emotions:

- The SAM questionnaire will be used for the self-assessment of emotions after the delivery of each visual stimulus to the subject. It is hypothesised that the evaluations of valence and arousal related to the videos of the referred people are different from the evaluations related to the videos of the unknown people.

For what concerns the automatic recognition of emotion:

- The oculomotor behaviour of a participant is strictly related to the emotional response to the provided stimulus. Thus, the three different stimuli provided (known and loved person, unknown person and resting), produce different oculomotor behaviours.
- The level of pupil dilation of a subject and the pupillary response is related to the emotional response to the provided stimulus.

1.5.3. OBJECTIVES

This study was designed in order to fulfil three main objectives:

1) Investigate the ability of machine learning models to discern the physiological responses elicited during the resting phases and the video phases.

2) Investigate the ability of machine learning models to discern the physiological response elicited by audio-visual stimuli corresponding to acquaintances and strangers.

3) Investigate the ability of machine learning models to discern the physiological response elicited by emotional audio-visual stimuli differentiated in terms of the level of valence (pleasure) and arousal (excitation) perceived by the participant.

2. METHODS

2.1. PARTICIPANTS

A convenient sample of 15 healthy subjects took part to this study. Among these, 6 participants are men and 9 participants are women, so 40% of the participants are men, while 60% are women. The age of the participants goes from a minimum of 24 to a max of 42 with a mean and standard deviation of 32.4 ± 6.46. The majority of participants are volunteers from the *Instituto de Investigación e Innovación en Bioingeniería* of the Polytechnic University of Valencia. All of them are volunteers, they do not receive any fee for their participation in the study. They have been selected considering the following inclusion and exclusion criteria.

Inclusion criteria:

- They must have more than 18 years

- They must be volunteers

Exclusion criteria:

- They must not have had neurological or cardiac illnesses
- They must have normal or corrected-to-normal vision and normal hearing

The participants who took part in the experimentation as well as their referred person captured in the videos had to sign a written consent regarding our use of their image in the study.

2.2. INSTRUMENTATION

2.2.1. SOFTWARE

In this subchapter the software used in this study are described. In particular the software used are:

- Matlab (Matworks, Massachusetts, United States): software used for the processing and filtering of the eye-tracking data, for the implementation of statistical analysis, for the training of machine learning models and for the graphical representation of the results.
- Adobe premiere pro (Adobe, California, United states): video editing software used for the editing of the audio-visual stimulation
- e-Prime (Psychology software tools, Inc., Pensilvania, United states): stimulus presentation software used for the design and execution of the experiment.
- Unity: (Unity technology, San Francisco, United states): cross-platform game engine used for the synchronisation of the eye-tracking data with the data acquired from the other instrumentation used (electroencephalogram, electrocardiogram and fNIRS).

2.2.2. AUDIO-VISUAL STIMULATION

The audio-visual stimulus consists in videos of people referred to the subject (relatives, friends etc.) answering questions about their personal life (memories, anecdotes, thoughts…) and in videos of people unknown to the subject answering the same questions about another participant. These videos were delivered to the subject through the virtual reality headcap HTC VIVE Pro Eye. Topics of the videos that will be included on the experiment will follow common criteria of adequacy, in particular they are meant to elicit high arousal and positive valence emotions in the subject. Participants loved ones, who record the videos, are provided with a list of 8 topics to talk about. Six videos from those are selected according to a predefined order of relevance. Videos are either self-recorded following a step-by-step guide or by an experimenter in such a way that the videos used

in the experimentation in different participants are as standardised as possible. An example of a video realised following the guidelines is shown in Figure 8 :

Figure 8 Example of well realised and set video

The guidelines for the recording of the videos are provided in Spanish to the participant and they are explicitly explained in chapter 1 of the annexes. Each one of the videos used as stimuli to the subject has a total duration of 55 seconds and is divided in two main parts: the first 5 seconds of the video the referred person or the unknown person introduce him/herself to the subject and in the remaining 50 seconds the referred person or the unknown person answer to one of the 8 questions forementioned. Since a total of 15 participants took part to the experimentation and since per each one of the participants a total of 12 videos are provided (6 of the unknown person and 6 of the known person), a total of 180 are used as audio-visual stimulation for a total duration of 9900 seconds in which the eye-tracking data is recorded. In chapter 2.3.2 additional information is given about the video processing and editing of the videos. Moreover, as it is going to be discussed in 2.3.3 each video is followed by a resting phase with the duration of 55 seconds in which eye-tracking data is recorded. Therefore, a total of 180 resting phases are recorded for a total duration of 9900 seconds.

2.2.3. DATA COLLECTION

In this subchapter a brief description of the data collected during the experimentation is going to be delivered. In particular, the data collected during the experimentation is regarding the eye-tracking data, the Self-Assessment Manikin and the questionnaires that the participant is asked to fill after the experimentation which are the Film IEQ (Immersive Experience Questionnaire for Film and TV) and the ad hoc questionnaire.

EYE-TRACKING:

The virtual reality headcap HTC VIVE Pro Eye, which embeds eye-tracking technology, was adopted for both audio-visual stimulation playback and eye-tracking data collection. Although this headcap is usually used to create immersive virtual reality environments, 2D videos are employed in this study. The use of a VR headcap instead of a simpler and cheaper desktop eye-tracking system

is due to the requirements needed in the project ACTIVA, within which this study is located. In this project, besides collecting more physiological signals, this same study is extended to patients with severe brain injuries who have an oscillating and low-level attention as well as low or inexistent motor control. In these cases, the instruction of keeping the focus on a regular screen for a prolonged period of time may not be possible. Thus, the use of the virtual reality headcap to display the stimuli facilitate the maximum immersion of the participants and also guarantees that even if they maintain their eyes closed during the acquisition, the auditory stimulus is still provided to him/her.

Specifically, the relative position of the gaze and the pupil diameter will be acquired during the experiment. The headcap will also be used to provide the audio-visual stimulation to the participant.

Figure 9 virtual reality headcap HTC VIVE Pro Eye

In Table 1 the specifications of the headset are shown, while in Table 2 it is possible to find the specification of the Eye-tracking device embedded in the headcap powered by Tobii.

SCREEN	Dual OLED 3.5” diagonal
RESOLUTION	1440x1600 pixels per eye
REFRESH RATE	90 Hz
FIELD OF VIEW	110 degrees
AUDIO	Hi-res-certified headset Hi-Res-certified headphone High-impedance headphone support Enhanced headphone ergonomics
INPUT	Dual integrated microphones
CONNECTIONS	USB-C 3.0 1.2, Bluetooth
SENSORS	StreamVR Tracking G-sensor Gyroscope Proximity Eye Comfort Setting (IPD) Eye tracking

Table 1 specifications of the virtual reality headset

GAZE DATA OUTPUT FREQUENCY	120 Hz
ACCURACY	0.5°-1.1°
CALIBRATION	5-point
TRACKABLE FIELD OF VIEW	110°
DATA OUTPUT	Timestamp, Gaze origin, Gaze direction, Pupil position, Pupil size, Eye openness
INTERFACE	HTC SRanipal SDK
SDK ENGINE COMPATIBILITY	Unity, Unreal

Table 2 eye-tracking specifications

Finally an additional remark must be made. Usually at the beginning of eye-tracking acquisitions, a calibration process is achieved in order to get reliable information about the direction of gaze of the subject. In this study the calibration process is not achieved because the same experimental protocol needs to be repeated as similar as possible in patients as a purpose of the ACTIVA project but in patients no calibration is feasible. Nevertheless, as it is going to be discussed later, in this study no areas of interest are used, so the information of the precise gaze point is not needed.

Self-assessment of emotions:

The evaluation of the emotions perceived by the subject is performed through the Self-Assessment Manikin (SAM). As it has already been discussed in chapter 1.2.2.1, the SAM is a self-reported scale used to directly assess the valence, arousal and dominance associated with the emotional response to a stimulus in a 9-point Likert scale. The evaluation phase is presented to the subject right after each audio-visual stimulus. The subject must select the score for valence, arousal and dominance relative to the previous stimulus. In particular, the action of selecting a precise score is achieved through one of the controllers connected to the headcap as the one shown in Figure 10.

Figure 10 HTC VIVE Pro Eye controller

During the evaluation phase the screen shown in Figure 11 is displayed to the subject. When the participant presses the button on the controller the green circle around one of the values in the valance, arousal and dominance scale respectively shifts towards the next value on the right. If the green circle gets to the value “9” and the subjects wants to select a lower score, he/she just needs to press again the button and the green cursor goes back to “1”. When the subject wants to select a score he/she just needs to press for an extended period of time the same button. While pressing, a blue line on the corner of the SAM starts filling the black gap on the side and when the black gap is completely filled the score is selected and the selection phase continues in the line behind until all the three scores have been selected.

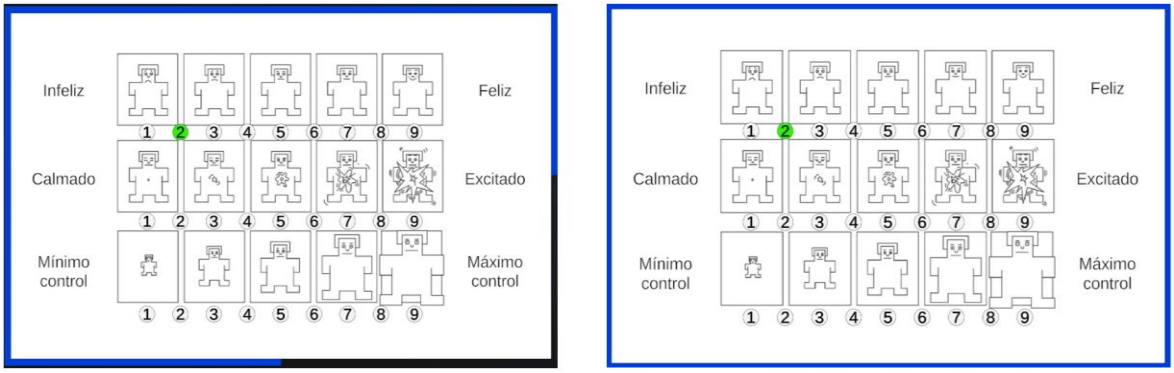

Figure 11 Example of SAM visualisation on the headcap’s screen; on the left the blue bar is filling while the subject is pressing the button, on the right the blue bar is completely filled

Video viewing experience:

The Immersive Experience Questionnaire for Film and Television (Film IEQ) is used to assess the video viewing experience. This questionnaire is based on the Immersive Experience Questionnaire (IEQ) by Jennet et al. [34] and adapted to video viewing by Rigby et al. [35]. A total of 24 questions are selected, each question is related to one of these four factors: (1) involvement, (2) captivation, comprehension (3) and real-word dissociation (4). It is asked to the subject to answer these questions using 1-7 Likert scale and finally overall immersion scores can be computed by first inverting responses to item 5, 6 and 7 (1 becomes 7, 2 becomes 6, 3 becomes 5, etc.) then summing all responses.

In our study, once the acquisition has come to an end, the subject is asked to fill the Film IEQ. The 24 questions together with the relative factor (from 1 to 4) are listed in chapter 2 of annexes.

Level of acknowledgement and influence of the instrumentation.

An ad-hoc questionnaire is proposed to the subject after the Film IEQ questionnaire. This questionnaire is formulated to check the level of acknowledgement about the experimentation the subject had before taking part to it and also to check what degree of influence the use of all the instrumentation used has had. Note that it is of relevant importance that the subject does not know

a priori what the experimentation is about in order to produce a higher emotional response. It is asked to the subject to answer these questions using 1-7 Likert scale. In the followings, the questions forming this questionnaire are listed:

1. Did you have idea of what the experiment consisted in before taking part in it?
2. Was the instrumentation uncomfortable during the experiment?
3. Do you think that your level of concentration during the experiment was influenced by instrumentation?

2.3. PROCEDURE

In this subchapter a general view of the procedure followed before and during the experimentation is presented. First of all, once the participants have been selected, we need to contact the referred person, the one related to the participants. It is asked to the referred person to record a video following the guidelines presented in chapter 2 of the annexes. Successively, once the video is available the video processing phase using the editor Adobe Premiere Pro starts. This phase consists in the editing of the videos in such a way they are as standardised as possible in terms of luminance, volume and size. Finally the experimentation phase can start.

2.3.1. VIDEO COLLECTION

In this subchapter, a general view of the procedure followed before and during the experimentation is presented. Before the experimentation, the participants were asked to provide the contact of an acquaintance who is emotionally linked to them. An experimenter of the ACTIVA project contacted the referred persons and asked them to record and provide a video following the guidelines presented in chapter 1 of the annexes. Once the videos were available to the experimenters, the luminance, volume and resolution of the videos were normalized. Once the videos were edited, the experimenters scheduled a meeting with the participants.

2.3.2. VIDEO EDITION

The video edition of the videos sent from the referred people is an essential step in the project pipeline. The videos presented to the participant during the experimentation are of two types: videos of their referred person and videos of the referred person to other participants used as unknown person. Since each one of the participants records the video by him/herself, differences in terms of luminance, volume and size may be present between different video types. These differences in the videos may imply differences in the data collected that may be later interpreted by the implemented emotion recognition algorithm as changes in the physiological response of the subject and this must be avoided.

Adobe Premiere Pro is used for the video edition. The video sent from the referred person consists in him/herself introducing him/herself to the participant and answering to a series of questions about the participant. The first step in the video processing consists in adjusting the size of this video in such a way that the subject in the video of the known person (from now on referred to as video of type A) and the subject in the already edited video of the unknown person (from now on referred to as video of type B) are placed in the same position putting particular attention in the size and the position of the face. In order to achieve this step it is important to place instead of the black background gap a mask whose colour matches as much as possible the colour of the background wall present in the video. Later, it is needed to compare the first frame of the A video with the first frame of the B video in order to detect the possible differences in luminance between the two types of videos. The editing features of the video are changed in order to match the luminance value of the A videos with the one of the B videos. Later, the volume levels are matched and kept between -24 and -6 dBs, while the background noise is removed. Once the matching between the two videos is achieved, the fragmentation and editing of the videos can start in order to build the sample videos that are used as stimuli in the experimentation. The total duration of the sample videos must be equal to 55 seconds and it is composed of 5 initial seconds in which the person introduce him/herself and 50 seconds in which the person answer one question about the participant. The initial 5 seconds of each one of the sample videos consist in the introduction of the person to the participants, as it is already explained in chapter 1 of the annexes. In particular, it is asked to the referred person to include in the video a section in which they say the name of the participant, their relationship with the participant and their name. In the successive 50 seconds of the video there will be the answer of the referred person to one of the questions listed in chapter 1 of the annexes.

2.3.3. EXPERIMENTATION

Afte the edition of the video, the participant is contacted and a date for the experimentation phase is panned. At this point, after the signature of the written consent by the participant, the experiment[1] itself can start. The procedure of the experimentation follows the next steps:

1. Experiment description (~2 minutes):

   A description of the experiment is be provided to the participant.

2. Experimental setup (~60 minutes):

   The participant will be equipped with all the recording devices which are the EEG, the fNIRs, the ECG and finally the HTC VIVE Pro eye headset as shown in Figure 12.

[1] It must be mentioned that during the acquisition of the eye-tracking data, other kinds of instrumentation including EEG (electroencephalograph), fNIRS (near-infrared spectroscopy) and EEG (electrocardiograph), were attached to the subject in order to gather additional physiological data which will be later exploited in a wider project. In particular, the fore-mentioned project is called "*ACTIVA: Alteraciones de la Consciencia: protocolos de Intervención y Valoración Activa*", it has been proposed by the Neuro Rehabilitation and Brain Research group, a group part of the Polytechnic City of Innovation of the *Universitat Politècnica de València.*

3. Resting phase (~1 minutes):

   An example of the task will be done as a training.

4. Emotional task (~32 minutes):

   I. ideo phase: A video will be displayed in the middle of the screen of the HTC VIVE Pro Eye headset and, at the same time, in an external screen. The order of the videos displayed is in random order to avoid any predictability.

   II. Evaluation: This phase consists in the evaluation of the video previously shown using the SAM.

   III. Resting phase: The participant can see in the screen a white cross placed at the middle of a black screen, so no stimulus is presented to the subject.. It is only asked to the participant to stay still, and, if possible, look at the cross in the screen of the HTC VIVE Pro Eye headset.

Figure 12 Example of experimental setup

The steps I to III will be repeated 12 times (6 videos of a loved one and 6 videos of an unknown person). If the subject asks to the experimenter to have a break from the experiment for any reason, the experiment will be paused. The entire experimentation procedure will last approximately one hour and 35 minutes. A summary of the protocol is shown in Figure 13.

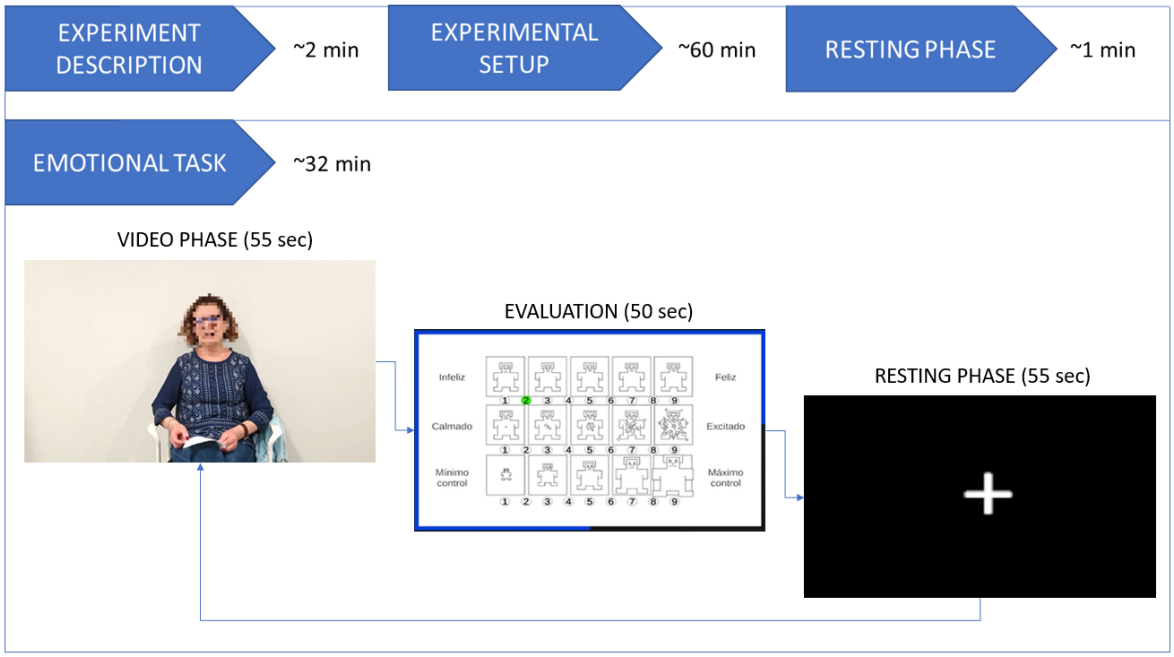

Figure 13 Scheme of the experimental protocol

2.4. DATA ANALYSIS

This chapter describes the processing of the eye-tracking data in order to extract the features of interest, the statistical analyses used for the features selection and the classification of the data using machine learning classifiers. The entire pipeline of the classification problem is shown in Figure 14.

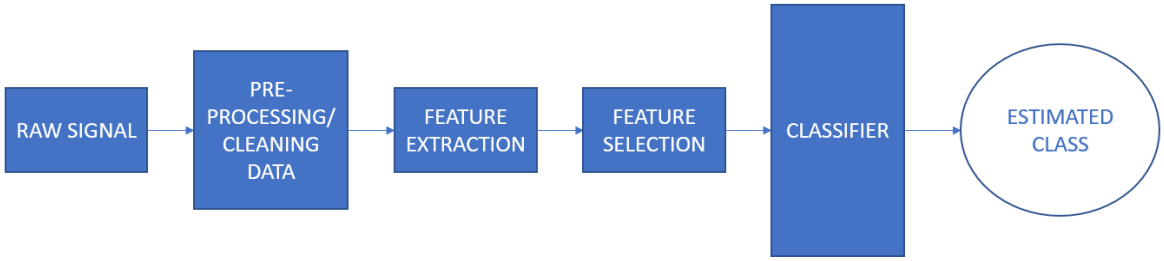

Figure 14 Simplified pipeline of the classification problem

2.4.1. PREPROCESSING OF THE DATA

The data obtained from the VR headset comprise x and y gaze coordinates expressed in centimetres and left and right pupil diameters expressed in millimetres. These variables are sampled by default at around 90 Hz of sampling frequency. Before further and feature-specific pre-processing steps are

done, a resampling procedure at 90 Hz has been implemented in order to have a constant sampling data. The pre-processing of the data consists in the pre-processing of the eye gaze coordinates, in the identification of blinking periods and in the artifact removal of the pupil dilation signal.

2.4.2. EYE GAZE COORDINATES

The pre-processing of the eye gaze coordinates consists in applying a low pass Butterworth filter of the 4th order with cut-off frequency equal to 5 Hz as suggested in [33]. This filtration step is important for the removal of high frequency noise which might be introduced by the electronic instrumentation. The difference between filtered and unfiltered eye gaze coordinates can be seen in Figure 15.

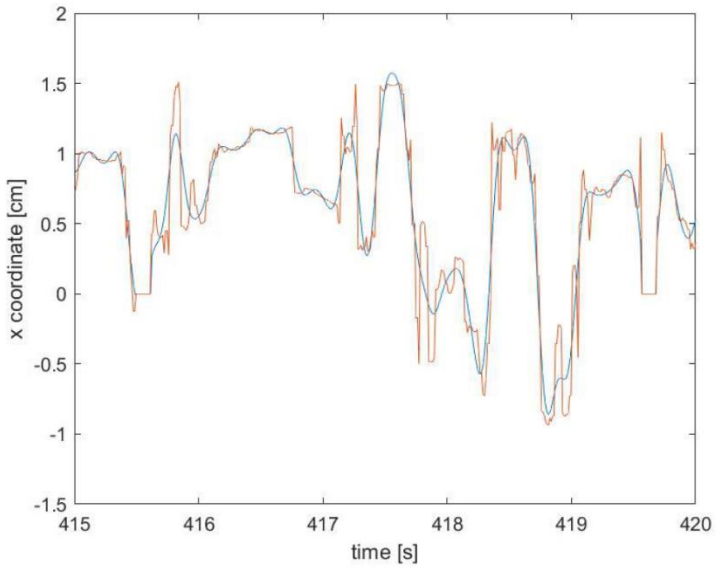

Figure 15 Example of X eye gaze coordinate unfiltered (in red) and filtered (in blue)

2.4.3. FIXATIONS AND SACCADES

The identification of saccades and fixations is a fundamental step in eye-movement analysis and it has a dramatic impact on the extraction of eye-movement metrics such as saccadic amplitude and fixation duration and so on higher-level analysis. In particular, different algorithms have been implemented for the fixations and saccades identification and they can give different results. Therefore, in this chapter the differences between this kind of algorithms are going to be discussed together with the description of the algorithm exploited in this study. Finally, the calibration phase needed to set the threshold values required by the algorithms is described.

Difference between identification algorithms:

The main differences between the identification algorithms lies in the way they use spatial and temporal information embedded in eye-tracking data, Salvucci et al. [36] propose a taxonomy of fixation identification algorithm. Especially, in their study Salvucci et al. identifies different

algorithms based on the way they account for spatial and temporal characteristics. For spatial characteristics they identify three criteria that distinguish three primary types of algorithms: velocity-based, dispersion based and area-based. Velocity based algorithms focus on the velocity information in the eye-tracking protocols, dispersion-based algorithms focus on the dispersion of fixation points while area-based algorithms identify points within given areas of interest (AOIs). On the contrary, for what concerns temporal characteristics, they include two criteria: whether the algorithm uses duration information and whether it is locally adapted. In particular, duration information refers to the duration of the identified fixations (considering that a fixation is rarely less than 100 ms) while local adaptability refers to the possibility of taking into account the influence that temporally adjacent points may have to the fixation identification. In [36] five different identification algorithms are explained and analysed: I-VT (velocity threshold identification), I-HMM (hidden Markov model identification), I-DT (dispersion-threshold identification), I-MST (minimum spanning trees identification), and I-AOI (area-of-interest identification). In this chapter only the I-DT algorithm is explained and analysed since it is the one that will be implemented later on this study.

I-DT algorithm:

Dispersion-based identification algorithms utilise the fact that fixation points tend to cluster closely together due to their relatively low velocity with respect to saccadic events. These algorithms require the definition of two parameters, the dispersion threshold and the duration threshold. The former is a value under which a set of points cannot be considered a fixation based on their dispersion in space, while the latter is a value under which a set of points cannot be considered a fixation based on their duration in time. The dispersion threshold and the duration threshold can be set a priori , The I-DT algorithm is a dispersion based identification algorithm that uses a moving window that spans consecutive data points checking for potential fixations. The moving window begins at the start of the gaze coordinate signals and initially spans a minimum number of points, determined by the given duration threshold and sampling frequency. I-DT then checks the dispersion of the points in the window by computing the value of a specific dispersion metric. In literature it is possible to find several metrics based upon spatial variance or area of samples, but in this study the dispersion metric proposed by Salvucci et al. is used and it is described in the equation (1).

$$D = [max(x) - min\,(x)] + [max(y) - min\,(y)] \qquad (1)$$

Where D is the dispersion value associated to the selected points, while x and y are the coordinates of the eye gaze in the time interval defined by the moving window. The dispersion value needs to be compared to the dispersion threshold. If the dispersion is below the dispersion threshold, the window represents a fixation. In this case, the window is expanded adding one sample at a time in the positive direction of the time axis until the dispersion of the window is above threshold. The final window is registered as a fixation. On the contrary, if the dispersion is above the dispersion threshold, the window is moved sample by sample to the positive direction of time, until the dispersion value associated to the new cluster of points is below the dispersion threshold.

The points which are not considered as a fixation are then categorised as part of a saccade since their spatial dispersion is higher than the threshold value. This process continues with window moving to the right until the end of the protocol is reached.

In this study the I-DT algorithm is used because it is a linear-time, potentially real-time algorithm that produces robust and accurate identification results [36]. The main disadvantage of I-DT is the need of setting a value to the thresholds required. These two thresholds highly influence the identification of fixations and saccades and choosing two different set of values may lead to completely different results. For instance, if a small dispersion threshold and a large duration threshold are chosen, it is possible to get to the identification of no fixations and so a completely wrong result. There are two different approaches for the setting of the dispersion threshold: in the first one, this threshold can be set to include 1/2° to 1° of visual angle if the distance from eye to screen is known, otherwise the dispersion threshold can be estimated from data-driven analysis through a calibration phase. As well as the dispersion threshold, the duration threshold, can be set a priori to a value between 100 and 200 ms ,since this is the minimum duration of a fixation as stated by Alastair G. et al in [37], or it can be determined through the calibration phase. The second approach will be used in this study and the calibration phase will be explained in the following subchapter.

Calibration phase:

For the setting of the threshold values needed by the I-DT algorithm, in this study it has been chosen to use a data-driven approach in order to find the threshold values that best fit our data and the chosen dispersion metric instead of using the values found in literature. This data-driven approach consists in a calibration phase proposed by J. Llanes-Jurado et al. in [38] in which a rule-based criteria are proposed to calibrate the threshold of the algorithm through different features, such as number of fixations and the percentage of points which belong to a fixation.

As a first step, it is needed to define the calibration criteria. In particular, this calibration criteria focus on a set of features that can be extracted by the overall dataset by average between all the subjects. Four features can be extracted: number of fixations, percentage of points classified as a part of a fixation, mean fixation time, and percentage of fixations inside area of interest (AoIs). In this study only the first three features are going to be exploited since no areas of interest are going to be defined. In the followings, it is possible to find the description of each one of the three features of interest, note that each feature must be computed per each subject and then averaged between all the subjects:

- **Number of fixations:** it consists in the average number of fixations that can be computed with the implemented I-DT algorithm per subject during the task. Starting from a low value of the dispersion threshold to a high value of it, it is expected to have an initial growth of this feature up to a maximum, followed by a decrease until we get for an extremely high value of the threshold the identification of a unique fixation. The parameters must be selected in such a way that the number of fixations exceed the maximum number of fixations, in fact the area close to the maximum is characterised by a high degree of instability.

- **Percentage of points classified inside a fixation:** it measures the percentage of points classified as part of a fixation with respect to the entire data set. Starting from a low value of the dispersion threshold to a high value of it, it is expected to have initially an exponential increase of this feature. Then, approximately in correspondence of the maximum number of fixations, it is expected that this feature gets to an elbow point. When the elbow point is reached, the growth of the curve becomes smoother until it gets to 100% of the points included in a fixation. The parameters must be selected in such a way that this feature exceed the elbow point and it is as high as possible.
- **Mean fixation time:** it measures the average fixation time per subject. This feature follows a linear relation with the dispersion threshold of the I-DT algorithm in 2D world-centered experiments [38], in fact the mean fixation time increases proportionally as the dispersion threshold increases. The reason of this proportionality is quite intuitive, increasing the dispersion threshold we have an increase in the number of points that are inside a fixation that leads to a higher mean fixation time. Thanks to this condition, it is possible to define an upper limit for the dispersion threshold and time window, in fact it is possible to set the maximum value of the mean fixation time depending on the task achieved in the experiment and considering that the upper limit established in the literature for this feature is of about 0.65 seconds [39].

These features are examined in terms of the dispersion thresholds and time windows in a grid search. The objective is to find a set of points that simultaneously satisfy the conditions imposed for each feature. This set of points would be the optimum to use for an I-DT algorithm. The grid used in this study to calibrate the algorithm goes from 0.1 to 0.5 with a step of 0.05 seconds for the minimum duration threshold as suggested by [38], while it goes from 0.1 to 2.5 with a step of 0.05 cm for the dispersion threshold. The values of the grid chosen for the dispersion threshold are defined in this study since no reference about the calibration was found in literature concerning the dispersion metric used. The calibration criteria are summarised in Table 3.

FEATURE	CRITERIUM
Number of fixations	After the maximum fixation number
Percentage of points classified inside a fixation	After the elbow point and as high as possible
Mean fixation time	Lower tha 0,65 seconds

Table 3 Calibration criteria

In the following figures it is possible to see the evolution of each one of the forementioned features in terms of the duration and dispersion thresholds got as a result of the calibration phase applied to the gaze coordinate data extracted in this study. In Figure 16-A it is possible to see that

in the first region of the dispersion threshold (approximately 0.1 up to 0.5 cm), the number of fixations increases up to a maximum, and this happens for all the values of the duration threshold. After that, the features decreases smoothly as both parameters get bigger. In Figure 16-B it is possible to see that for a dispersion threshold that goes from 0 up to 0.6, the curve grows exponentially until an elbow point. The percentage of points inside a fixation decreases as the time window lengthens but increases with the increment in the dispersion threshold. Figure 16-C shows that the mean fixation time depends linearly on the dispersion threshold for small values of duration threshold, while for higher values it presents a region of instability in the low values region of the dispersion threshold. The dependency of the mean fixation time on the duration threshold is still linear but it is less evident.

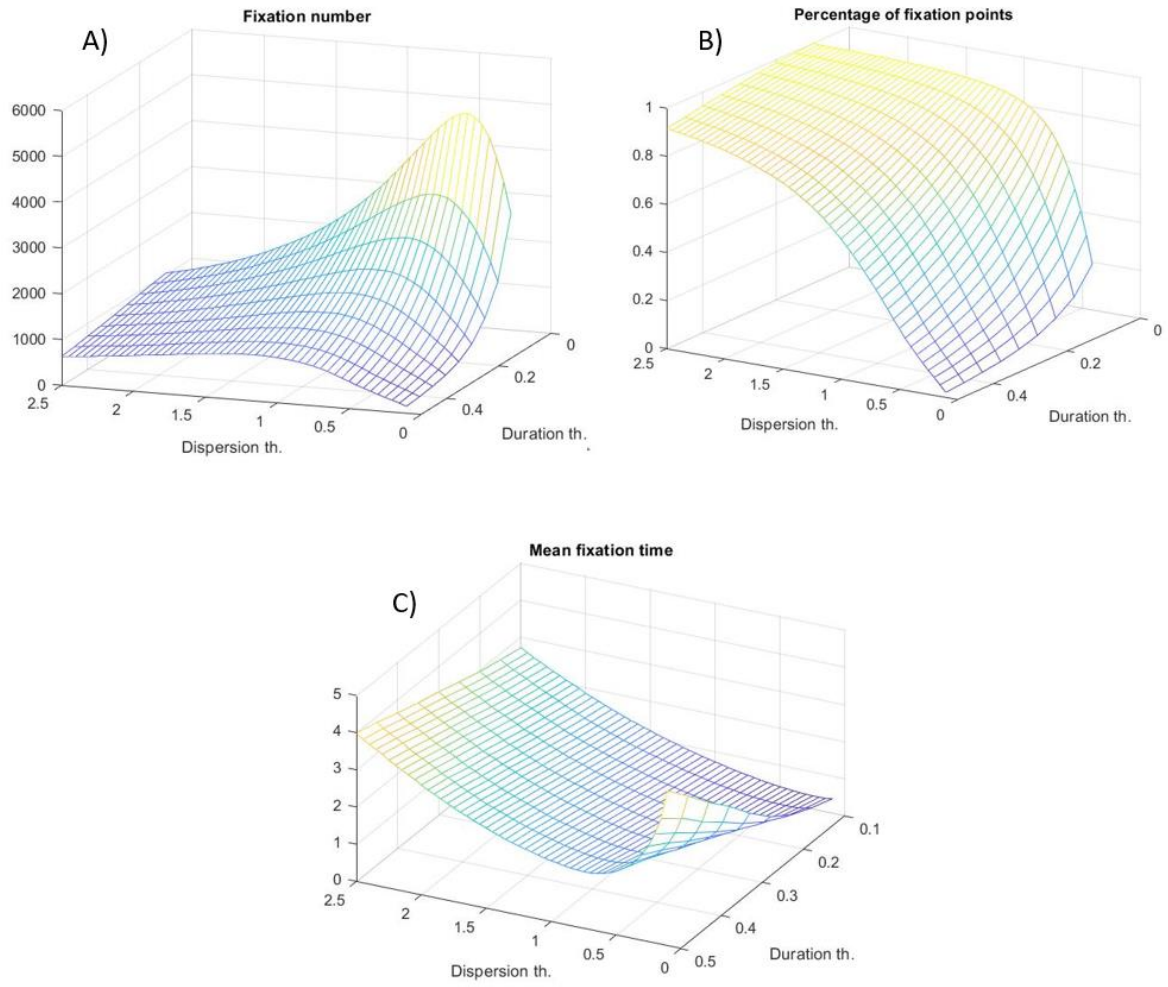

Figure 16 Graphical representation of the calibration results representing the calibration features in function of the dispersion and duration thresholds; A) Fixation number B) Percentage of fixation points C) Mean fixation time

At the end of the calibration process, a dispersion threshold of 1.0 cm and a duration threshold of 0.1 seconds were chosen since these values satisfies the calibration criteria.

2.4.4. BLINK IDENTIFICATION

Blink identification consists in identifying the intervals of time in which the subject has closed his/her eyelids during the experimentation. In all the time frames in which the eyes are closed we have a loss of data due to the fact that the eye-tracker is no more able to acquire the eye-movement metrics such as the eye-gaze coordinates as well as the pupil dilation. As a result, the output of the eye-tracker in those time frames present a specific and recognisable pattern as shown in Figure 17. In particular, if the eye-gaze coordinates are considered, the x and y coordinates are set at zero in the time frames in which the eyelids are closed. Thus, the blink identification is based on searching the instants of time in which both the x and y coordinates are set at zero. Moreover it is important to check if the duration of the period of time in which the x and y coordinates are equal to zero is consistent with the duration of a physiological spontaneous eye blink. In particular, in this study this interval is identified as a blink if its duration is smaller than 400 ms since it is the maximum duration of a typical blink as stated by C. Ranti in [21].

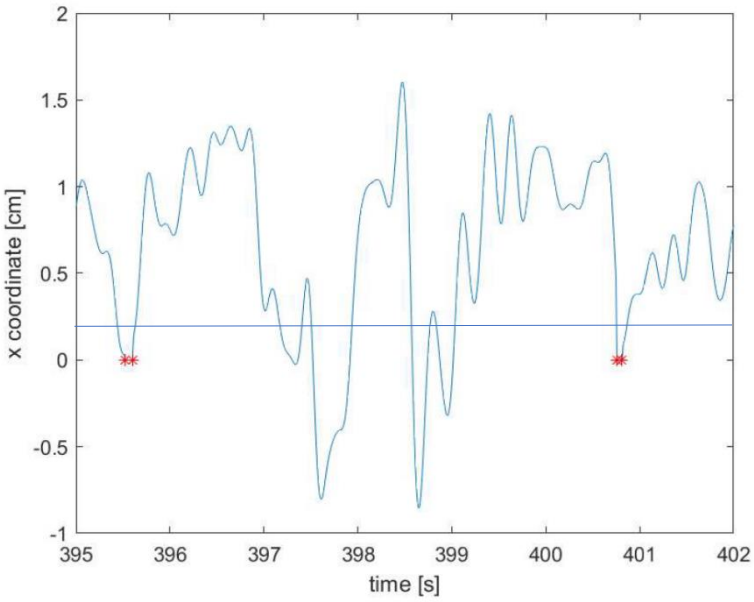

Figure 17: Example of blink identification algorithm applied to filtered x coordinate

An example of blink identification algorithm applied to the x coordinate is shown in Figure 17. The blinks can be identified by the red stars placed in correspondence of the beginning and ending of the blink period. Here, it is possible to see that the subject has blinked three times in a total of seven seconds (from second 395 to second 402 of the acquisition).

2.4.5. PUPIL DILATION

Raw pupil dilation data can not be directly given as input to the feature extraction process but it needs a pre-processing step. In fact, although eye-tracking is a robust data, pupil dilation data can

also contain samples that are purely the result of noise or artifacts and so carry no useful information. Therefore, pupil dilation data needs a filtering pipeline able to identify and remove this kind of artefacts.

The pipeline that is going to be described in this subchapter and that is implemented in this study is described by M. E. Kret in [40]. In particular, the author provides together with the explanation in the paper the already implemented code that is used in this study. The types of artifacts that are mainly present in pupil size samples are dilation speed outliers and edge artifacts, trend-line deviation outliers and temporally isolated samples. Moreover, those samples that are characterised by a pupil diameter smaller than or greater to specific boundaries are rejected (e.g. a pupil diameter value smaller than 1.5 mm and greater than 9 mm). In Figure 18 it is possible to see some examples of the pupil diameter artifacts previously mentioned.

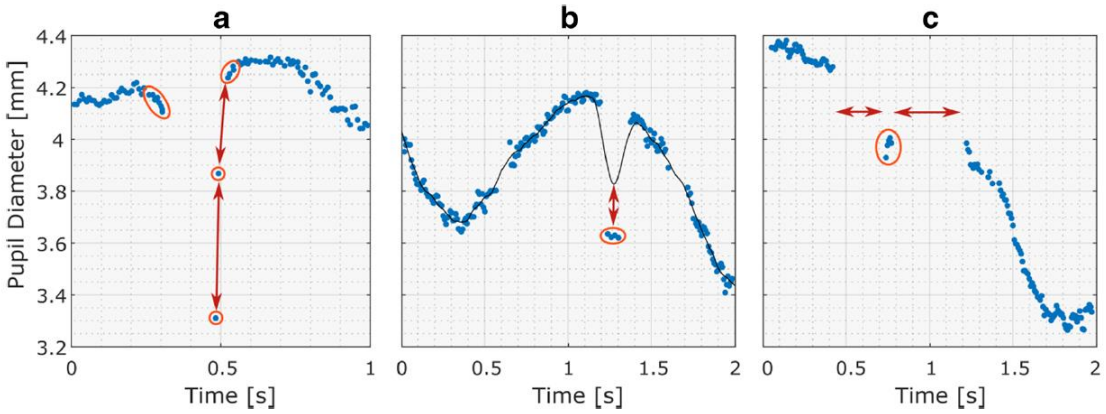

Figure 18 Raw pupil diameter data showing the different kind of artifacts; a) dilation speed outliers and edge artifacts b) trend-line deviation outliers (trend-line shown with a solid black line) c) temporally isolated samples [40]

Dilation speed outliers are those samples that present a larger pupil diameter variation with respect to the physiological one. However, the changes between samples exclusively due to changes in the pupil size are generally less than those resulting from artefacts such as blinks. In order to recognise these outliers, it is possible to compute the associated velocity per each one of the pupil diameter samples and then compute the median absolute deviation (MAD). MAD is a robust and outlier resilient data dispersion metric and can be computed as shown in (2). Then the next step consists in computing the threshold by multiplying the MAD by a constant n and summing to it the median dilation speed as shown in (3). For the best results, the parameters of the filtering approach such as n in equation (3) should be chosen empirically so that they best fit a particular dataset since no “one size fits all” set of rejection criteria exists.

$$MAD = median(|d'[i] - median(d')|) \qquad (2)$$

$$Threshold = median(d') + n * MAD \quad (3)$$

In these equations d'[i] represents the dilation speed associated to the i^{th} sample. Samples with dilation speed higher than the threshold can be considered as dilation speed outliers. There may be also other artefacts around gaps in the data, especially if these gaps are the result of eye blinks, which are called edge artifacts and which may cause pupil size underestimation due to eyelid occlusion. To get rid of these artifacts around gaps it is possible reject samples within 50 ms of gaps, with gaps being defined as contiguous missing data sections larger than 75 ms. For what concerns trend-line deviation outliers, they are small groups of clearly invalid samples that are clustered together and that can be identified by their strong departure from the signal's trend line. This trend line can be generated by interpolating and smoothing the data that remain after the previous filtering steps. Outliers in absolute trend-line deviations can then be identified and rejected in a similar manner to dilation speed outliers by feeding the absolute trend-line deviations into equations (2) and (3). Subsequently, a new trend line can be generated using the remaining samples, and the outlier detection process can be repeated on all samples considered in the first deviation filter step. This multi-step approach allows for the reintroduction of valid samples that were previously rejected due to the invalid samples that made the trend line diverge from the real one. Raw pupil size samples that are characterised by high sparsity must be rejected too. Indeed, a proper pupil size signal is fairly solid, with continuous gaps during blinks and look-away moments. Secluded samples are likely to be the result of noise or a momentary eye-tracker glitch, such as erroneous pupil detection during shut eyes. A sparsity filter is needed, it can be built as a filter that splits the pupil size signal at the samples that border a gap larger than a first criterion and then rejects the resulting sections that are smaller than a second criterion. Setting these criteria at 40 and 50 ms, respectively, appears to adequately rid the raw eye-tracking data of invalid secluded samples [40].

After the initial pre-processing for the artifact removal, the linear interpolation of the pupil diameters is done to re-construct the lost samples and then the mean pupil diameteris computed by averaging them. Now, two different types of steps are performed depending on the features that are going to be extracted successively in the feature extraction phase: the pupil dilation normalisation and the estimation of the power spectral density associated to the pupil dilation signal. There are different possible ways to normalise the pupil diameter, but in this study it has been chosen to exploit the min-max normalisation. This kind of normalisation can be described by equation (4):

$$PD_{scaled} = \frac{PD - min(PD)}{max(PD) - min(PD)} \quad (4)$$

In this equation PD refers to all the samples in time of the pupil diameter in a specific video section while PD_{scaled} is the normalised pupil diameter in the same section. In Figure 19 it is possible to see an example of pupil diameter normalisation applied to one video section in the data set.

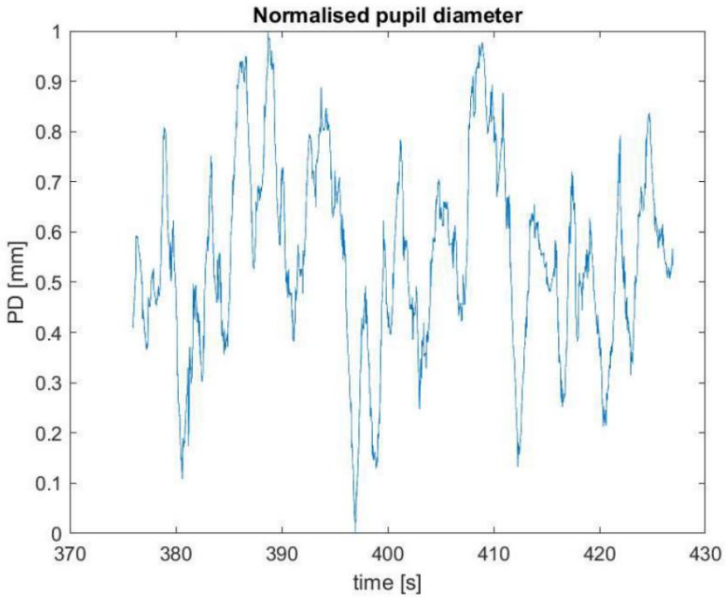

Figure 19 Example of pupil diameter normalisation applied to one section of the entire signal

The power spectral density can be extracted directly from the pre-processed pupil diameter after removing the mean to discard the high power zero frequency content (DC offset) and after an initial filtration phase achieved in order to remove the possible high frequency noise. A low-pass Butterworth filter of the 4^{th} order with cut-off frequency at 5 Hz is applied. For the estimation of the PSD, the so-called Welch method has been used. For the signal segmentation necessary to achieve the Welch method a Hamming window has been used whose length has been set at 20 times the sampling frequency of the signal which in our case is equal to 90 Hz.

2.4.6. DISCARDED DATA

Before starting the feature extraction phase, it is extremely important to discard the data that can lead to errors in the classification process. In particular, in this study it has been decided to discard those sections (video or resting sessions) in which an unacceptable part of the data is missing. The missing data can be due to two different factors: the eye-tracker may have had some technical issues that can have avoided it from properly acquiring the data, or the subject may have closed his/her eyes for a prolonged period of time. The discarding criteria chosen is this one, the missing samples of each section can not be higher than 30% of the total samples present in the acquisition. After the acquisition of all the data coming from all the subjects who took part in the experiments, it has been observed that a total of 33 sections must be discarded following the criteria forementioned. All the 33 discarded sections belong to the resting phase, it can be explained by the

fact that some of the subjects had the tendency to close their eyes during this phase. Moreover, it has also been noticed that all the acquisitions coming from one of the subjects must be discarded. In this case the data was corrupted and unusable due to some technical issues with the eye-tracker.

2.4.7. FEATURE EXTRACTION

The successive step to the signal processing phase is the extraction from the pre-processed signals of the features that are going to be given as input to the machine learning models. Feature extraction consists in transforming the already pre-processed data into numerical features that preserve the information in the original data set. Manually feature extraction has been used in this study, meaning that the identification and description of the features has been manually achieved choosing the features that are relevant for the emotion recognition problem. In this subchapter the feature extraction process is described and all the extracted features are listed. The features extraction process is summarised in a schematical way in Figure 20.

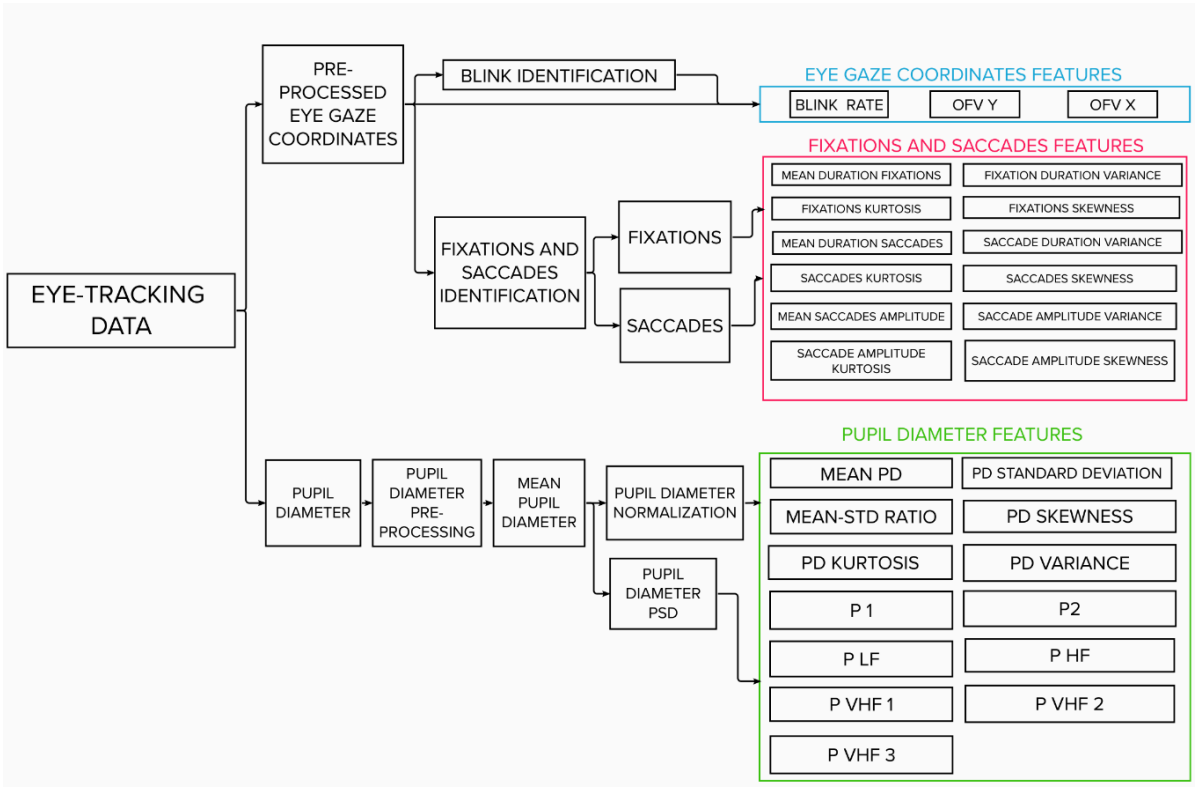

Figure 20 Block scheme of the feature extraction process starting from eye-tracking raw data; OFV x and OFV y stand for x and y coordinate of the overall fixation vector respectively, PD stands for pupil dilation, P1 is the power of pupil dilation in [0,0.45] Hz band, P2 is the power of pupil dilation in [0.45,5] Hz band P LF is power of pupil dilation in [0.04,015] Hz band P HF is the power of pupil dilation in [0.15,0.45] Hz band, P VHF 1 is the power of pupil dilation in [0.45,1] Hz band, P VHF 2 is the power of pupil dilation in [1,2.5] Hz band, P VHF 3 is the power of pupil dilation in [2.5,5] Hz band

The features that have been extracted in this study can be grouped in three different categories: features extracted from the eye gaze coordinates, features extracted from fixations and saccades

and features extracted from the pupil diameter signals. In Figure 20 the entire features extraction scheme starting from the raw eye-tracking data is shown.

Gaze coordinate features:

Regarding the eye gaze coordinates features, the features that can be extracted are the blink rate and the x and y coordinates of the overall fixation vector (OFV). Once we have identified the blinks' intervals it is possible to determine the estimated total number of blinks in the selected interval of time and then to determine the blink rate that can be defined as in equation (5):

$$Blink\ rate = \frac{Total\ number\ of\ blinks}{Period\ of\ time} \quad (5)$$

Besides, it is possible to extract the coordinates of the overall fixation vector (OFV) that is introduced in [33]. This feature considers the number, position and duration of fixations and can be described by equation (6):

$$\boldsymbol{OFV} = \sum_{i=1}^{N} t_i * \boldsymbol{v}_i \quad (6)$$

Where the time instant t_i is equal to the instant of time in which the fixation point is placed and $\boldsymbol{V}_i$ is a vector with an origin in the center of the screen (X_c,Y_c), and the end at the fixation point (X_i,Y_i) as described in equation (7). Note that with the subscript i the i^{th} fixation point is selected and that in our case the coordinates of the center of the screen corresponds to the point (0,0) as shown in equation (8).

$$\boldsymbol{v}_i = (X_i - X_c, Y_i - Y_c) \quad (7)$$

$$\begin{cases} X_c = 0 \\ Y_c = 0 \end{cases} \quad (8)$$

Fixations and saccades features:

On the other hand, the features extracted from the fixations and saccades, consists of a total of twelve which are: Mean duration of the fixations, Variance of the fixations' duration, Fixations kurtosis, Fixations skewness, Mean duration of the saccades, Variance of the saccades duration, Saccades kurtosis, Saccades skewness, Mean saccades amplitude, Variance of the saccades amplitude, Saccades amplitude kurtosis, Saccades amplitude skewness.

The average of the duration of fixations, the variance of the duration of fixations, the kurtosis of fixation points and the skewness of fixation points can be extracted once the points belonging to a fixation are identified. The average of the duration of fixations and the variance of the duration of fixations can be computed once the duration of the fixations is determined per each one of the

identified fixations and they correspond to the mean value and the variance of the duration of fixations. The fixation kurtosis and the fixation skewness can be determined directly from the fixation points and they are equal to the kurtosis and skewness of the duration of fixations. In particular, Kurtosis and Skewness are two measures that can be used to describe the shape of a data distribution. The kurtosis of a probability distribution is a scaled version of the fourth moment of the distribution and this number is related to the tails of the distribution. Hence, higher kurtosis' values correspond to greater extremity of deviations, also known as outliers. If the kurtosis of a data distribution is close to zero, it means that it approximately follows a gaussian probability density distribution, while the kurtosis of a non-Gaussian data distribution can be higher or lower than zero. In the case that the kurtosis is higher than zero, the data distribution is said to be super-gaussian, while if it is lower than zero it is said to be sub-gaussian. On the other hand, skewness is a measure of the asymmetry of the probability distribution around its mean. The skewness of a data distribution is negative if the left tail is longer and in this case the data distribution is said to be left-skewed, while it is positive the right tale is longer. In the latter case the data distribution is said to be right skewed. The kurtosis and skewness of a data distribution defined by a set of N samples x_i displaced in a vector X are computed by the equation (9) and (10) respectively:

$$Kurtosis(X) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - mean(X))^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - mean(X))^2)^2} \quad (9)$$

$$Skewness(X) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - mean(X))^3}{(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - mean(X))^2})^3} \quad (10)$$

The remaining features are relative to the saccades point. In particular, the average duration of saccades and the variance of the duration of saccades can be extracted once the durations of each saccade are computed and they correspond to the mean value and the variance of the saccades' duration. The saccade kurtosis and the saccade skewness are equal to the kurtosis and skewness of the duration of saccades while the average saccade amplitude, the variance of the saccade amplitudes, the kurtosis of the saccade amplitudes and the skewness of the saccade amplitudes are relative to the amplitude of the saccades. In this study it has been chosen to approximate the saccades' amplitude with the Euclidean norm between fixation points as shown in equation (11).

$$Saccade's\ amplitude = \sqrt[2]{(X_{i+1} - X_i)^2 - (Y_{i+1} - Y_i)^2} \quad (11)$$

The subscript i+1 indicate the first sample of the successive fixation, while the subscript i indicates the last sample of the previous fixation. Once the saccades' amplitudes have been computed per each one of the saccades, we can extract the mean value, the variance, the skewness and the kurtosis of the amplitudes of the saccades.

Pupil diameter features:

Finally, let us consider the last group of features called pupil diameter features. These features are all the features that can be extracted from the pupil diameter measure and they are: Mean pupil diameter, Standard deviation of the pupil diameter, Mean-standard deviation ratio, Variance of pupil diameter, Skewness of pupil's diameter, Kurtosis of pupil diameter, P1, P2, P LF, P HF, P VHF 1, P VHF 2, P VHF 3

All these features must be extracted after the initial processing of the pupil diameter. In particular, as an output from the eye-tracker, we have the raw pupil diameter of the left eye and the one of the right eye. Once the pre-processing of the pupil dilation signal described in chapter 2.4.5 is performed, the feature extraction process can be applied to extract two kinds of features: the ones that can be computed after the pupil diameter normalisation, and the ones that can be extracted after the estimation of the pupil diameter power spectral density. The pupil diameter normalisation is needed in order to be able to compare the pupil diameter of the different video and rest sections. In fact, as it has already been discussed in chapter 1.3.1, the pupil diameter's value is not only correlated to the physiological response of the ANS of the participant to the selected stimulus, but it is also correlated to the level of luminance of the stimulus itself. Even though the videos have been realised in the most standardised way possible by balancing the values of luminance as close as possible, it is possible that a slight difference in luminance level is present. This effect is especially present if we compare the resting phase with the video phase since the stimulus presented in the resting phase is a black screen with a white cross in the center of the screen. This stimulus is then characterised by a low value of luminance if compared to the video stimuli, implicating an average higher value in the pupil diameter.

The features extracted from the normalised pupil diameter are the following: the mean pupil diameter ,the standard deviation of the pupil diameter, the mean-standard deviation ratio, the pupil diameter's variance, the pupil's diameter's skewness and the pupil diameter's kurtosis. In particular, the mean-standard deviation ratio (or coefficient of variation) is the ratio between the mean and the standard deviation of the pupil dilation values. For what concerns then the remaining features they are all those features that can be extracted from the pupil dilation power spectral density (PSD). In particular, each one of these features corresponds to the power contained by the pupil diameter signal in specific frequency bands that have been suggested by F. Onorati et al. in [41]. P1 and P2 are the spectral power computed in two macro frequency bands. In particular P1 corresponds to the spectral power of the pupil dilation signal in the frequency band that goes from 0 up to 0.45 Hz, P2 is the spectral power in the frequency band that goes from 0.45 up to 5 Hz. Exploiting these two frequency bands it is possible to reconstruct the tonic and phasic components of the pupil dilation signal as shown in Figure 21. In the low frequency band up to 0.45 Hz, it is possible to define two further bands: P LF is the spectral power in the low frequency band that goes from 0.04 to 0.15 Hz, P HF is the spectral power in the low frequency band that goes from 0.15 up to 0.45 Hz. Finally, to explore the frequency contributions from 0.45 Hz up to 5 Hz, referred to as very high frequency (VHF) band, it is possible to compute the spectral power in 3 additional frequency bands: P VHF1 is the spectral power contained in the frequency band that goes from 0.45 up to 1 Hz, P VHF2 is the spectral power contained in the frequency band that goes from 1 up to 2.5 Hz while P VHF3 is the spectral power contained in the frequency band that goes from 2.5 up

to 5 Hz. Note that the above-mentioned frequency bands are caught between 0 and 5 Hz, in fact the spectral content for the pupil dilation signal is up to 4-5 Hz [41]. Moreover, for the estimation of the frequency contribution in the VHF band, it has been chosen to apply a high-pass Butterworth filter of the 4th order with cut-off frequency at 0.2 Hz to eliminate the high-power low frequency content.

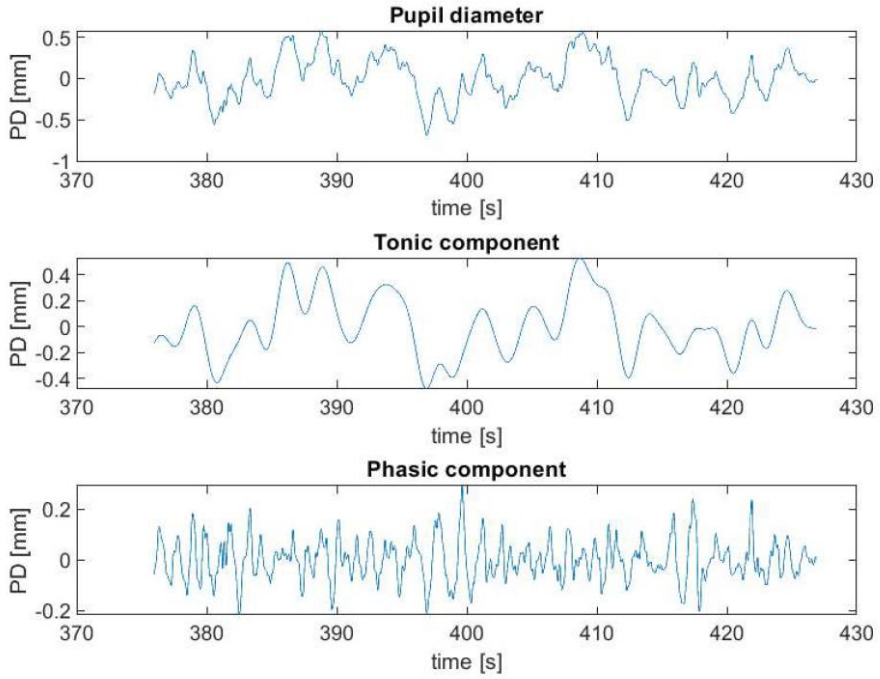

Figure 21 Tonic and phasic components extracted from one eye's pupil diameter in a section of the signal; 1) Pre-processed pupil diameter with removed mean 2) Tonic component 3) Phasic component

Outliers removal:

It must be mentioned that a feature optimisation step can be achieved before the statistical analysis. Sometimes a dataset may contain values that are outside the range of what is expected and unlike the other data. These are called outliers. Machine learning modelling can be improved by removing these outlier values. Therefore, outliers have been identified and later removed using the Median Absolute Deviation (MAD). This method consists in considering as an outlier each value of the data set which is higher in absolute value to three scaled MAD from the median. A MAD is defined as specified in equation (12) and in (13).

$$MAD = c * median(|A - median(A)|) \quad (12)$$

Where:

$$c = \frac{-1}{\sqrt{2} * erfcinv(\frac{3}{2})} \quad (13)$$

The input A is a vector with all the values of a specific feature and the function “erfcinv” indicates the inverse complementary error function.

2.4.8. FEATURE SELECTION

Feature selection is the process of reducing the number of features used as input when developing a predictive model. It is desirable to reduce the number of input variables to reduce the computational cost of modelling, as a large number of variables can slow the development and training of models and require a large amount of system memory. Additionally, a feature selection is also needed in order to improve the performance of the model, in fact, if input variables that are not relevant to the target variable are included, the performance of the model may be worse.

In this study, statistical-based feature selection methods are used. It involves evaluating the relationship between each feature using statistics and selecting those input variables features that are most informative and non-redundant, facilitating the learning and generalization of the classifiers. The choice of statistical tests depends on the data type of the input and on the number of the output classes. In our case, all the features are numerical. For that reason, it has been decided to use a statistical hypothesis test such as the two samples t-test in the case of binary classification since this kind of test takes as input only two variables, and the so-called one-way ANOVA (Analysis of Variance) in the case of multi-label classification since this kind of test can take as input more than two variables. In particular, if a binary classification is needed, there are two groups for the same feature, while if a multi-label classification is needed, there are a number of groups for the same feature equal to the number of defined classes (e.g. blink rate feature is selected, then it is possible to define a group of values for the blink rate per each one of the classes selected). It is important to consider that both the two samples t-test and the one-way ANOVA can be applied to specific data only if the analysed data follows a normal distribution. For that reason, a normality *test* must be used to determine whether sample data has been drawn from a normally distributed population. In this study it has been decided to use a one-sample Kolmogorov-Smirnov test. If the normality assumption is not fulfilled by the data the non-parametric tests are needed. The non-parametric tests are “distribution-free” and, as such, they can be used for the testing of non-normal variables. In this study it has been chosen to use the Kruskal Wallis Test in the case in which the one-way ANOVA must be replaced, and the Wilcoxon rank sum test in the case in which the two-samples t-test must be replaced. The main difference between the Kruskal Wallis test and the ANOVA lies in the fact that the first one is used in order to check if there is a significant difference between the medians of the input features, while the second one is used in order to check if there is a significant difference between the means of the input features. In the same way, the Wilcoxon rank sum test is used in order to check if there is a significant difference between the medians of the input features, while the two-samples t-test is used in order to check if there is a significant difference between the means of the input features. In Figure 22 the block scheme of the statistical analysis process explained in detail is shown. In the case of multi-label classification, a total of three groups is defined as a matter of example.

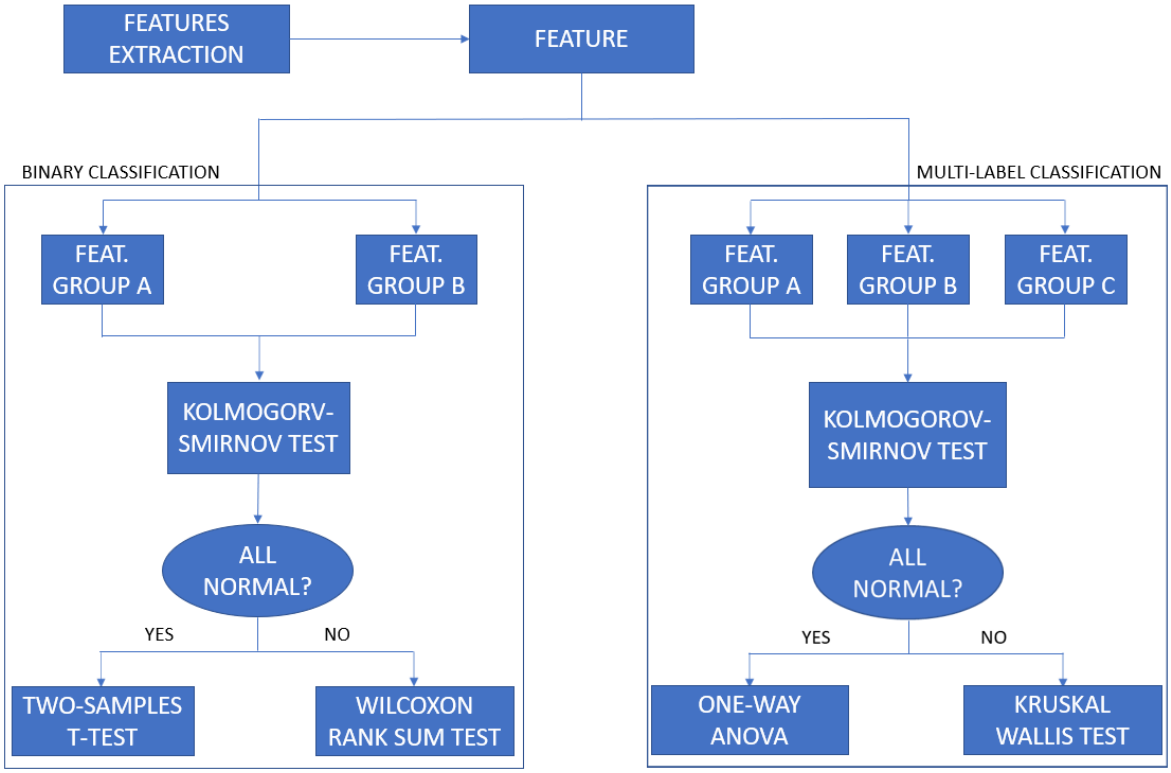

Figure 22 Block scheme of the statistical-based feature selection used

2.5. CLASSIFICATION

The classification algorithms can be categorised between supervised and unsupervised learning. In particular, in supervised learning, the input data is already labelled by the user (supervisor), on the other hand, in unsupervised learning the user gives as input unlabelled data. In this study a Supervised Learning approach has been chosen. In classification problems that use machine learning supervised algorithms, a model is trained with a dataset which contains the selected features together with the class which each sample is part of. Once the model is trained, its performance must be validated with a test dataset. In this chapter it will be discussed the description of the defined classes and the machine learning models that have been used.

2.5.1. CLASSES

As it has been previously discussed in , three different classification problems has been defined in this study, the first one is the classification between the audio-visual stimulation of the subject (video phase) and the resting phase, the second one is the classification of the type of video provided to the subject (video of a known person or video of an unknown person) while the third one consists in the classification of the level of arousal and valence of the emotional response of the subject during the video display.

The classification between the video phase (or audio-visual stimulation) and the resting phase requires the implementation of a binary classifier. In particular, the classes that are defined for this classification are “rest” for the resting phase and “video” for the video phase.

Regarding the classification of the type of video provided to the subject, a binary classification has been used. The first class defined is “known” in the case the video displayed is of a referred person and it is labelled as type A videos whereas the second class is “unknown” in the case the video displayed is of an unknown person and it is labelled as type B.

For what concerns the classification of the level of arousal and valence, it has been decided to divide the valence-arousal plane, presented in the circumplex model of emotions defined by James Russel and previously explained in chapter 1.2.1, into four parts:

1. HIGH AROUSAL – HIGH VALENCE
2. HIGH AROUSAL – LOW VALENCE
3. LOW AROUSAL – LOW VALENCE
4. LOW AROUSAL – HIGH VALENCE

Thus, a total of four classes are built and a multi-label classification must be achieved. In particular, high arousal or valence is defined as those evaluations in which their value is higher or equal to 5, while low means that their value is lower than 5.

2.5.2. MACHINE LEARNING MODELS

In order to find the machine learning model that best fits the problem faced in this study a series of different machine learning models have been trained and validated. In particular, the classification learner app available in “Statistics and Machine learning toolbox” of MATLAB has been used. The Classification Learner app trains models to classify data. Using this app, it is possible to explore supervised machine learning using various classifiers and it is possible to explore the data, select features, specify validation schemes, train models, and assess results.

As it has been discussed previously in chapter 1.3.4, SVM classifiers are the most used in emotion recognition so they are the ones chosen to explore in this study. Moreover, the LDA and k-NN classifiers are also used in this study. In the followings, these three models will be briefly explained.

Linear discriminant analysis (LDA) is machine learning model used to classify pattern between two classes; even if it can be extended to classify multiple patterns as stated by Vaibhaw et al. in [42]. LDA is based on the assumption that all the selected classes are linearly separable and its purpose is the identification of several hyperplanes in the feature space exploited to distinguish the classes. Then the LDA projects the data onto this hyperplane in such a way as to maximize the separation of the two categories. Two criteria must be considered simultaneously in order to create the hyperplanes:

- Maximizing the distance between the means of two classes;
- Minimizing the variation between each category

The k-NN classifier is based on the principle that in most of the cases similar features are displaced in close points in the feature space [42]. k-NN is based on a distance function that measures similarity between two data samples. Usually the standard Euclidean distance between two data samples x and y is used as stated by L. Jiang in [43]. Given a data sample x, k-NN assigns the most common class of k nearest neighbours of x to x. Depending on the value of the parameter k we can have three different models: fine k-NN, medium k-NN and coarse k-NN. In the classification learner app a value of k = 1, k = 10 and k = 100 are respectively used in fine k-NN, medium k-NN and coarse k-NN. All these types of k-NN models are trained in this study.

As stated by W. S. Noble in [44], SVM aims to create a decision boundary between two classes that for their prediction starting from a feature vector. As explained by C. Cortes et al. in [45], this decision boundary is known as hyperplane and its orientation is found in order to place it as distant as possible data points from each of the classes. The objective of training an SVM model is to find the parameters of the hyperplane that best separate the data. The SVM algorithm was originally proposed to build a linear classifier, but, as shown by S. Huang et al. in [46] , an alternative use for SVM is the kernel method, which gives the possibility to model higher dimensional non-linear models. In a non-linear problem, a kernel function could be used to add additional dimensions to the raw data and thus make it a linear problem in the resulting higher dimensional space. The choice of kernel function greatly affects the performance of an SVM model and there is no way to figure out which kernel best fits a classification problem. Therefore, the optimal kernel function can only be chosen through different trials from a fixed set of kernels by using cross-validation. Even though SVM does not support multiclass classification, the multiclass problem can be achieved through multiple binary classifications. In this way the SVM models can be used also in multiclass problems. Depending on the type of used Kernel function, it is possible to build various SVM models. In particular, in this study the following models are exploited: linear SVM, quadratic SVM, cubic SVM, fine gaussian SVM, medium gaussian SVM and coarse gaussian SVM.

2.5.3. VALIDATION AND EVALUATION METRICS

In machine learning, model validation refers to as the procedure where a trained model is assessed with a testing data set. The testing data set is a part of the overall dataset from which also the training dataset is inferred. The principal reason for utilizing the testing dataset is to test the speculation capacity of a prepared model. It is possible to choose between different validation techniques but one of the most widely used is the cross-validation. As stated by F. Maleki in [47], Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In typical cross-validation, the training and validation sets must cross over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-validation is k-fold cross-validation. K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called folds. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test

set. This approach is a very popular cross-validation approach because the output is less biased than other methods. In this study it has been decided to use a k-fold validation with k equal to five. This method gives a good estimate of the predictive accuracy of the final model trained with all the data. It requires multiple fits but makes efficient use of all the data, so it is recommended for small data sets [47].

Once the models have been trained and validated, confusion matrices from the validated results can be obtained. From these matrices, several parameters can be extracted. If a binary classification is considered, the two generical classes can be called positive (P) class and negative (N) class. The predictions got as an output of the classification process are then called true positive, if the predicted class is positive and the actual class is positive, true negative, if the predicted class is negative and the actual class is negative, false positive if the predicted class is positive but the actual class is negative and false negative if the predicted class is negative but the actual class is positive. In Figure 23 a general confusion matrix scheme is shown.

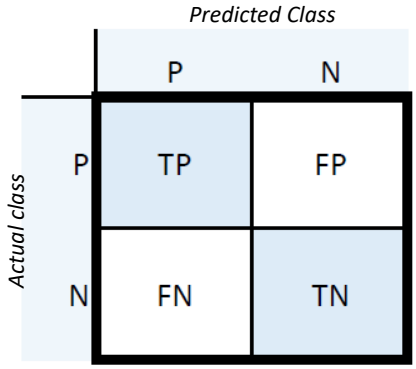

Figure 23 Confusion matrix example; P (positive), N (negative), TP (true positives), TN (true negatives), FP(false positives), FN (false negatives)

Evaluation metrics such as accuracy can be extracted from these parameters. However, as shown by A. Luque in [48], the accuracy in validation alone can be misleading if you have an unequal number of observations in each class. For this reason, other evaluation metrics are needed to evaluate the performance of the models. In this study, the metrics used are accuracy (A), precision (P), recall (R), and F1-score (F1).

- **Accuracy (A)**

Accuracy is a metric that quantifies the ratio between the correctly classified predictions and all the predictions carried out. In particular, it is equal to the sum of the true positives and true negatives divided by the total number of positive and negative predicted samples [48]. The formula describing precision is shown in equation (14):

$$Accuracy = \frac{True\ positives + true\ negatives}{Total\ number\ of\ predictions} \quad (14)$$

- **Precision (P)**

Precision is a metric that quantifies the number of correct positive predictions made. It is calculated as the ratio of correctly predicted positive examples divided by the total number of positive examples that were predicted [48]. The formula describing precision is shown in equation (15):

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (15)$$

The result is a value between 0.0 for no precision and 1.0 for full or perfect precision, this is true also for recall and F1-score. On the other hand, recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

- **Recall (R)**

Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions, in fact, as it is shown in (16), it is calculated as the number of true positives divided by the total number of true positives and false negatives [48].

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (16)$$

- **F1-score (F1)**

For what concerns the F1-score it is a measure of a test's accuracy and it is calculated from the precision and recall of the test since it is the harmonic mean of precision and recall [48]. The formula describing the F1-score is shown in equation (17).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (17)$$

3. RESULTS

In this chapter the results drawn from this study are going to be shown in the following order: 1) the results of the questionnaires provided to the participants; 2) the results of the classification between stimulation and non-stimulation periods; 3) the results of the classification between stimuli featuring acquaintances or strangers and 4) the results of the classification between stimuli of specific self-reported pleasure (valence) and excitation (arousal).

3.1. FILM IEQ AND AD-HOC QUESTIONNAIRES

As shown in chapter 2.2.3, the Film IEQ questionnaire is a questionnaire made up of 24 questions which are used to assess the video viewing experience and the degree of immersion of the subject during the delivery of the audio-visual stimulation. As a result, the average overall immersion score (mean value of the overall immersion scores for each one of the subjects) is equal to 5.41±0.58 out of 7, stating a high degree of immersion of the subjects.

For what concerns the ad-hoc questionnaire and as it has already been explained in chapter 2.2.3, it is a questionnaire formulated to check the level of acknowledgement about the experimentation the subject had before taking part to it and also to check what degree of influence the use of all the instrumentation used has had. The average score per each one of the questions are 2±1.35 out of 7 for question related to the previously acknowledgment of the study, 4.93±1.92 out of 7 and 4.26±1.86 out of 7 for the questions related to the level of influence of the instrumentation in the study. These values show that the subjects were almost unaware of the topic of the experimentation as it was expected and wanted. Then, the average score of the answers of the second and third questions shows that the instrumentation had a medium-high influence on the concentration of the subjects due to its discomfort.

AD HOC QUESTIONNAIRE		
1)	Did you have idea of what the experiment consisted in before taking part in it?	2±1.35
2)	Was the instrumentation uncomfortable during the experiment?	4.93±1.92
3)	Do you think that your level of concentration during the experiment was influenced by instrumentation?	4.26±1.86

Table 4 Mean and standard deviation of the results of the ad hoc questionnaire per each one of the questions

3.2. STIMULI AND RESTING PHASE CLASSIFICATION

In this chapter the results of the comparison between the stimulation and non-stimulation periods are going to be shown. In particular, we will examine the values of the extracted features explained in chapter 2.4.7 and the results of the statistical-based feature selection. Finally, the results of the machine learning classification models will be shown.

3.2.1. FEATURE EXTRACTION

The pre-processed data is divided in the two groups of interest: stimuli phase, which comprises videos of both the known and unknown person, and resting phase, which comprises the periods where the subjects were exposed to the black screen with a white cross.

As it has been explained in chapter 2.4.7, a total of 28 features are obtained. The mean and standard deviation values of the features are shown in Table 5. Once the features have been extracted, statistical-based feature selection is done to discard all features that are not relevant. The first step of the statistical analysis is checking the normality of the extracted features through the Kolmogorov-Smirnov test. The results of this test showed that all features do not follow a normal distribution, with p-values <0.001, much lower than the threshold of significance of 0.05. Given these results, Wilcoxon rank sum test is chosen to test if there is a statistically significant difference between the resting and video phase in each feature. As it is shown in Table 5, a total of 18 features out of 28 present a significant difference. These features are the average duration of fixations ($p = 1.50x10^{-6}$), the variance of fixation duration ($p = 1.67x10^{-9}$), the y coordinate of the overall fixation vector ($p = 2.25x10^{-21}$), the kurtosis of saccade duration ($p = 3.14x10^{-10}$), the skewness of saccade duration ($p = 3.06x10^{-5}$), the average amplitude of saccades ($p = 7.87x10^{-8}$), the variance of saccade amplitude ($p = 5.07x10^{-8}$), the skewness of saccade amplitude ($p = 2.35x10^{-5}$), the mean-standard deviation ratio ($p = 2.27x10^{-12}$), the pupil dilation skewness ($p = 2.77x10^{-20}$), the pupil dilation kurtosis($p = 1.71x10^{-4}$) and all the features relative to the power of the pupil dilation in different frequency bands: P1 is the power in the frequency band [0,0.45] Hz ($p = 0.02$), P2 is the power in the frequency band [0.45,5] ($p =1.04x10^{-20}$), Hz P LF is the power in the frequency band [0.04,0.15] Hz ($p = 1.07x10^{-4}$), P HF is the power in the frequency band [0.15,0.45] Hz ($p = 4.67x10^{-4}$), P VHF1 is the power in the frequency band [0.45,1] Hz ($p = 2.18x10^{-17}$), P VHF2 is the power in the frequency band [1,2.5] Hz ($p = 3.24x10^{-29}$), P VHF3 is the power in the frequency band [2.5,5] Hz ($p = 3.24x10^{-29}$). In particular, the p values associated to all these selected features is smaller than 0.01 with the exception of the p value of P1.

Feature	Video Mean Std.	Rest Mean Std.	p values
Blink rate	20.20 ± 9.49	22.42 ± 9.6	0.13
Average duration of fixations**	**0.55 ± 0.17**	**0.7 ± 0.29**	**1.50 e-06**
Variance of fixation duration**	**0.24 ± 0.18**	**0.42 ± 0.32**	**1.67 e-09**
Kurtosis of fixation duration	8.40 ± 3.95	7.883 ± 3.6	0.31
Skewness of fixation duration	2.14 ± 0.73	2.09 ± 0.79	0.62
OFV x	-77.57 ± 754.61	-52.11 ± 674.47	0.25
OFV y**	**1.90 e+03 ± 1.67 e+03**	**0.17 e+03 ± 1.10 e+03**	**2.25 e-21**
Average duration of saccades	4.31 e-02 ± 1.71 e-02	0.04 ± 1.87 e-02	0.48
Variance of saccades duration	0.15 e-02 ± 0.11 e-02	0.15 e-02 ± 0.10 e-02	0.75
Kurtosis of saccade duration**	**3.80 ± 2.08**	**2.41 ± 0.92**	**3.14 e-10**
Skewness of saccade duration**	**1.17 ± 0.71**	**82 ± 0.75**	**3.06 e-05**
Average amplitude of saccades**	**0.99 ± 0.45**	**0.7 ± 0.48**	**7.87 e-08**
Variance of saccade amplitude**	**0.65 ± 0.56**	**0.34 ± 0.33**	**5.07 e-08**
Kurtosis of saccade amplitude**	**2.81 ± 1.08**	**2.3 ± 0.96**	**2.35 e-05**
Skewness of saccade amplitude	0.70 ± 0.6	0.58 ± 0.62	0.07
Std. of pupil dilation	0.18 ± 2,14 e-02	0.19 ± 3,09 e-02	0.16
Mean-Std. Ratio**	**2.77 ± 0.49**	**3.37 ± 76**	**2.27 e-12**
Pupil dilation skewness**	**-0.03 ± 0.43**	**-0.67 ± 0.61**	**2.77 e-20**
Pupil dilation kurtosis**	**2.96 ± 0.55**	**3.39 ± 0.96**	**1.71 e-04**
Variance of pupil dilation	3.32 e-02 ± 0,74 e-02	3.57 e-02 ± 1,14 e-02	0.13
P1*	**2.64 e-02 ± 1.53 e-02**	**3.68 e-02 ± 2.93 e-02**	**0.02**
P2**	**0.64 e-02 ± 0.33 e-02**	**0.31 e-02 ± 0.18 e-02**	**1.04 e-20**
P LF**	**1.33 e-02 ± 0.80 e-02**	**2.13 e-02 ± 1.68 e-02**	**1.07 e-04**
P HF**	**1.10 e-02 ± 0.70 e-02**	**0.83 e-02 ± 0.59 e-02**	**4.67 e-04**
P VHF1**	**0.47 e-02 ± 0.26 e-02**	**0.23 e-02 ± 0.15 e-02**	**2.18 e-17**
P VHF2**	**1.80 e-03 ± 9.58 e-04**	**0.67 e-03 ± 3.74 e-04**	**3.24 e-29**
P VHF3**	**1.62 e-04 ± 6.81 e-05**	**1.12 e-04 ± 5.57 e-05**	**4.69 e-12**
Average pupil dilation	0.50 ± 0.08	0.62 ± 0.08	/

* p<0.05, ** p<0.01

Table 5 Values of mean and standard deviation of all the extracted features for the stimuli phase (video) and resting phase (rest) and relative p value got as output of the Wilcoxon rank sum test. In bold text are highlighted the features that present a significant difference (p<0.05); OFV x and OFV y stand for x and y coordinates of the overall fixation vector, std. stands for standard deviation, , P1 is the power of pupil dilation in [0,0.45] Hz band, P2 is the power of pupil dilation in [0.45,5] Hz band P LF is power of pupil dilation in [0.04,015] Hz band P HF is the power of pupil dilation in [0.15,0.45] Hz band, P VHF 1 is the power of pupil dilation in [0.45,1] Hz band, P VHF 2 is the power of pupil dilation in [1,2.5] Hz band, P VHF 3 is the power of pupil dilation in [2.5,5] Hz band

The average pupil dilation feature does not have an associated p value in Table 5 since in the statistical analysis implemented on the features used in the classification between stimulation and non-stimulation periods, this feature was not considered. The reason behind this choice lies in the fact that a significant difference in the average pupil dilation in this case could be due a difference in luminance and not to a different physiological response of the subject. The features that presented a statistically significant difference were then normalized. The normalized mean value and standard deviation are shown in the Figure 25.

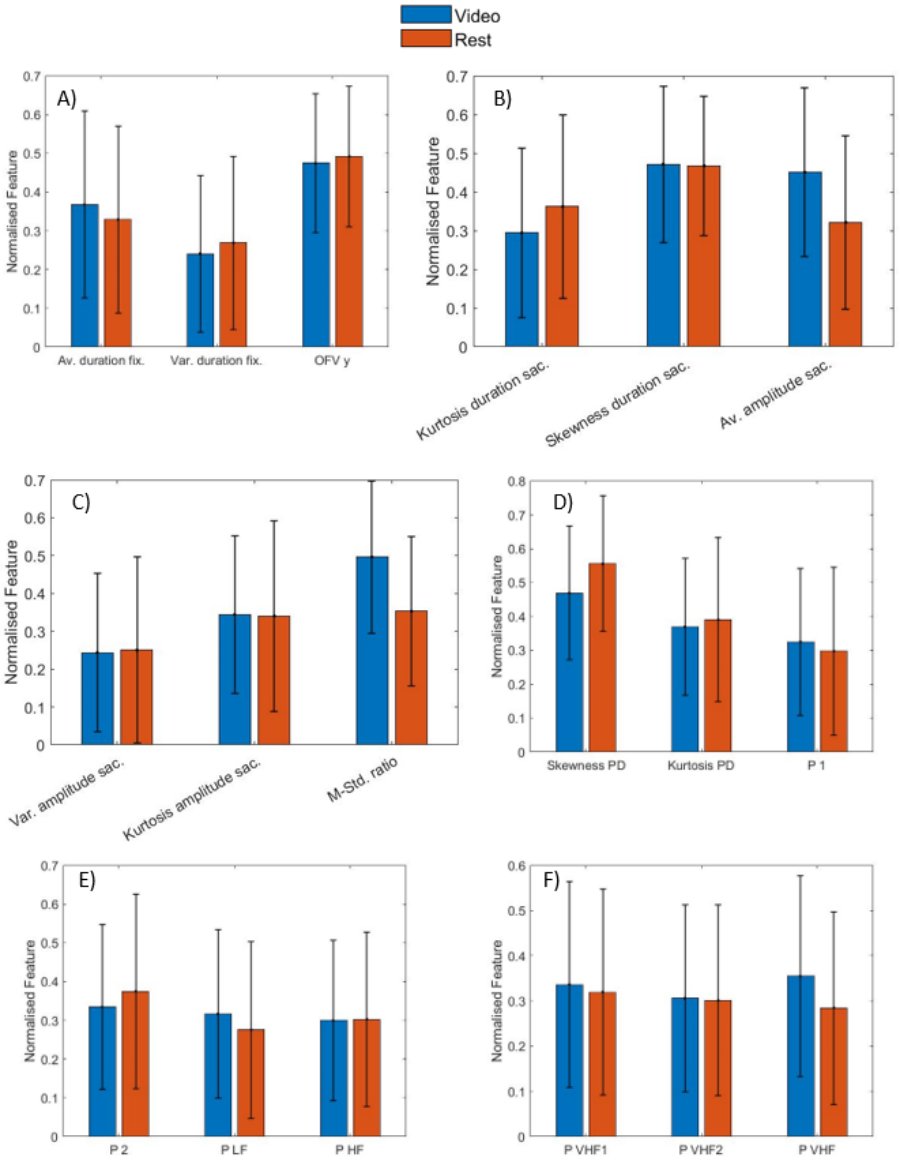

Figure 25 graphical representation of the values of all the features whose relative p value is lower than 0.05: A) Average duration of fixations (av. Duration fix.), variance of fixation duration (var. duration fix.), y coordinate of overall fixation vector (OFV y) B) Kurtosis of saccade duration (kurtosis duration sac.), skewness of saccade duration (skewness duration sac.), average amplitude of saccades (av. Amplitude sac.) C) Variance of saccade amplitude (var. amplitude sac.), kurtosis of saccade amplitude (kurtosis amplitude sac.), mean-standard deviation ratio of pupil dilation(m-std ratio) D) Pupil dilation skewness (skewness PD), pupil dilation kurtosis (kurtosis PD), power of pupil dilation in [0,0.45] Hz band (P1) E) power of pupil dilation in [0.45,5] Hz band (P2), power of pupil dilation in [0.04,015] Hz band (P LF), power of pupil dilation in [0.15,0.45] Hz band (P HF) F) power of pupil dilation in [0.45,1] Hz band (P VHF 1), power of pupil dilation in [1,2.5] Hz band (P VHF 2), power of pupil dilation in [2.5,5] Hz band (P VHF 3)

3.2.2. CLASSIFICATION

18 features were obtained after the feature extraction and selection phases. The successive step consists in the training and validation of the machine learning models. In this case, since we need to discern between two classes (resting phase and video phase), a binary classifier is needed. The machine learning models that are going to be exploited are 10: Linear Discriminant Analysis (LDA), linear Support Vector Machines (SVM), quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, coarse Gaussian SVM fine k-Nearest Neighbour (k-NN), medium k-NN and finally coarse k-NN. In Table 7 the accuracy of each one of the trained models obtained through the 5 fold cross-validation technique is shown.

Accuracy			
LDA	**92.70%**	Medium Gaussian SVM	**94.40%**
Linear SVM	92.70%	Coarse Gaussian SVM	92.40%
Quadratic SVM	94.10%	Fine k-NN	86.80%
Cubic SVM	93.10%	Medium k-NN	**92.70%**
Fine Gaussian SVM	55.80%	Coarse k-NN	86.80%

Table 7 Accuracy values for all the trained models; in bold the best accuracy values associated to each one of the model types (LDA, SVM, k-NN)

As it is possible to see in the previous table, most of the trained models present degree of accuracy higher than 90%. In particular, the accuracy of the LDA model is equal to 92.70%, the accuracy of the linear SVM model is equal to 92.70%, the accuracy of the quadratic SVM model is equal to 94.10%, the accuracy of the cubic SVM model is equal to 93.10%, the accuracy of the Fine Gaussian SVM model is equal to 55.80%, the accuracy of the Medium Gaussian SVM model is equal to 94.40%, the accuracy of the coarse Gaussian SVM model is equal to 92.40%, the accuracy of the fine k-NN model is equal to 86.80%, the accuracy of the medium k-NN model is equal to 92.7% while the accuracy of the coarse k-NN model is equal to 86.80%. The model characterised by the highest accuracy between all the trained models is the medium Gaussian SVM, while the model characterised by the lowest accuracy is the fine Gaussian SVM

Together with the accuracy a set of three performance metrics is computed per each one of the forementioned models: precision, recall and F1 score. In particular, these performance metrics are computed for both the classes. In Table 8 the values of precision, recall and F1 score of the trained models is shown.

	LDA			Linear SVM		
Classes	P	R	F1	P	R	S
Video	0.94	0.93	0.93	0.96	0.91	0.94
Rest	0.91	0.92	0.94	0.89	0.94	0.92
Mean	0.93	0.93	0.94	0.93	0.93	0.93

	Quadratic SVM			Cubic SVM		
Classes	P	R	F1	P	R	S
Video	0.96	0.94	0.95	0.95	0.93	0.94
Rest	0.92	0.95	0.93	0.92	0.93	0.92
Mean	0.94	0.95	0.94	0.94	0.93	0.93

	Fine Gaussian SVM			Medium Gaussian SVM		
Classes	P	R	F1	P	R	S
Video	1	0.56	0.71	0.98	0.93	0.95
Rest	0,74 e-02	1	1,47 e-02	0.9	0.97	0.93
Mean	0.50	0.78	0.36	0.94	0.95	0.94

	Coarse Gaussian SVM			Fine k-NN		
Classes	P	R	F1	P	R	S
Video	0.93	0.93	0.93	0.90	0.87	0.88
Rest	0.91	0.92	0.91	0.83	0.87	0.85
Mean	0.92	0.93	0.92	0.87	0.87	0.87

	Medium k-NN			Coarse k-NN		
Classes	P	R	F1	P	R	S
Video	0.98	0.9	0.94	0.95	0.84	0.88
Rest	0.87	0.97	0.91	0.77	0.92	0.84
Mean	0.93	0.94	0.93	0.86	0.88	0.86

Table 8 Values of the per class precision (P), recall (R), F1 score (F1) and mean value of precision, recall and F1 score associated to all the trained models

The LDA model presents an average F1 score (mean value between the two per class F1 score) equal to 0.94. The SVM model that present the highest average F1 score are the quadratic SVM (0.94) and the medium Gaussian SVM (0.94). The k-NN model that presents the highest average F1 score is the medium k-NN (0.93). The model that shows the lowest average F1 score is the fine Gaussian SVM.

In Figure 26, the confusion matrix of the method which present the highest accuracy is going to be shown. In particular, it is possible to see the confusion matrix of the medium Gaussian SVM.

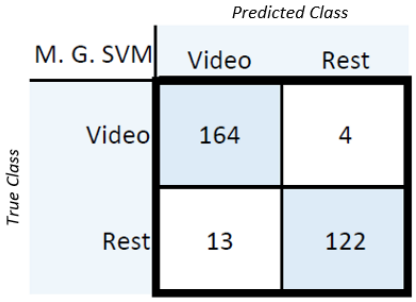

Figure 26 Confusion matrix relative to the trained model which present the highest accuracy, in this case the medium Gaussian support vector machine (M. G. SVM). The class video refers to the stimulation periods, while the class rest refers to the non-stimulation periods

1.1. ACQUAINTANCES AND STRANGERS CLASSIFICATION

In this chapter the results of the comparison between stimuli featuring acquaintances or strangers. In particular, we will examine the values of the extracted features explained in chapter 2.4.7 and

the results of the statistical-based feature selection. Finally, the results of the machine learning classification models will be shown.

1.1.1. FEATURE EXTRACTION

The pre-processed data is divided in the two groups of interest: stimuli featuring acquaintances, which comprises videos of the known person, and stimuli featuring strangers, which comprises videos of the unknown person.

A total of 28 features are obtained through the feature extraction as it has been explained in chapter 2.4.7. The mean and standard deviation values of the features are shown in Table 9.

Feature	A Mean ± Std.	B Mean ± Std.	p value
Blink rate	19.55 ± 8.72	20.84 ± 10.22	0.44
Average duration of fixations*	**0.53 ± 0.17**	**0.57 ± 0.16**	**0.03**
Variance of fixation duration**	**0.17 ± 0.13**	**0.31 ± 0.19**	**1.23 e-08**
Kurtosis of fixation duration*	**7.68 ± 3.36**	**9.12 ± 4.36**	**0.03**
Skewness of fixation duration*	**1.98 ± 0.68**	**2.29 ± 0.74**	**0.01**
OFV x	-89.64 ± 921.11	-65.5 ± 544.98	0.36
OFV y	1.73 e+03 ± 1.74 e+03	2.07 e+03 ± 1.57 e+03	0.14
Average duration of saccades	4.11 e-02 ± 1.54 e-02	4.52 e-02 ± 1.85 e-02	0.09
Variance of saccades duration*	**1.32 e-03 ± 9.43 e-04**	**1.70 e-03 ± 1.10 e-03**	**0.02**
Kurtosis of saccade duration*	**3.35 ± 1.61**	**4.23 ± 2.4**	**0.03**
Skewness of saccade duration	1.08 ± 0.67	1.26 ± 0.75	0.14
Average amplitude of saccades**	**0.81 ± 0.35**	**1.17 ± 0.48**	**4.10 e-08**
Variance of saccade amplitude**	**0.40 ± 0.31**	**0.9 ± 0.63**	**1.03 e-08**
Kurtosis of saccade amplitude**	**2.59 ± 1.02**	**3.02 ± 1.09**	**2.97 e-03**
Skewness of saccade amplitude**	**0.57 ± 0.58**	**0.83 ± 0.6**	**4.32 e-03**
Std. of pupil dilation	1.79 e-01 ± 1.88 e-02	1.84 e-01 ± 2.37 e-02	0.11
Mean-Std. Ratio*	**2.84 ± 0.47**	**2.7 ± 0.49**	**0.05**
Pupil dilation skewness	-0.04 ± 0.42	-0.02 ± 0.43	0.72
Pupil dilation kurtosis	2.98 ± 0.49	2.94 ± 0.6	0.18
Variance of pupil dilation	3.22 e-02 ± 0.65 e-02	3,41 e-02 ± 0.82 e-02	0.10
P1*	**2.33 e-02 ± 1.23 e-02**	**2.96 e-02 ± 1.74 e-02**	**0.03**
P2	0.63 e-02 ± 0.31 e-02	0.66 e-02 ± 0.36 e-02	0.75
P LF**	**1.14 e-02 ± 0.63 e-02**	**1.52 e-02 ± 0.91 e-02**	**8.99 e-03**
P HF	9,80 e-03 ± 0.53 e-02	1.21 e-02 ± 0.81 e-02	0.22
P VHF1	4.60 e-03 ± 0.25 e-02	0.47 e-02 ± 0.28 e-02	0.73
P VHF2	0.19 e-02 ± 1.00 e-03	0.16 e-02 ± 8.87 e-04	0.17
P VHF3	1.62 e-04 ± 7.21 e-05	1.62 e-04 ± 6.45 e-05	0.77
Average pupil dilation	0.51 ± 7.98 e-02	0.49 ± 0.08	0.43

* p<0.05, ** p<0.01

Table 9 Values of mean and standard deviation of all the extracted features for the stimuli featuring acquaintances (A) and strangers (B) and relative p value got as output of the Wilcoxon rank sum test. In bold text are highlighted the features that present a significant difference (p<0.05); OFV x and OFV y stand for x and y coordinates of the overall fixation vector, std. stands for standard deviation, , P1 is the power of pupil dilation in [0,0.45] Hz band, P2 is the power of pupil dilation in [0.45,5] Hz band P LF is power of pupil dilation in [0.04,015] Hz band P HF is the power of pupil dilation in [0.15,0.45] Hz band, P VHF 1 is the power of pupil dilation in [0.45,1] Hz band, P VHF 2 is the power of pupil dilation in [1,2.5] Hz band, P VHF 3 is the power of pupil dilation in [2.5,5] Hz band

Statistical-based feature selection is achieved after the features extraction phase in order to discard all features that are not relevant. The first step of the statistical analysis is checking the normality of the extracted features through the Kolmogorov-Smirnov test. The results of this test showed that all features do not follow a normal distribution, with p-values <0.001, much lower than the threshold of significance of 0.05. Therefore, non-parametric Wilcoxon rank sum test is chosen to test if there is a statistically significant difference between the resting and video phase in each feature. As it is shown in Table 9, a total of 13 features out of 28 present a significant difference. These features are the average duration of fixations (p = 0.03), the kurtosis of fixation duration (p = 0.03), the skewness of fixation duration (p = 0.01), the variance of saccades duration (p = 0.02), the kurtosis of saccade duration (p = 0.03), the mean-standard deviation ratio (p = 0.05) and P1, the power associated to the pupil dilation in the frequency band [0.45 , 5] Hz (p = 0.03), while the significantly different features that present a p value smaller than 0.01 are the variance of fixation duration ($p = 1.23x10^{-8}$), the average amplitude of saccades ($p = 4.10x10^{-8}$), the variance of saccade amplitude ($p = 1.03x10^{-8}$), the kurtosis of saccade amplitude ($p = 2.97x10^{-3}$), the skewness of saccade amplitude ($p = 4.32x10^{-3}$) and P LF, the power associated to the pupil dilation in the frequency band [0.04 , 0.15] Hz ($p = 8.99 \quad x10^{-3}$).

In order to have a graphic representation of the features belonging to two different populations with different mean, in Figure 27 the bar plot showing the mean value and standard deviation for each one of these features is presented. Note that the features have been normalised with the min-max normalisation technique in order to better visualise them since they may have different orders of magnitude.

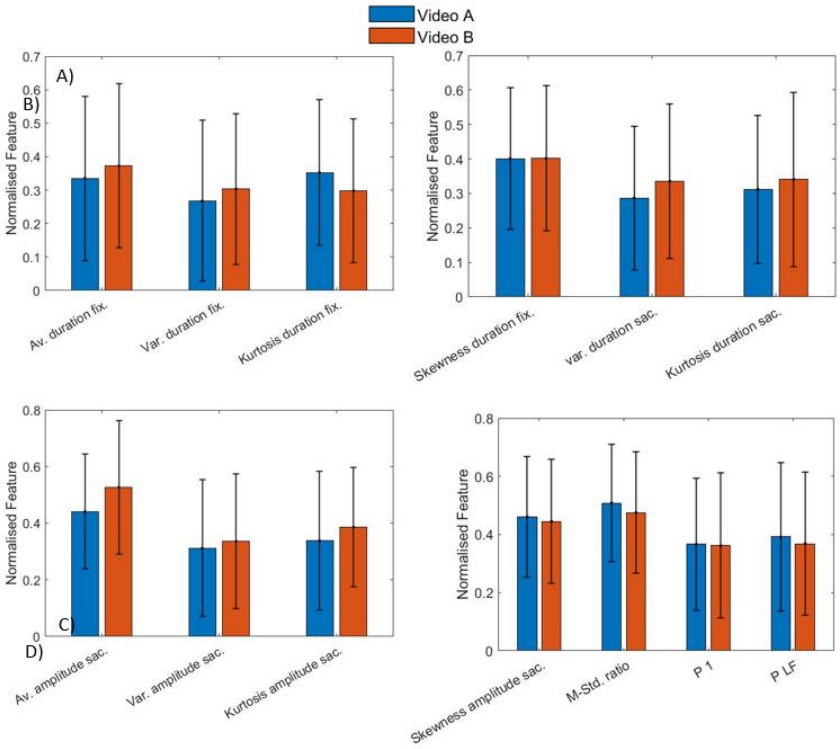

Figure 27 graphical representation of the values of all the features whose relative p value is lower than 0.05: A) Average duration of fixations, variance of fixation duration, kurtosis of fixation duration B) skewness of fixation duration, variance of saccade duration, kurtosis of saccade duration C) Average saccade amplitude, variance of saccade amplitude, kurtosis of saccade amplitude D) skewness of saccade amplitude, mean-standard deviation ratio, power of pupil dilation in [0,0.45] Hz band, power of pupil dilation in [0.04,015] Hz band, power of pupil dilation in [0.15,0.45] Hz

1.1.2. CLASSIFICATION

The successive step after the feature extraction and feature selection of the 13 features consists in the training and validation of the machine learning models. In this case, since we need to discern between two classes (video of type A and video of type B), a binary classifier is needed. The machine learning models that are going to be exploited are 10: Linear Discriminant Analysis (LDA), linear Support Vector Machines (SVM), quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, coarse Gaussian SVM fine k-Nearest Neighbour (k-NN), medium k-NN and finally coarse k-NN. In Table 10 the accuracy of each one of the trained models obtained through the 5 fold cross-validation technique is shown.

Accuracy			
LDA	**75.00%**	Medium Gaussian SVM	**79.20%**
Linear SVM	78.60%	Coarse Gaussian SVM	78.00%
Quadratic SVM	70.80%	Fine k-NN	**74.40%**
Cubic SVM	72.60%	Medium k-NN	73.80%
Fine Gaussian SVM	63.10%	Coarse k-NN	61.30%

Table 10 Accuracy values for all the trained models; in bold text the accuracy values associated to each one of the model types (LDA, SVM, k-NN)

As it is possible to see in the previous table, most of the trained models present a degree of accuracy between 70% and 80%. In particular, the accuracy of the LDA model is equal to 75.00%, the accuracy of the linear SVM model is equal to 78.60%, the accuracy of the quadratic SVM model is equal to 70.80%, the accuracy of the cubic SVM model is equal to 72.60%, the accuracy of the Fine Gaussian SVM model is equal to 63.10%, the accuracy of the Medium Gaussian SVM model is equal to 79.20%, the accuracy of the coarse Gaussian SVM model is equal to 78.00%, the accuracy of the fine k-NN model is equal to 74.40%, the accuracy of the medium k-NN model is equal to 73.80% while the accuracy of the coarse k-NN model is equal to 61.30%. The model characterised by the highest accuracy between all the trained models is the medium Gaussian SVM, while the model characterised by the lowest accuracy is the coarse k-NN.

Together with the accuracy a set of three performance metrics is computed per each one of the forementioned models: precision, recall and F1 score. In particular, these performance metrics are computed for both the classes. In Table 11 the values of precision, recall and F1 score per each one of the trained models is shown.

	LDA			Linear SVM	
Classes	P	R	F1	P	R
A	0.73	0.73	0.73	0.73	0.73
B	0.73	0.73	0.73	0.73	0.73
Mean	0.73	0.73	0.73	0.73	0.73

	Quadratic SVM			Cubic SVM	
Classes	P	R	F1	P	R
A	0.81	0.77	0.79	0.77	0.71
B	0.76	0.80	0.78	0.68	0.75
Mean	0.79	0.79	0.79	0.73	0.73

	Fine Gaussian SVM			Medium Gaussian	
Classes	P	R	F1	P	R
A	0.32	0.84	0.47	0.79	0.76
B	0.94	0.58	0.72	0.74	0.77
Mean	0.63	0.71	0.60	0.77	0.77

	Coarse Gaussian SVM			Fine k-NN	
Classes	P	R	F1	P	R
A	0.85	0.75	0.79	0.75	0.74
B	0.71	0.82	0.76	0.74	0.75
Mean	0.78	0.79	0.78	0.75	0.75

	Medium k-NN			Coarse k-NN	
Classes	P	R	F1	P	R
A	0.87	0.69	0.77	0.98	0.57
B	0.61	0.82	0.7	0.25	0.91
Mean	0.74	0.76	0.74	0.62	0.74

Table 11 Values of the per class precision (P), recall (R), F1 score (F1) and mean value of precision, recall and F1 score associated to all the trained models

The LDA model presents an average F1 score (mean value between the two per class F1 score) equal to 0.73. The SVM model that present the highest average F1 score is the quadratic SVM (0.79). The k-NN model that presents the highest average F1 score is the fine k-NN (0.75). The model that shows the lowest average F1 score is the coarse k-NN.

In Figure 28, the confusion matrix of the method which present the highest accuracy is going to be shown. In particular, it is possible to see the confusion matrix of the medium Gaussian SVM.

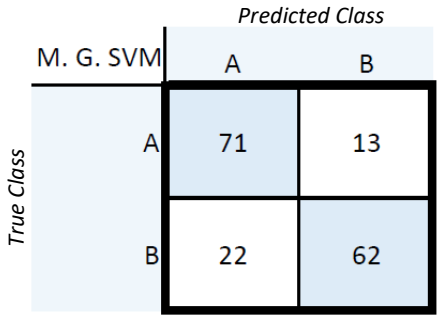

Figure 28 Confusion matrix relative to the trained models which present the highest accuracy, in this case medium Gaussian support vector machine (M. G. SVM). The class A refers to the videos featuring the acquaintances, while class B refers to the videos featuring the strangers

1.2. VALENCE AND AROUSAL VIDEO CLASSIFICATION

In this chapter the results of the comparison between the videos belonging to different emotional categories on the base of the arousal and valence assessed values are going to be shown. In particular we will examine the values of the extracted features explained in 2.4.7 and the results of the statistical analysis implemented in order to check the significant differences present in the

different classes. Moreover, the results of the machine learning classification models will be shown and analysed in order to check if the models are able to discern between the different cases. The categorisation criteria for the placement of each audio-visual stimulus in a label is shown in chapter 2.5.1 In Figure 29 it is possible to see the scatter plot of the distribution of the videos in the valence arousal plane:

Figure 29 Distribution of the videos in the valence-arousal plane; in each square there is the number of videos with the relative values of valence and arousal, the blank square corresponds to zero videos

For what concerns the videos of type A, a total of 66 videos belong to the category 1 (high valence, high arousal) and a total of 18 videos belong to category 4 (high valence, low arousal). For what concerns the videos of type B, a total of 8 videos belong to category 1, a total of 21 videos belong to category 3 (low valence, low arousal), while a total of 55 videos belong to category 4. Summing up the videos of type A and of type B per each one of the classes, we have that a total of 74 videos belong to category 1, a total of zero videos belong to category 2 (low valence, high arousal), a total of 21 videos belong to category 3, while a total of 73 videos belong to category 4. These results are in accordance with the expectations, since the videos were designed to be pleasant (high valence). Moreover it was expected to have a higher emotional involvement (high arousal) for the videos of type A, and a lower one (low arousal) for videos of type B. No videos belong to category 2, so at the end the machine learning models need to discern between 3 classes.

1.2.1. FEATURE EXTRACTION

In this case, the pre-processed data is divided in three groups of interest: category 1, which comprises videos with high valence and high arousal, category 3, which comprises videos with low valence and low arousal, and category 4, which comprises videos with high valence and low arousal.

A total of 28 features are extracted per each one of the groups of interest. The mean and standard deviation values of the features are shown in Table 12.

Feature	Category 1 Mean ± Std.	Category 3 Mean ± Std.	Category 4 Mean ± Std.	p value
Blink rate	20.11 ± 8.92	24.21 ± 9.46	19.13 ± 9.89	0.10
Average duration of fixations**	**0.53 ± 0.18**	**0.67 ± 0.18**	**0.54 ± 0.14**	**2.24 e-03**
Variance of fixation duration**	**0.17 ± 0.12**	**0.38 ± 0.21**	**0.27 ± 0.19**	**1.21 e-07**
Kurtosis of fixation duration	7.87 ± 3.34	8.91 ± 5.24	8.79 ± 4.08	0.48
Skewness of fixation duration	2.04 ± 0.69	2.21 ± 0.83	2.22 ± 0.72	0.50
OFV x	-91.25 ± 905.49	-67.52 ± 423.98	-66.59 ± 662.72	0.65
OFV y	1.75 e+03 ± 1.68 e+03	1.98 e+03 ± 1.63 e+03	2.03 e+03 ± 1.67 e+03	0.42
Average duration of saccades	4.27 e-02 ± 1.57 e-02	3.74 e-02 ± 1.95 e-02	4.52 e-02 ± 1.75 e-02	0.25
Variance of saccades duration	0.14 e-02 ± 9.53 e-04	0.13 e-02 ± 0.11 e-02	0.17 e-02 ± 0.11 e-02	0.36
Kurtosis of saccade duration	3.67 ± 1.81	3.92 ± 2.17	3.88 ± 2.32	0.92
Skewness of saccade duration	1.16 ± 0.69	1.3 ± 0.8	1.14 ± 0.71	0.79
Average amplitude of saccades**	**0.85 ± 0.36**	**0.94 ± 0.5**	**1.15 ± 0.48**	**1.35 e-04**
Variance of saccade amplitude**	**0.42 ± 0.31**	**0.75 ± 0.65**	**0.85 ± 0.64**	**2.95 e-04**
Kurtosis of saccade amplitude	2.70 ± 1.07	2.74 ± 1.22	2.94 ± 1.05	0.22
Skewness of saccade amplitude	0.60 ± 0.60	0.75 ± 0.74	0.79 ± 0.54	0.13
Std. of pupil dilation	0.19 ± 1,90 e-02	0.18 ± 2.37 e-02	0.18 ± 2.33 e-02	0.82
Mean-Std. Ratio	2.82 ± 0.52	2.69 ± 0.55	2.74 ± 0.43	0.56
Pupil dilation skewness	-4.73 e-02 ± 0.43	-3.99 e-02 ± 0.53	-1,58 e-02 ± 0.39	0.90
Pupil dilation kurtosis	2.95 ± 0.53	3.11 ± 0.69	2.92 ± 0.52	0.58
Variance of pupil dilation	3.31 e-02 ± 0.67 e-02	3.28 e-02 ± 0.86 e-02	3.33 e-02 ± 0.79 e-02	0.85
P1	2.39 e-02 ± 1.23 e-02	2.50 e-02 ± 1.55 e-02	2.94 e-02 ± 1.76 e-02	0.23
P2	0.66 e-02 ± 0.33 e-02	0.60 e-02 ± 0.28 e-02	0.64 e-02 ± 0.36 e-02	0.83
P LF	1.19 e-02 ± 0.64 e-02	1.32 e-02 ± 0.90 e-02	1.48 e-02 ± 0.90 e-02	0.21
P HF	1.04 e-02 ± 0.60 e-02	1.03 e-02 ± 0.60 e-02	1.17 e-02 ± 0.80 e-02	0.93
P VHF1	0.47 e-02 ± 0.25 e-02	0.45 e-02 ± 0.24 e-02	0.47 e-02 ± 0.28 e-02	0.97
P VHF2	0.19 e-02 ± 1.00 e-03	0.15 e-02 ± 0.66 e-03	0.17 e-02 ± 0.94 e-03	0.41
P VHF3	1.67 e-04 ± 7.63 e-05	1.43 e-04 ± 5.35 e-05	1.62 e-04 ± 6.28 e-05	0.48
Average pupil dilation	0.51 ± 8.24 e-02	0.48 ± 9.30 e-02	0.5 ± 7.57 e-02	0.48

* p<0.05, ** p<0.01

Table 12 Values of mean and standard deviation of all the extracted features for the video and resting phase and relative p value got as output of the Wilcoxon rank sum test. In bold text are highlighted the features that present a significant difference (p<0.05). OFV x and OFV y stand for x and y coordinates of the overall fixation vector, std. stands for standard deviation, , P1 is the power of pupil dilation in [0,0.45] Hz band, P2 is the power of pupil dilation in [0.45,5] Hz band P LF is power of pupil dilation in [0.04,015] Hz band P HF is the power of pupil dilation in [0.15,0.45] Hz band, P VHF 1 is the power of pupil dilation in [0.45,1] Hz band, P VHF 2 is the power of pupil dilation in [1,2.5] Hz band, P VHF 3 is the power of pupil dilation in [2.5,5] Hz band

Statistical-based feature selection is needed to discard all features that are not relevant. The Kolmogorov-Smirnov test is used to check the normality of the extracted features. The results of this test showed that all features do not follow a normal distribution, with p-values <0.001, much lower than the threshold of significance of 0.05. Given these results, Kruskal Wallis test is chosen to test if there is a statistically significant difference between the resting and video phase in each feature. As it is shown in Table 12, a total of 4 features out of 28 present a significant difference. These features are the average duration of fixations ($p = 2.24x10^{-3}$), the variance of fixation duration ($p = 1.21x10^{-7}$), the average amplitude of saccades ($p = 1.35x10^{-4}$), the variance of saccade amplitude ($p = 2.95x10^{-4}$). The Bonferroni multiple comparison test showed that 2 features are suitable for distinguishing between category 1 and category 3: these features are the average duration of fixations ($p = 1.40x10^{-3}$) and the variance of fixation duration ($p = 1.33x10^{-6}$). Moreover it showed that 3 features are suitable for distinguishing between category 1 and category 4: these features are the variance of fixation duration ($p = 1.48x10^{-4}$), the average amplitude of saccades ($p = 7.67x10^{-}$

[5]) and the variance of saccade amplitude ($p = 1.91 \times 10^{-4}$). Finally the Bonferroni multiple comparison test showed that just one feature is suitable for distinguishing between category 3 and category 4: the average duration of fixation ($p = 0.01$).

In order to have a graphic representation of the features belonging to the three different populations with different mean, in Figure 30 the bar plot showing the mean value and standard deviation for each one of these features is presented. Note that the features have been normalised with the min-max normalisation technique in order to better visualise them since they may have different orders of magnitude.

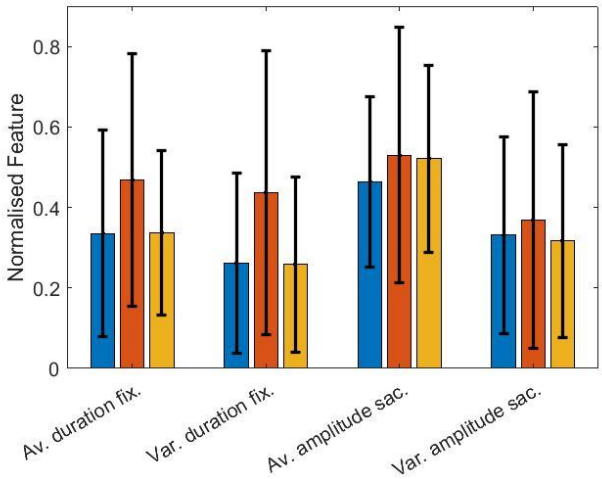

Figure 30 Graphical representation of the values of all the features whose relative p value is lower than 0.05: Average duration of fixations, variance of fixation duration, average amplitude of saccades, variance of saccade amplitude

1.2.2. CLASSIFICATION

The machine learning models are now trained given as input the 4 forementioned input features. In this case, since we need to discern between three classes (category 1, category 3, category 4), a multi-label classifier is needed. The machine learning models that are going to be exploited are 10: Linear Discriminant Analysis (LDA), linear Support Vector Machines (SVM), quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, coarse Gaussian SVM fine k-Nearest Neighbour (k-NN), medium k-NN and finally coarse k-NN. In Table 13 the accuracy of each one of the trained models obtained through the 5 fold cross-validation technique is shown.

	Accuracy		
LDA	**66.10%**	Medium Gaussian SVM	**64.30%**
Linear SVM	**64.30%**	Coarse Gaussian SVM	63.10%
Quadratic SVM	63.10%	Fine k-NN	54.80%
Cubic SVM	61.30%	Medium k-NN	**61.90%**
Fine Gaussian SVM	58.30%	Coarse k-NN	54.80%

Table 13 Accuracy values for all the trained models; in bold text the accuracy values associated to each one of the model types (LDA, SVM, k-NN)

As it is possible to see in the previous table, the accuracy of the LDA model is equal to 66.10%, the accuracy of the linear SVM model is equal to 64.30%, the accuracy of the quadratic SVM model is equal to 63.10%, the accuracy of the cubic SVM model is equal to 61.30%, the accuracy of the Fine Gaussian SVM model is equal to 58.30%, the accuracy of the Medium Gaussian SVM model is equal to 64.30%, the accuracy of the coarse Gaussian SVM model is equal to 63.10%, the accuracy of the fine k-NN model is equal to 54.80%, the accuracy of the medium k-NN model is equal to 61.90% while the accuracy of the coarse k-NN model is equal to 54.80%. The model characterised by the highest accuracy between all the trained models is the LDA, while the model characterised by the lowest accuracy is the fine Gaussian SVM.

Together with the accuracy a set of three performance metrics is computed per each one of the forementioned models: precision, recall and F1 score. In particular, these performance metrics are computed for all the classes. In Table 14 the values of precision, recall and F1 score are grouped and shown.

	LDA			Linear SVM		
Classes	P	R	F1	P	R	F1
Category 1	0.84	0.68	0.75	0.84	0.67	0.74
Category 3	0.19	0.40	0.25	0.00	0.00	0.00
Category 4	0.62	0.67	0.64	0.63	0.61	0.62
Mean	0.55	0.58	0.55	0.49	0.43	0.45

	Quadratic SVM			Cubic SVM		
Classes	P	R	F1	P	R	F1
Category 1	0.77	0.66	0.71	0.73	0.71	0.72
Category 3	0.19	0.40	0.26	0.24	0.25	0.24
Category 4	0.61	0.63	0.63	0.60	0.61	0.61
Mean	0.52	0.56	0.53	0.52	0.52	0.52

	Fine Gaussian SVM			Medium Gaussian SVM		
Classes	P	R	F1	P	R	F1
Category 1	0.57	0.72	0.64	0.82	0.64	0.72
Category 3	0.00	0.00	0.00	0.19	0.67	0.30
Category 4	0.77	0.56	0.65	0.59	0.64	0.61
Mean	0.45	0.43	0.43	0.53	0.65	0.54

	Coarse Gaussian SVM			Fine k-NN		
Classes	P	R	F1	P	R	F1
Category 1	0.88	0.62	0.72	0.57	0.72	0.64
Category 3	0.00	0.00	0.00	0.00	0.00	0.00
Category 4	0.56	0.65	0.60	0.77	0.51	0.62
Mean	0.48	0.42	0.44	0.45	0.41	0.42

	Medium k-NN			Coarse k-NN		
Classes	P	R	F1	P	R	F1
Category 1	0.82	0.64	0.72	0.88	0.62	0.73
Category 3	0.19	0.67	0.3	0.00	0.00	0.00
Category 4	0.59	0.64	0.61	0.56	0.65	0.6
Mean	0.53	0.65	0.54	0.48	0.42	0.44

Table 14 Values of the per class precision (P), recall (R), F1 score (F1) and mean value of precision, recall and F1 score associated to all the trained models; category 1 is the high-arousal, high-valence category, category 3 is the low-arousal, low-valence category, category 4 is the low-arousal, high valence category

The LDA model presents an average F1 score (mean value between the three per class F1 score) equal to 0.43. The SVM model that presents the highest average F1 score is the medium Gaussian SVM (0.54). The k-NN model that presents the highest average F1 score is the medium k-NN (0.54). The model that shows the lowest average F1 score is the fine Gaussian SVM. It is possible to see that for some of the models (linear SVM, fine Gaussian SVM and coarse k-NN), the classification process misclassified all the samples belonging to the third category, in fact the values of the per class precision, recall and F1 score relative to the third class are all equal to zero. The presence of these null values makes the values of the average precision, recall and F1 score drop to a low value. In particular, the linear SVM model present a high accuracy compared to the accuracies of the other SVM models, but it completely misclassifies the third category leading to a low value of the average F1 score. Moreover, it is possible to note that in all the models trained, the values of per class precision, recall and F1 score relative to category 3 are lower than the per class precision, recall and F1 score relative to category 1 and category 4.

In Figure 31, the confusion matrix of the method which present the highest accuracy is going to be shown. In particular, it is possible to see the confusion matrix of the LDA model

LDA	Cat.1	Cat. 3	Cat.4
Cat. 1	62	2	10
Cat. 3	5	4	12
Cat. 4	24	4	45

Figure 31 Confusion matrix relative to the trained model which present the highest accuracy, in this case the linear discriminant analysis (LDA); the classes cat. 1, cat. 2, cat. 3 refers to category 1 (high valence-high arousal), category 3 (low valence-low arousal) and category 4 (high valence-low arousal)

1.3. HIGH AND LOW AROUSAL VIDEO CLASSIFICATION

As a result of the classification between the three emotional classes analysed in chapter …, it has been shown that the different machine learning models trained are not able to provide a good classification of the category 3 since the dataset is unbalanced due to a smaller number of samples belonging to category 3. However, the classification between the videos belonging to category 1 (high valence and high arousal) and category 4 (high valence and low arousal) gives robust results. Therefore, in this chapter the results of the comparison between the videos belonging to category 1 and the videos belonging to category 4 are going to be shown. In particular, we will examine the values of the extracted features explained in chapter 2.4.7 and the results of the statistical-based feature selection. Finally, the results of the machine learning classification models will be shown.

1.3.1. FEATURE EXTRACTION

The pre-processed data is divided in the two groups of interest: category 1, which comprises videos with high valence and high arousal and category 4, which comprises videos with high valence and low arousal.

A total of 28 features are obtained through the feature extraction as it has been explained in chapter 2.4.7. The mean and standard deviation values of the features are shown in the columns relative to category 1 and category 4 of Table 12. Statistical-based feature selection is achieved after the features extraction phase in order to discard all features that are not relevant. The first step of the statistical analysis is checking the normality of the extracted features through the Kolmogorov-Smirnov test. The results of this test showed that all features do not follow a normal distribution, with p-values <0.001, much lower than the threshold of significance of 0.05. Therefore, non-parametric Wilcoxon rank sum test is chosen to test if there is a statistically significant difference between the resting and video phase in each feature. A total of 4 features out of 28 present a significant difference. These features are the variance of fixation duration ($p = 4.13x10^{-5}$), the average amplitude of saccades ($p= 2.40x10^{-5}$), the variance of saccade amplitude ($p = 6.02x10^{-5}$), the skewness of saccade amplitude ($p = 0.04$). It is interesting to note that the result of the Wilcoxon rank sum test highlights a significant different that was not determined through the Bonferroni multiple comparison test applied in chapter ….. This feature is the skewness of saccade amplitude.

In order to have a graphic representation of the features belonging to two different populations with different mean, in Figure 32 the bar plot showing the mean value and standard deviation for each one of these features is presented. Note that the features have been normalised with the min-max normalisation technique in order to better visualise them since they may have different orders of magnitude.

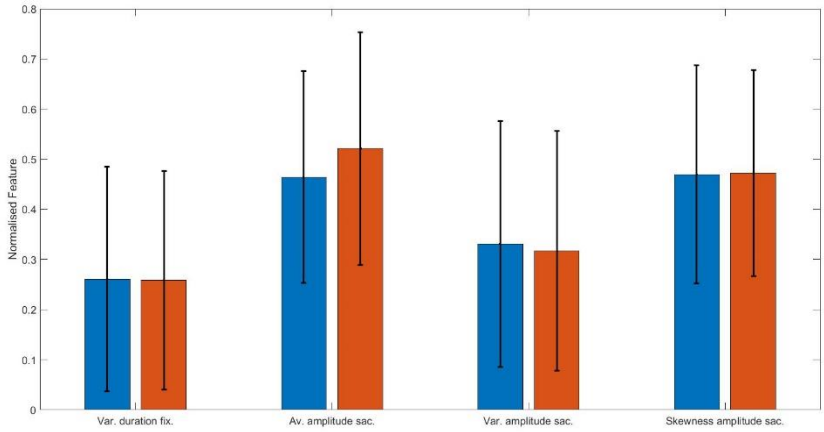

Figure 32 Graphical representation of the values of all the features whose relative p value is lower than 0.05: variance of the duration of fixations, average amplitude of saccades, variance of saccade amplitude, skewness of saccade amplitude

1.3.2. CLASSIFICATION

The next step consists in the training and validation of the machine learning models. In this case, since we need to discern between two classes (high and low arousal), a binary classifier is needed. The machine learning models that are going to be exploited are 10: Linear Discriminant Analysis (LDA), linear Support Vector Machines (SVM), quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, coarse Gaussian SVM fine k-Nearest Neighbour (k-NN), medium k-NN and finally coarse k-NN. In Table 15 the accuracy of each one of the trained models obtained through the 5 fold cross-validation technique is shown.

Accuracy			
LDA	**71.40%**	Medium Gaussian SVM	**71.40%**
Linear SVM	71.40%	Coarse Gaussian SVM	70.70%
Quadratic SVM	68.00%	Fine k-NN	63.30%
Cubic SVM	64.60%	Medium k-NN	**66.70%**
Fine Gaussian SVM	66.70%	Coarse k-NN	53.10%

Table 15 Accuracy values for all the trained models; in bold text the accuracy values associated to each one of the model types (LDA, SVM, k-NN)

As it is possible to see in the previous table, the accuracy of the LDA model is equal to 71.40%, the accuracy of the linear SVM model is equal to 71.40%, the accuracy of the quadratic SVM model is equal to 68.00%, the accuracy of the cubic SVM model is equal to 64.60%, the accuracy of the Fine Gaussian SVM model is equal to 66.70%, the accuracy of the Medium Gaussian SVM model is equal to 71.40%, the accuracy of the coarse Gaussian SVM model is equal to 70.70%, the accuracy of the fine k-NN model is equal to 63.30%, the accuracy of the medium k-NN model is equal to 66.70% while the accuracy of the coarse k-NN model is equal to 53.10%. The model characterised by the highest accuracy between all the trained models is the LDA and the medium Gaussian SVM, while the model characterised by the lowest accuracy is the coarse k-NN.

Together with the accuracy a set of three performance metrics is computed per each one of the forementioned models: precision, recall and F1 score. In particular, these performance metrics are computed for all the classes. In Table 16 the values of precision, recall and F1 score are grouped and shown.

	LDA			Linear SVM		
Classes	P	R	F1	P	R	F1
Category 1	0.80	0.69	0.74	0.80	0.69	0.74
Category 4	0.63	0.75	0.69	0.63	0.75	0.69
Mean	0.72	0.72	0.72	0.72	0.72	0.72

	Quadratic SVM			Cubic SVM		
Classes	P	R	F1	P	R	F1
Category 1	0.78	0.65	0.67	0.70	0.69	0.65
Category 4	0.58	0.66	0.62	0.59	0.65	0.68
Mean	0.68	0.66	0.65	0.65	0.67	0.67

	Fine Gaussian SVM			Medium Gaussian SVM		
Classes	P	R	F1	P	R	F1
Category 1	0.61	0.68	0.74	0.82	0.68	0.74
Category 4	0.73	0.77	0.68	0.60	0.77	0.68
Mean	0.67	0.73	0.36	0.71	0.73	0.71

	Coarse Gaussian SVM			Fine k-NN		
Classes	P	R	F1	P	R	F1
Category 1	0.86	0.66	0.75	0.62	0.64	0.63
Category 4	0.55	0.80	0.62	0.64	0.63	0.64
Mean	0.71	0.73	0.69	0.63	0.64	0.64

	Medium k-NN			Coarse k-NN		
Classes	P	R	F1	P	R	F1
Category 1	0.74	0.65	0.69	0.96	0.52	0.67
Category 4	0.59	0.69	0.64	0.10	0.70	0.17
Mean	0.67	0.67	0.67	0.53	0.61	0.42

Table 16 Values of the per class precision (P), recall (R), F1 score (F1) and mean value of precision, recall and F1 score associated to all the trained models; category 1 is the high-arousal, high-valence category, category 4 is the low-arousal, high valence category

The LDA model presents an average F1 score (mean value between the three per class F1 score) equal to 0.72. The SVM model that presents the highest average F1 score is the linear SVM (0.72). The k-NN model that presents the highest average F1 score is the medium k-NN (0.67). The model that shows the lowest average F1 score is the fine Gaussian SVM (0.36).

In Figure 33, the confusion matrix of the methods which present the highest accuracy and average F1 score is going to be shown. In particular, it is possible to see the confusion matrices of the LDA model, since the average F1 score of the LDA model (0.72) is higher than the average F1 score of the medium Gaussian SVM (0.71).

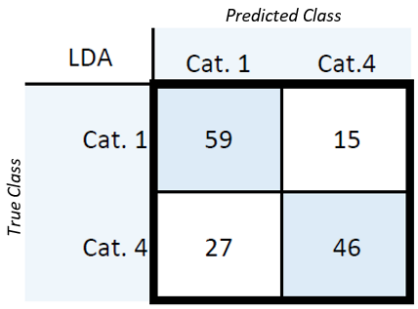

Figure 33 Confusion matrix relative to the trained model which present the highest accuracy and the higher average F1-score, in this case the linear discriminant analysis (LDA); the classes cat. 1, cat. 2, cat. 3 refers to category 1 (high valence-high arousal), category 3 (low valence-low arousal) and category 4 (high valence-low arousal)

2. DISCUSSION

The main objective of this study is to verify if the difference in the features extracted by eye-tracking data can be exploited for: the classification between the physiological response of the subject to a neutral resting phase and a video phase (stimuli and resting phase classification), the classification between the physiological response of the subject to an audio-visual stimulation representing a referred loved person and an unknown person (known and unknown video classification) and the

classification between the physiological response of the subject to audio-visual stimuli categorised in distinct emotional responses in terms of emotion recognition (valence and arousal classification and high arousal-low arousal classification). In order to achieve these objectives, an analogous procedure is carried out: the signal processing of the eye-tracking raw data, a feature extraction phase, a feature selection phase and a classification phase are done. Once these steps are achieved, it is possible to analyse which of the oculomotor and pupil dilation variables can be exploited for each one of the classifiers and their performances. In this chapter the results obtained in this study in each one of the four classifications are going to be discussed.

2.1. STIMULI AND RESTING PHASE CLASSIFICATION

After the feature extraction, it has been shown in chapter 3.2.1 that most of the features extracted (18 features out of 28) from the eye-tracking data present a significant different between the video phase and resting phase with *p* value smaller than 0.05. In particular, these features belong to both the oculomotor variables and the pupil dilation variables.

The average duration of fixations found in the resting phases is found to be significantly higher than the one found during the stimulation periods. Moreover, a significant difference is also found in the features relative to the saccades. In particular, the saccadic activity encountered during the stimulation periods is on average higher than during the resting phase confirming that the gaze coordinates are subjected to a higher dispersion during the stimulation periods. This may be due to the fact that during the video phase the attention of the subject is captured from different spatial regions in the delivered video. As B. Mahanama et al. state in [17], saccades tend to be inhibited during engaged visual attention on stationary stimuli implicating a higher average fixation duration, thus our results respect this statement since the resting phase can be considered as a stationary visual stimulus representing a white cross centered in the origin of the video frames. The y coordinate of the overall fixation vector characterising the audio-visual stimulation periods is found to be lower during the resting phase. Since the overall fixation vector embeds information about the number, position and duration of the fixations, the previous assumption may be explained by a high concentration of fixations around the white cross placed in the origin of the video phase.

The power computed in different frequency bands of the estimated spectral density of the pupil dilation signal is generally higher in the case of the audio-visual stimulation. The only case in which this result is not present is the case of the low frequency band (P LF). Therefore, it is possible to state that the phasic component of the pupil dilation is more evident during the audio-visual stimulation, while the tonic component of the pupil dilation is more evident in the resting phase.

In conclusion, it is possible to state that the ocular physiological response of the subject can be exploited for the classification between audio-visual stimulation and no stimulation. This can be accredited by the fact that these features are shared in the other classification processes with the exception of the skewness of fixation duration and the skewness of fixation duration for the known-unknown categorisation and of the skewness of saccade amplitude for the high arousal-low arousal categorisation.

For what concerns the classification performances, the trained machine learning models present an accuracy higher than 86.80% with the exception of the fine Gaussian SVM which presents an accuracy equal to 55.80%. In particular, the model that present the highest accuracy is the medium Gaussian SVM with 94.40%. As it is possible to see, the highest accuracy is given by a SVM model which is the kind of model that is most widely used in literature as in [49], [33], [50],[51], [52], [53]. The values of the average F1 scores are higher than 0.86 with the exception of the fine Gaussian SVM (0.36), confirming the high performance of the classifiers.

For what concerns the values of per class precision, recall and F1 score, it is possible to see in Table 8 that these values relative to the two classes (video and rest) present in most of the trained machine learning models just a slight difference, implicating that the models are able to classify the two classes with almost the same degree of accuracy.

2.2. KNOWN AND UNKNOWN VIDEO CLASSIFICATION

A total of 13 features out of 28 were selected during the features selection relative to the classification between videos featuring acquaintances and strangers. These features belong to both the oculomotor variables and the pupil dilation variables.

The features extracted from fixations that are characterised by a significant difference present a p value not smaller than 0.01 with the exception of the variance of fixation duration, while the features extracted from the saccades that are characterised by a significant difference present a p value much smaller than 0.001. This result permits us to state that in the classification between videos featuring acquaintances and strangers, the features extracted from the saccadic events present a more marked significant difference with respect to the features extracted from fixations. In particular, the average duration of fixations is slightly lower in the videos of the acquaintances while the variance of duration fixations is higher in the case of videos featuring strangers. Moreover, the videos featuring the acquaintances present a lower value of all the features extracted from saccadic events. Therefore, these differences between the two types of videos may be interpreted as a higher visual attention during the viewing of videos featuring acquaintances.

The pupil responses associated to the two classes of videos present a significant difference in the low frequency bands of the pupil dilation signal. In particular, a slight significant difference is encountered in the power associated to the frequency band [0, 0.45] Hz, and a significant difference is encountered in the power associated in the frequency band [0.04, 0.15] Hz. In both cases, the computed power is lower in the videos featuring acquaintances, implicating a more evident tonic component of the pupil dilation in the case of the videos featuring strangers.

For what concerns the classification performances, it is possible to check in Table 11 that most of the trained machine learning models present an accuracy higher than 70.80% with the exception of the fine Gaussian SVM (63.10%) and the coarse k-NN (61.30%). In particular, the model that present the highest accuracy is the medium Gaussian SVM with 79.20%. As it is possible to see,

as well as in the case of the rest-video classification, the highest accuracy is given by an SVM model which is the kind of model that is most widely used in literature as in the studies [49], [33], [53], [50],[51], [52]. The accuracy obtained for this classification problem can be compared to the accuracies obtained with previous studies in which machine learning models are exploited for a classification algorithm in which positive (or pleasant), negative (or unpleasant) and neutral stimuli were presented to the participants. Indeed, the videos of the acquaintances can be tagged as positive stimuli, while the videos of the strangers can be tagged as neutral stimuli. In Table 17 , the type of stimuli provided, the number and types of classes, the machine learning models and the accuracies of these studies are shown.

Works	Stimuli type	Classes	ML models	Accuracy
Our research	Videos	2 (positive, neutral)	M. G. SVM	79.2
Soleymani et al. [50]	Videos	3 (unpleasant, neutral, pleasant)	SVM	66.6
W. Zheng et al. [53]	Videos	3 (positive, neutral, negative)	Linear SVM	45,78 ± 11,03 58,90 ± 10,25
Y. Lu et al. [49]	Videos	3 (positive, neutral, negative)	Linear SVM	77,80 ± 14,61
Alhargan et al. [52]	Games	3 (negative, neutral, positive)	SVM	56,1 ± 5,9
Alhargan et al. [51]	Games	3 (negative, neutral, positive)	SVM	58.7

Table 17 Comparison between different emotion recognition studies in whose defined classes are positive (or pleasant), neutral and negative ()or unpleasant. Stimuli type, number and type of the classes, machine learning models and relative accuracies are shown

As shown in the previous figure, the performance obtained in this study is higher or comparable with the accuracy obtained in the study of Soleymani et al. [50], W.Zheng et al. [33], Alhargan et al. [52], and it can be considered comparable to the accuracy obtained in the study of Y. Lu et al. [49]. Nevertheless, it must be considered that a direct comparison between the accuracy of our study and of the forementioned study is not possible, since in our case, for the classification between videos of the acquaintances (positive stimuli) and of strangers (negative stimuli) a binary classifier was built, while in the other forementioned studies a multi-label classifier with three output classes was built.

For what concerns the values of per class precision, recall and F1 score, it is possible to see in Table 11 that these values relative to the two classes (known and unknown) present in most of the trained machine learning models just a slight difference, implicating that the models are able to classify the two classes with almost the same degree of accuracy.

2.3. VALENCE AND AROUSAL VIDEO CLASSIFICATION

The feature selection of the comparison among videos with different levels of valence and arousal achieved 4 features out of 28. In particular, these features belong to the oculomotor variables.

First, a significant difference in the average fixation duration feature was encountered. Specifically, as a result of the Bonferroni multiple comparison test, a significant difference between videos belonging to category 1 (high valence – high arousal) and category 3 (low valence – low arousal) as well as between category 4 (high valence – low arousal) and category 3 was found. The average value of fixation duration is then found to be higher in the videos considered to have a low-valence (unpleasant) level and this result can be found also in other studies in which an emotion recognition is implemented such as the study of A. Alhargan et al. [29] and the study of Y. Lu et al [32]. The variance of fixation duration presents a significant difference between category 1 and category 3 as well as between category 1 and category 4. No significant difference was found for this feature between category 3 and 4. For this reason it is possible to state that the variance of fixation duration is significantly in the videos characterised by a high arousal if compared to the videos characterised by a low level of arousal. At the same time, a significant difference is encountered for the average amplitude of saccades between videos characterised by a high and a low value of arousal (category 1 and category 4) where the average amplitude of saccades is found to be higher in high arousal videos. This same significant difference was encountered by P. Tarnowski in [30]. For what concerns the variance of saccade amplitude, this feature is found to be significantly smaller in the videos characterised by high valence and low arousal if compared to the videos characterised by high valence and high arousal. This same significant difference was encountered by P. Tarnowski in [30].

No significant difference is found in the variables related to the pupil dilation, in contrast with the previous studies which exploited only this kind of variables ([53], [52]) and with the studies in which both the oculomotor variables and the pupil dilation features are used as input variables to the machine learning models. For what concerns pupil dilation variables, in [50] a significant difference was found in the standard deviation of the pupil dilation, in [33] it was found in the variance of pupil dilation and in the average pupil diameter, in [49], [51] and [52] in the mean pupillary responses. The lack of significant differences in the pupil dilation variables compared with the results of the forementioned studies can be due to the different design of our study: in this study the stimuli delivered to the subjects were designed in order to produce a strong and positive emotion (high arousal and high valence) in the case of the video of the known person or, in contrast, a neutral response in the case of the video of the unknown person, so no negative emotions or low arousal emotions were meant to be elicited. On the contrary, in the forementioned studies, the emotional response of the subject covers all the areas of the valence-arousal plane (low, moderate, high values of valence and arousal).

For what concerns the classification performances, it is possible to check in Table 10 that most of the trained machine learning models present an accuracy is higher than 61.10% with the exception of the fine Gaussian SVM (58.30%), the fine k-NN (54.80%) and the coarse k-NN (54.80%).

In particular, the model that presents the highest accuracy and that can be considered the best trained model for this classification problem is the LDA with 66.10%. In the case of the work of W. Zheng et al. [53] , Y. Lu et al. [49], Alhargan et al. [52], the accuracies are provided as the mean value and the standard deviation of the machine learning models trained using the data extracted from each one of the participants. Moreover, it must also be noticed that in [51] and [52], slightly higher accuracies (some percentage points) are reached by applying the Hilbert transform to the pupil dilation signal. The accuracy reached by the algorithm implemented in this study is higher than the accuracy reached in some studies present in literature and it is lower than others. Nevertheless, in order to compare the listed studies, it must be considered that the design of the carried out experimentations is different as well as the chosen classes. The type of stimuli provided, the number and types of classes, the machine learning models and the accuracies of each one of the cited studies is shown in Table 18.

Works	Stimuli type	Classes	ML models	Accuracy
Our research	Videos	3 (High arousal-high valence, low arousal-low valence, low arousal-low valence)	LDA	66,1
Soleymani et al. [50]	Videos	3 (calm, medium aroused, activated)	SVM	71,1
Soleymani et al. [50]	Videos	3 (unpleasant, neutral, pleasant)	SVM	66,6
Tarnowski et al. [33]	Videos	3 (high arousal-high valence, low arousal,-moderate valence, high arousal-high valence)	SVM	80
W. Zheng et al. [53]	Videos	3 (positive, neutral, negative)	Linear SVM	45,78 ± 11,03 58,90 ± 10,25
Y. Lu et al. [49]	Videos	3 (positive, neutral, negative)	Linear SVM	77,80 ± 14,61
Alhargan et al. [52]	Games	3 (low arousal, neutral arousal, high arousal)	SVM	70,0 ± 5,8
Alhargan et al. [52]	Games	3 (negative, neutral, positive)	SVM	56,1 ± 5,9
Alhargan et al. [51]	Games	3 (low arousal, neutral arousal, high arousal)	SVM	71,4
Alhargan et al. [51]	Games	3 (negative, neutral, positive)	SVM	58,7

Table 18 Comparison between different emotion recognition studies. Stimuli type, number and type of the classes, machine learning models and relative accuracies are shown

The accuracy achieved in this study is higher than the accuracy achieved in some of the cited studies and lower than others. In particular, the accuracy achieved in this study (66.1%) is higher than the accuracy achieved by W. Zheng et al. in [53], Alhargan et al. in [52] (in the case of negative, neutral and positive classification) and in [51] (in the case of negative, neutral and positive

classification). On the contrary, it is lower than or comparable to the accuracy achieved by Soleymani et al. in [50] , Tarnowski et al. in [33], Y. Lu et al. in [49], and Alhargan et al. in [51] (in the case of low, neutral and high arousal classification) and in [52] (in the case of low, neutral and high arousal classification).

For what concerns the values of per class precision, recall and F1 score, it is possible to see in Table 14 the values relative to category 1 and category 4 present just a slight difference, while the values relative to category 3 are much lower, stating that the classifier is not able to classify in the right way the samples belonging to category 3. This property of the trained models may be due to the fact that the dataset used to train and test the models is unbalanced (21 samples belonging to category 3 against 74 and 73 samples belonging to category 1 and 4 respectively).

2.4. HIGH AND LOW AROUSAL VIDEO CLASSIFICATION

After the feature extraction, it has been shown in chapter 1.3.1 that 4 features out of 28 of the features extracted from the raw eye-tracking data present a significant different between the video phase and resting phase with *p* value smaller than 0.05. In particular, these features belong to the oculomotor variables. The variance of fixation duration was found to be smaller in the high arousal videos with respect to the one encountered in the low arousal videos. Moreover, the average amplitude of saccades, the variance of saccade amplitude and the skewness of saccade amplitude was found to be significantly smaller in the high arousal videos. It is interesting to note that the result of the Wilcoxon rank sum test highlights a significant difference in the skewness of saccade amplitude that was not determined through the Bonferroni multiple comparison test applied in chapter 1.2.1.

On the other hand, no significant difference was encountered in the pupil dilation variables. This evidence is in contrast with the results found in previous studies ([50], [51], [52]) where a significant difference in the mean pupillary responses was found. As in the previous classification problem (valence and arousal video classification), this difference in the results can be due to the different design of our study. Indeed, no emotions with extremely low arousal were meant to be elicited in this study, since the stimuli delivered to the subjects were designed in order to produce a strong and positive emotion (high arousal and high valence) in the case of the video of the acquaintances or a neutral response in the case of the video featuring strangers.

For what concerns the classification performances, it is possible to check in Table 13 that most of the trained machine learning models present an accuracy is higher than 63.30% with the exception of the coarse k-NN (53.10%). In particular, the models that present the highest accuracies are the LDA with 71.40% and the medium Gaussian SVM with 71.40%. The model chosen to be the best model for this classification problem is the LDA since the average F1 score achieved using this model (0.72) is higher than the one achieved using the medium Gaussian SVM (0.71). In order to compare the results achieved in this study with the results present in literature, it is possible to check Table 19.

Works	Stimuli type	Classes	ML method	Accuracy
Our research	Videos	2 (High arousal, low arousal)	LDA / M.G. SVM	71,4
Soleymani et al. [50]	Videos	3 (calm, medium aroused, activated)	SVM	71,1
Alhargan et al. [52]	Games	3 (low arousal, neutral arousal, high arousal)	SVM	70,0 ± 5,8
Alhargan et al. [51]	Games	3 (low arousal, neutral arousal, high arousal)	SVM	71,4

Table 19 Comparison between different classification studies relative to the degree of arousal. Stimuli type, number and type of the classes, machine learning models and relative accuracies are shown

In Table 19, for each one of the cited studies, the type of stimuli provided, the number and types of classes, the machine learning models and the accuracies are shown. Note that the studies presented in Table 19 also carried out a classification process relative to the levels of valence of the presented stimuli, but these cases are not taken into account for the comparison with this classification problem. In the case of the work of Alhargan et al. [52], the accuracies are provided as the mean value and the standard deviation of the machine learning models trained using the data extracted from each one of the participants. It must be considered that the results encountered in our study cannot be directly compared to the results encountered in the forementioned studies, because in our case we carry out a binary classification, while in the other cases a multi-label classification with three classes was achieved. Nevertheless, the values of accuracy achieved in this study, in the study of Soleymani et al. in [50] and in the study of Alhargan et al. in [51] and in [52] are comparable.

For what concerns the values of per class precision, recall and F1 score, it is possible to see in chapter Table 16 that these values relative to the two classes (category 1 and category 4) present in most of the trained machine learning models just a slight difference, implicating that the models are able to classify the two classes with almost the same degree of accuracy.

2.5. LIMITATIONS AND FUTURE STUDIES

Various limitations have been identified in this study. First, the difference in age and in the distribution of genders in the group of participants could influence the carried out results due to functional differences in the oculomotor system and in the autonomous nervous system of the subjects, even if no evidence about that was found in literature. Second, the number of participants to the study is limited and this bring to a limited dataset; in future studies, the employment of a considerable number of participants and so of a bigger dataset can lead to more generalized results. Furthermore, a specific experimental protocol should have been designed per each one of the carried out classifications. In particular, the design of the experimentation used in this study

perfectly fits the classification between the rest and video phase and the classification between videos featuring known and unknown people, while for the classification between emotional audio-visual stimuli differentiated in terms of the level of valence and arousal a different approach should have been used. Indeed, for the latter case future studies should opt for a different set of videos intended to elicit a wider set of emotions characterised by a wider arousal and valence range of values. A higher variability in the arousal values and in the valence values can lead to a higher performance of the machine learning models and so to better results.

3. CONCLUSIONS

This study aims at assessing the ability of machine learning algorithms to classify different emotional audio-visual stimuli exploiting eye-tracking data acquired from a group of healthy subjects. The main steps followed for the realisation of this project consist of data collection, data pre-processing and cleaning, features extraction and selection and, finally, the classifications. In particular, three different kinds of classifications were implemented and validated. First, the classification between audio-visual stimulation and resting phase. Concerning this classification, the machine learning model that provided the best performance in terms of accuracy was the medium Gaussian Support Vector Machine model with an accuracy of the 94.40%. Second, the classification between audio-visual stimuli featuring acquaintances and strangers. Similarly to the previous classification problem, the machine learning model that provided the best performance was the medium Gaussian Support Vector Machine model that in this case provided an accuracy 79.20%. Third, the classification between emotional audio-visual stimuli differentiated in terms of the level of valence (pleasure) and arousal (excitation) perceived by the participant. The chosen classes for this classification problem were high valence – high arousal stimuli, low valence – low arousal stimuli and high valence – low arousal stimuli. Together with this last classification, the classification between high arousal and low arousal stimuli was performed. In both these classification problems, the machine learning model presenting the best performance was the Linear Discriminant Analysis model with an accuracy of the 66.10%, in the former case, and with an accuracy of the 71.40% in the latter. In conclusion, in agreement with the general hypothesis, it is possible to state that the ocular response to different emotional audio-visual stimuli can be identified and classified through the analysis and exploitation of eye-tracking data in machine learning algorithms. Nevertheless, additional research about the optimisation of experimental protocols and machine learning algorithms is needed in order to provide a better performance of the classification between audio-visual stimuli characterised by different levels of valence and arousal.

The formulation of specific hypothesis was needed for the assessment of the general hypothesis. Concerning the specific hypothesis the following conclusions are drawn:

- It is fulfilled the hypothesis that the view of a video of a loved person answering questions about the participant provides a strong emotional response since most of these videos are assessed by the participants to belong in the area of the valence-arousal plane relative to high arousal and high valence.

- It is fulfilled the hypothesis that the view of a video of a completely unknown person produces to the subject a neutral emotional response different from the one produced by the video of a referred person. This is evidenced by the fact that the classification between the two types of videos provides good results.
- It is fulfilled the hypothesis that the assessment of the level of valence and arousal related to the videos featuring acquaintances are different from the assessment related to the videos of the unknown people. This is evidenced by the fact that, as a result of the SAM questionnaire, most of the videos featuring a known person are categorised in category 1 (high arousal and high valence), while most of the videos of the unknown person are categorised in category 4 (low arousal and high valence).
- It is fulfilled the hypothesis that the oculomotor behaviour of a participant is strictly related to the emotional response to the provided stimulus. This is evidenced by the fact that in all the classifications performed in this study, a significant difference is found in part of the features belonging to the oculomotor variables.
- It is partially rejected the hypothesis that the level of pupil dilation of a subject and the pupillary response is related to the emotional response to the provided stimulus. This is evidenced by the fact that, in the classification between emotional audio-visual stimuli differentiated in terms of the level of valence (pleasure) and arousal (excitation), no features belonging to the pupil dilation variables are exploited since they are not denoted by a significant difference. Nevertheless, in the classification between resting phase and audio-visual stimulation, and in the categorisation between videos featuring acquaintances and strangers, presents a significant difference was encountered in features extracted from the pupil dilation signal.

BUDGET

1. INTRODUCTION

The practical realisation of a project is strictly related to its financial request. In particular, it is fundamental to produce a budget estimate for the project itself in which all the financial requests are taken into account. In this study three sources of expense have been identified: the cost of employment of the people who collaborated in the project and the costs related to the used software and hardware. Finally, all the financial requests are grouped and a final budget is presented.

2. PARTIAL BUDGET

2.1. EMPLOYMENT COST

The financial budget related to the employment cost consists in the salary of three tutors and of the author of the thesis:

- Tutor: Valeriana Naranjo Ornedo, professor at the Polytechnic University of Valencia and doctor in telecommunications.
- Cotutor: Roberto Llorens Rodríguez, researcher of the Polytechnic University of Valenia and doctor in telecommunications.
- Cotutor: Anny Michelle Maza Pino, PhD student at Neurorehabilitation and Brain Research Group of the Polytechnic University of Valencia.
- Student: Alessandro Profili, student of the Master in Biomedical Engineering.

The workload, the hourly wage and the total cost for each one of the employees is shown in Table 20.

Worker	Workload (hours)	Hourly wage (€)	Total cost (€)
Tutor	30	29.5	885
Cotutor	30	23.35	700.5
Cotutor	30	17.2	516
Student	500	12.5	6250

Table 20 partial budget for the wage of each one of the employees to the project

2.2. SOFTWARE

For the realisation of the project a set of different software. First, the programming software Matlab was used for the processing and filtering of the eye-tracking data, for the implementation of statistical analysis, for the training of machine learning models and for the graphical representation of the results. The video editing software Adobe Premiere Pro was used for the editing of the audio-visual stimulation while the stimulus presentation software e-Prime was used for the design and execution of the experiment. Finally the cross-platform game engine Unity was exploited for the for the synchronisation of the eye-tracking data with the data acquired from the

other instrumentation used (electroencephalogram, electrocardiogram and fNIRS). Finally, for the drafting of the TFM Microsoft Office 365 was used. In Table 21 the partial budget relative to each one of the software used is shown.

Software	Units	Quantity	Cost per unit (€)	Total cost (€)
Matlab R2022a	year	1	840	840
Adobe Premiere Pro	month	5	48.39	241.95
e-Prime	-	-	1995	1995
Unity	year	1	399	399
Microsoft office 365	month	5	7	35

Table 21 partial budget needed for the software used

2.3. HARDWARE

The hardware exploited in this project consists in a personal computer used for the data storage, data processing and analysis and for the drafting of the TFM, and the eye-tracker HTC Vive Pro.

Hardware	Cost (€)
Personal computer	479.99
HTC Vive Pro	1373.47

Table 22 partial budget needed for the hardware used

3. TOTAL BUDGET

The total budget needed for the accomplish this project is defined by the sum of the partial budgets relative to the employment cost, the software cost and the hardware cost. The value added tax (VAT) is added to the estimated total budget. In Table 23 the total budget of the project is shown. The total budget for the accomplishment of the project is equal to sixteen thousand five hundred ninety-six and twenty-five euros.

Object	Cost
Employment cost	8351.5
Software	3510.95
Hardware	1853.46
	13715.91
VAT (21%)	2880.34
Total budget	16596.25

Table 23 Total budget for the TFM

ANNEXES

1. VIDEO RECORDING GUIDELINES

The guidelines for the recording of the videos will be provided in Spanish and consist in the following considerations:

- This interview will consist in your answers to several open questions related to the participant. As it has been previously explained, your answers will be recorded in video and they will be showed to the participant in a later session.

- The aim of this document, which is provided previously to the recording session, is that you can think on suitable answers to each question of the interview in advance. Please, try to think about answers with a minimum duration of 1 minute. There is no problem if the answer is longer.

- The information that you will provide will be confidential and it will be no judged by anyone. Please, try to answer the questions sincerely in order the participant can identify the narratives. Try to answer clearly, calmly and in a medium volume tone.

- It is preferable that you answer all the questions but, if there is any topic that you prefer not to talk about, just inform the researcher. Likewise, if you feel unable to continue the recording session, it will be temporally stopped and it will be resumed at your convenience.

- When you describe facts, try to avoid memories near the accident since these may be not remembered. However, you may speak to the participant taking into account the accident has happened.

- The questions should be answered in first person.

- Before the answers, you should say the next sentence: “Hello [name of the participant], I am [relationship with the participant], [your name].”

  E.g. “Hello Juan, I am your sister, Alicia”.
  E.g. “Hello Juan, I am mum” (you can omit your name if your relationship is unequivocal)

In the followings, the list of the questions that each one of the subject’s relatives need to answer is presented.

1) “Describe an anecdote you have with him/her. “

   It is asked to the relative to explain a funny memory, a curious or unusual event that has happened to the subject, etc.

E.g.: "When I was young, the whole family was at home and we couldn't find you and it turned out that you were hiding in the washing machine..."

2) “Describe things he/she is passionate about.”

It is asked to the relative to explain hobbies or interests that the subject particularly enjoys, such as music, sports, being with friends, or even pets. *E.g.: "What you are most passionate about is football. You used to watch all the matches and go crazy when your team scored a goal..."*

3) “Describe his/her dreams and aspirations.”

It is asked to the relative to explain what the subject wanted to do, what life plans the subject had, etc.
E.g.: "You were looking for a flat with your boyfriend because you wanted to move in together. And you were looking for one with a terrace so you could have your plants and take care of them..."

4) “Describe the positive qualities and strengths he/she has.”

It is asked to the relative to explain what are the things he/she likes about the subject that make him/her stand out from others.
E.g.: "Lucia, you are a very nice person. You are always in a good mood. You always have a smile for everyone. You are able to make anyone's day when they are sad..."

5) “Describe an achievement or act that made you proud of him/her.”

It is asked to the relative to explain how he/she felt when the subject managed to finish something that was hard to achieve or a reaction that surprised him/her for how good it was.

E.g.: "Once, when you were with friends, you found a wallet full of money, and you took it to the police so they could find the owner"

6) “Describe a memory or experience that you would like to relive with him/her.”

It is asked to the relative to explain a relevant moment, a holiday that he/she remembers fondly, a trip, or even a stage of life, and to explain how he/she felt.
E.g.: "I remember a holiday in our hometown when I was 5 years old. We used to go for a walk in the countryside, and we used to sit and have a snack next to a fountain. I would like to relive that time to see you so happy"

7) “Describe how you would like your future to be. What would you like to happen, however unlikely it may seem?”

It is asked to the relative to imagine and explain what he/she wants to happen and what

his/her dreams are, related to the participant.
E.g.: "What I would like to happen is that you wake up one day and look at my eyes. I would like to be able to be with you and walk on the beach..."

8) “What would you say to him/her if you were completely sure that he/she was listening to you?”

   It is asked to the relative to explain what he/she wants the subject to know and how he/she feels.

Finally it is provided to the subject’s relative a series of considerations about how to record the videos in order to get videos as standardised as possible. First of all, in the field of view of the camera it must appear only the person sat on a chair. His/her position must be centered, meaning that he/she must be placed in the center of the video. In the scene there must be only the person in front of a wall of a light colour, preferably white. Other objects such as tables, photos, posters, must not be present. It is preferrable to not wear nothing that could interfere with the correct visualisation and listening of the video. For example big necklaces or bracelets. For what concerns the position of the person in the scene, it is important that only the superior part of his/her body is visible (from the head to the superior part of the legs in such a way that it is possible to see the hands) in the same way it is shown in in Figure 34:

Figure 34 Example of a properly realised video

It is extremely important that the video is taken in an internal room with controlled light (distant from light sources such as windows or balconies) and without surrounding noise (people getting in or getting out from the room, noisy surrounding rooms etc.). Finally, it is asked to the person to sit on a chair without wheels as shown in Figure 35. In this way the position of the subject is as still as possible.

Figure 35 On the left, an example of appropriate chair, on the right an example of chair with wheels that must not be used

2. FILM IEQ QUESTIONS

The list of the 24 questions present in the Film IEQ questionnaire are here shown:

1) To what extent did the movie, TV show, or clip hold your attention? (1)
2) To what extent did you feel you were focused on the movie, TV show, or clip? (1)
3) How much effort did you put into watching the movie, TV show, or clip? (1)
4) Did you feel that you were trying your best to follow the events of the movie, TV show, or clip? (1)
5) To what extent did you feel consciously aware of being in the real world whilst watching? (2)
6) To what extent were you aware of yourself in your surroundings? (2)
7) To what extent did you notice events taking place around you? (2)
8) To what extent could you picture yourself in the scene of the events shown in the movie, TV show, or clip? (4)
9) To what extent did you feel like you were separated from your real-world environment? (4)
10) To what extent did you feel that the movie, TV show, or clip was something you were experiencing, rather than something you were just watching? (4)
11) To what extent was your sense of being in the environment shown in the movie, TV show, or clip stronger than your sense of being in the real world? (4)

12) To what extent did you find the concepts and themes of the movie, TV show, or clip challenging? (3)
13) To what extent did you feel motivated to keep on watching? (1)
14) To what extent did you find the concepts and themes easy to understand?(3)
15) To what extent did you feel like you were making progress towards understanding what was happening, and what you thought might happen at the end? (3)
16) How well do you think you understood what happened? (3)
17) To what extent were you interested in seeing how the events in the movie, TV show, or clip would progress? (1)
18) How much did you want the events in the movie, TV show, or clip to unfold successfully for the main characters involved? (1)
19) Were you in suspense about how the events would unfold? (1)
20) At any point did you find yourself become so involved that you wanted to speak to the movie, TV show, or clip directly? (4)
21) To what extent did you enjoy the cinematography, graphics and/or imagery? (1)
22) How much would you say you enjoyed watching the movie, TV show, or clip? (1)
23) When it was over, were you disappointed that you had to stop watching? (1)
24) Would you like to watch more of this, or similar content, in the future? (1)

BIBLIOGRAPHY

1. REFERENCES

[1] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors (Switzerland)*, vol. 20, no. 8. MDPI AG, Apr. 01, 2020. doi: 10.3390/s20082384.

[2] W. L. Zheng, W. Liu, Y. Lu, B. L. Lu, and A. Cichocki, "EmotionMeter: A Multimodal Framework for Recognizing Human Emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019, doi: 10.1109/TCYB.2018.2797176.

[3] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan. 2010, doi: 10.1109/T-AFFC.2010.1.

[4] A. M. Al-Kaysi *et al.*, "Predicting tDCS treatment outcomes of patients with major depressive disorder using automated EEG classification," *Journal of Affective Disorders*, vol. 208, pp. 597–603, Jan. 2017, doi: 10.1016/j.jad.2016.10.021.

[5] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors (Switzerland)*, vol. 20, no. 8. MDPI AG, Apr. 01, 2020. doi: 10.3390/s20082384.

[6] Sherer K. R. and Ekman P., *Approaches to emotion*, 1st ed. 1984.

[7] R. Plutchik, "The nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice.," *American Scientist*, vol. 89, no. 4, pp. 344–50, 2001.

[8] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, doi: 10.1037/h0077714.

[9] M. Murdoch, M. R. Partin, D. Vang, and S. M. Kehle-Forbes, "The Psychological Risk of Minimal Risk Activities: A Pre- and Posttest Study Using the Self-Assessment Manikin," *Journal of Empirical Research on Human Research Ethics*, vol. 14, no. 1, pp. 15–22, Feb. 2019, doi: 10.1177/1556264618810302.

[10] Mehrabian Albert and Russel J. A., "An approach to environmental psychology," *The MIT press*, 74AD.

[11] M. M. Bradley and P. J. Lang, "MEASURING EMOTION: THE SELF-ASSESSMENT MANIKIN AND THE SEMANTIC DIFFERENTIAL," 1994.

[12] Lang P. J., Greenwald M. K., Bradley M. M., and Hamm A. O., "Looking at pictures: evaluative, facial, visceral, and behavioral responses. Psychophysiology," 1993.

[13] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proceedings of ASRU 2005: 2005 IEEE Automatic Speech Recognition and Understanding Workshop*, 2005, vol. 2005, pp. 381–385. doi: 10.1109/ASRU.2005.1566530.

[14] L. Shu *et al.*, "A review of emotion recognition using physiological signals," *Sensors (Switzerland)*, vol. 18, no. 7. MDPI AG, Jul. 01, 2018. doi: 10.3390/s18072074.

[15] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, “Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?,” *Developmental Cognitive Neuroscience*, vol. 25. Elsevier Ltd, pp. 69–91, Jun. 01, 2017. doi: 10.1016/j.dcn.2016.11.001.

[16] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, *Multimodal Emotion Recognition using EEG and Eye Tracking Data*. 2014. doi: 10.0/Linux-x86_64.

[17] B. Mahanama *et al.*, “Eye Movement and Pupil Measures: A Review,” *Frontiers in Computer Science*, vol. 3. Frontiers Media S.A., Jan. 11, 2022. doi: 10.3389/fcomp.2021.733531.

[18] V. D. Costa and P. H. Rudebeck, “More than Meets the Eye: The Relationship between Pupil Size and Locus Coeruleus Activity,” *Neuron*, vol. 89, no. 1. Cell Press, pp. 8–10, Jan. 06, 2016. doi: 10.1016/j.neuron.2015.12.031.

[19] J. Kaminer, A. S. Powers, K. G. Horn, C. Hui, and C. Evinger, “Characterizing the spontaneous blink generator: An animal model,” *Journal of Neuroscience*, vol. 31, no. 31, pp. 11256–11267, Aug. 2011, doi: 10.1523/JNEUROSCI.6218-10.2011.

[20] S. Negi and R. Mitra, “Fixation duration and the learning process: an eye tracking study with subtitled videos,” *Journal of Eye Movement Research*, vol. 13, no. 6, pp. 1–15, 2020, doi: 10.16910/jemr.13.6.1.

[21] C. Ranti, W. Jones, A. Klin, and S. Shultz, “Blink Rate Patterns Provide a Reliable Measure of Individual Engagement with Scene Content,” *Scientific Reports*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-64999-x.

[22] A. Maffei and A. Angrilli, “Spontaneous blink rate as an index of attention and emotion during film clips viewing,” *Physiology and Behavior*, vol. 204, pp. 256–263, May 2019, doi: 10.1016/j.physbeh.2019.02.037.

[23] S. Mathôt, J. Fabius, E. van Heusden, and S. van der Stigchel, “Safe and sensible preprocessing and baseline correction of pupil-size data,” *Behavior Research Methods*, vol. 50, no. 1, pp. 94–106, Feb. 2018, doi: 10.3758/s13428-017-1007-2.

[24] C. Daluwatte, J. H. Miles, J. Sun, and G. Yao, “Association between pupillary light reflex and sensory behaviors in children with autism spectrum disorders,” *Research in Developmental Disabilities*, vol. 37, pp. 209–215, Feb. 2015, doi: 10.1016/j.ridd.2014.11.019.

[25] G. L. Lohse and E. J. Johnson, “A Comparison of Two Process Tracing Methods for Choice Tasks,” 1996.

[26] T. Partala and V. Surakka, “Pupil size variation as an indication of affective processing,” *International Journal of Human Computer Studies*, vol. 59, no. 1–2, pp. 185–198, 2003, doi: 10.1016/S1071-5819(03)00017-X.

[27] Babiker A., Faye I., and Malik A., “Pupillary behavior in positive and negative emotions,” *IEEE International Conference on Signal and Image Processing Applications*, pp. 379–383, 2013.

[28] Alhargan A., Cooke N., and Binjammaz T., “Affect Recognition in an Interactive Gaming Environment using Eye Tracking,” *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

[29] A. Alhargan, N. Cooke, and T. Binjammaz, “Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals,” in *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Nov. 2017, vol. 2017-January, pp. 479–486. doi: 10.1145/3136755.3137016.

[30] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, “Eye-Tracking Analysis for Emotion Recognition,” *Computational Intelligence and Neuroscience*, vol. 2020, 2020, doi: 10.1155/2020/2909267.

[31] Soleymani M., Pantic M., and Pun T., “Multimodal Emotion Recognition in Response to Videos ,” 2015.

[32] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, “Combining Eye Movements and EEG to Enhance Emotion Recognition.”

[33] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, “Eye-Tracking Analysis for Emotion Recognition,” *Computational Intelligence and Neuroscience*, vol. 2020, 2020, doi: 10.1155/2020/2909267.

[34] C. Jennett *et al.*, “Measuring and defining the experience of immersion in games,” *International Journal of Human Computer Studies*, vol. 66, no. 9, pp. 641–661, Sep. 2008, doi: 10.1016/j.ijhcs.2008.04.004.

[35] J. M. Rigby, S. J. J. Gould, D. P. Brumby, and A. L. Cox, “Development of a questionnaire to measure immersion in video media: The Film IEQ,” in *TVX 2019 - Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, Jun. 2019, pp. 35–46. doi: 10.1145/3317697.3323361.

[36] D. D. Salvucci and J. H. Goldberg, “Identifying Fixations and Saccades in Eye-Tracking Protocols.”

[37] Alastair G. Gale and Johnson Frank, *Theoretical and Applied Aspects of Eye Movement Research, Selected/Edited Proceedings of The Second European Conference on Eye Movements*. Academic Press, Elsevier, 1984.

[38] J. Llanes-Jurado, J. Marín-Morales, J. Guixeres, and M. Alcañiz, “Development and calibration of an eye-tracking fixation identification algorithm for immersive virtual reality,” *Sensors (Switzerland)*, vol. 20, no. 17, pp. 1–15, Sep. 2020, doi: 10.3390/s20174956.

[39] D. E. Irwin, “Memory for Position and Identity Across Eye Movements,” 1992.

[40] M. E. Kret and E. E. Sjak-Shie, “Preprocessing pupil size data: Guidelines and code,” *Behavior Research Methods*, vol. 51, no. 3, pp. 1336–1342, Jun. 2019, doi: 10.3758/s13428-018-1075-y.

[41] F. Onorati, R. Barbieri, M. Mauri, V. Russo, and L. Mainardi, “Reconstruction and analysis of the pupil dilation signal: Application to a psychophysiological affective protocol,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2013, pp. 5–8. doi: 10.1109/EMBC.2013.6609423.

[42] Vaibhaw, J. Sarraf, and P. K. Pattnaik, “Brain-computer interfaces and their applications,” in *An Industrial IoT Approach for Pharmaceutical Industry Growth: Volume 2*, Elsevier, 2020, pp. 31–54. doi: 10.1016/B978-0-12-821326-1.00002-4.

[43] L. Jiang, Z. Cai, D. Wang, and S. Jiang, “Survey of Improving K-Nearest-Neighbor for Classification.”

[44] W. S. Noble, “What is a support vector machine?,” 2006. [Online]. Available: http://www.nature.com/naturebiotechnology

[45] C. Cortes, V. Vapnik, and L. Saitta, “Support-Vector Networks Editor,” Kluwer Academic Publishers, 1995.

[46] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, “Applications of support vector machine (SVM) learning in cancer genomics,” *Cancer Genomics and Proteomics*, vol. 15, no. 1. International Institute of Anticancer Research, pp. 41–51, Jan. 01, 2018. doi: 10.21873/cgp.20063.

[47] F. Maleki, N. Muthukrishnan, K. Ovens, C. Reinhold, and R. Forghani, “Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment,” *Neuroimaging Clinics of North America*, vol. 30, no. 4. W.B. Saunders, pp. 433–445, Nov. 01, 2020. doi: 10.1016/j.nic.2020.08.004.

[48] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognition*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[49] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, “Combining Eye Movements and EEG to Enhance Emotion Recognition.”

[50] Institute of Electrical and Electronics Engineers, *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on : date 21-24 Sept. 2015.*

[51] A. Alhargan, N. Cooke, and T. Binjammaz, “Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals,” in *ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Nov. 2017, vol. 2017-January, pp. 479–486. doi: 10.1145/3136755.3137016.

[52] Institute of Electrical and Electronics Engineers, *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) : 23-26 Oct. 2017.*

[53] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, *Multimodal Emotion Recognition using EEG and Eye Tracking Data*. 2014. doi: 10.0/Linux-x86_64.