INVESTIGATING COGNITIVE AND EMOTIONAL FACTORS THAT TRIGGER PUPIL DILATION IN HUMAN COMPUTER INTERACTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

CEREN UYANIK CIVEK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING

JULY 2018

Approval of the thesis:

INVESTIGATING COGNITIVE AND EMOTIONAL FACTORS THAT TRIGGER PUPIL DILATION IN HUMAN COMPUTER INTERACTION

submitted by **CEREN UYANIK CIVEK** in partial fulfillment of the requirements for the degree of **Master of Science in Biomedical Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar _____
Dean, Graduate School of **Natural and Applied Sciences**

Assoc. Prof. Dr. Ergin Tönük _____
Head of Department, **Biomedical Engineering**

Assist. Prof. Dr. Didem Gökçay _____
Supervisor, **Medical Informatics**

Dr. Serdar Baltacı _____
Co-supervisor

Examining Committee Members:

Assoc. Prof. Dr. Vilda Purutçuoğlu
Department of Statistics, METU _____

Assist. Prof. Dr. Didem Gökçay
Department of Medical Informatics, METU _____

Assoc. Prof. Dr. Senih Gürses
Department of Engineering Science, METU _____

Assist. Prof. Dr. Fikret Arı
Department of Electrical & Electronics Engineering,
Ankara University _____

Assist. Prof. Dr. Selen Pehlivan
Department of Computer Engineering, TED University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: CEREN UYANIK CIVEK

Signature :

ABSTRACT

INVESTIGATING COGNITIVE AND EMOTIONAL FACTORS THAT TRIGGER PUPIL DILATION IN HUMAN COMPUTER INTERACTION

Uyanık Civek, Ceren
M.S., Department of Biomedical Engineering
Supervisor : Assist. Prof. Dr. Didem Gökçay
Co-Supervisor : Dr. Serdar Baltacı

July 2018, 105 pages

Human-computer interaction can be enhanced if emotional arousal of the user can be predicted. Measurement of pupil dilation is an effective indicator to achieve a successive classification for categorizing the psychological state of a user. In this study, rather than trying to identify several psychological states, we focused on the identification of stress. There exist several factors that shift the state of a computer user from relaxation to stress. In this study, we mainly focused on the cognitive factors that cause stress by increasing the difficulty level of a task. We also evaluated the effect of color on emotional responses. We hypothesized that assigning more difficult tasks and looking at a colored image produce higher pupil dilation signals than the signals in a neutral state. In order to evaluate the effectiveness of these factors, we conducted experiments including two phases by using TOBII T120 eye-tracker system. In the first phase, a baseline was constructed by showing neutral IAPS images to record measurements during neutral emotion. In the second phase, some modifications on stimuli were made to increase cognitive load. Pupil measurements collected during these experiments were used to train supervised classifiers for categorizing

stressful versus neutral states of the computer users. Both collective and individual subject-based analyses were performed. Better classification results are obtained for individual subject-based classification.

Keywords: Cognitive Factors, Emotional Factors, Eye-Tracking, Pupil Dilation

ÖZ

İNSAN-BİLGİSAYAR ETKİLEŞİMİNDE GÖZBEBEĞİ BÜYÜME REFLEKSİNİ TETİKLEYEN ZİHİNSEL VE DUYGUSAL FAKTÖRLERİN ARAŞTIRILMASI

Uyanık Civek, Ceren
Yüksek Lisans, Biyomedikal Mühendisliği Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. Didem Gökçay
Ortak Tez Yöneticisi : Dr. Serdar Baltacı

Temmuz 2018 , 105 sayfa

İnsan-bilgisayar etkileşimi kullanıcıların duygusal uyarılmalarının tespit edilmesi ile iyileştirilebilir. Gözbebeğindeki büyüme-küçülme reflekslerinin anlık ölçümü, kullanıcıların psikolojik durumlarının başarılı bir şekilde tanımlanıp sınıflandırılması amacıyla kullanılan etkili bir yöntemdir. Bu çalışmada, farklı psikolojik durumların tanımlanması yerine stresin belirlenmesi üzerine yoğunlaşmaktayız. Gevşeme durumundan gergin duruma geçmeye sebep olacak birçok faktör bulunmaktadır. Bu çalışmada temel olarak, bir görevin zorluk düzeyini artırarak strese neden olan zihinsel faktörlere odaklandık. Ayrıca rengin etkisini duygusal bir stres faktörü olarak değerlendirdik. Daha zor görevlerin verilmesi ve renkli bir görüntüye bakmak, nötr durumdaki gözbebeğine kıyasla gözbebeğini daha fazla büyüttüğü varsayımında bulunduk. Bu faktörlerin etkinliğini değerlendirmek amacıyla TOBII T120 göz takip sistemi ile iki aşamalı deneyler yaptık. İlk aşamada kullanıcıların duyguları nötr düzeyde tutmak için nötr IAPS görüntüleri gösterilerek bir referans oluşturuldu. İkinci aşamada, zihinsel yükü arttırmak için uyarıcı üzerinde bazı değişiklikler yapıldı. Bu deneyler

sırasında toplanan gözbebeği ölçümleri bilgisayar kullanıcılarının stresli ve nötr durumlarını kategorize etmek için eğitimli sınıflandırma algoritmalarının geliştirilmesinde kullanıldı. Hem toplu hem de kullanıcı bazlı analizler yapıldı. Kullanıcı bazlı analizlerde daha yüksek sınıflandırma oranları elde edildi.

Anahtar Kelimeler: Zihinsel Faktörler, Duygusal Faktörler, Göz İzleme, Göz Bebeği Büyümesi

To my family

ACKNOWLEDGMENTS

I would first like to present my eternal gratitude to my advisor, Assoc. Prof. Didem Gokcay, with my most sincere feelings. The door to Prof. Gokcay's office was always open whenever I had trouble or had a question about my research. She consistently allowed this paper to be my own work, but steered me in the right direction whenever she thought I needed it. I also would like to thank my thesis co-advisor, Dr. Serdar Baltaci. His priceless effort to guide me throughout my studies helped me a lot to achieve this thesis. I am very grateful for completing my degree under their supervision. I believe that it would not be possible for me to achieve this work without their excellent support.

Besides my advisors, I would like to thank the rest of my thesis committee: Assoc. Prof. Dr. Senih Gürses, Assoc. Prof. Dr. Vilda Purutcuoglu, Assist. Prof. Dr. Fikret Ari and Assist. Prof. Dr. Selen Pehlivan, for their encouragement, insightful comments, and enlightening discussions.

Also, I like to thank Fatma Gülhan Saracaydin and Ayse Elvan Gündüz for the initial data collection and the participants in the experiments, who have willingly shared their precious time.

I would also like to thank TUBITAK for supporting me through BIDEB 2210- A Scholarship Program.

Finally, I must express my very profound gratitude to my parents and to my husband for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. I will be grateful forever for your love. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ANS	Autonomic Nervous System
BDI	Beck Depression Inventory
EEG	Electroencephalography
EOG	Electro Oculography
HCI	Human Computer Interaction
IAPS	International Affective Picture System
ID	Index of Difficulty
POG-VOG	Photo-Video Oculography
QEEG	Quantitative Electroencephalography
SCR	Skin Conductance Report
SDK	Software Development Kit
SPSS	Statistical Package for the Social Sciences (Software)
SVM	Support Vector Machine
TEPR	Task-Evoked Pupillary Response
WEKA	Waikato Environment for Knowledge Analysis (Software)

CHAPTER 1

INTRODUCTION

Human Computer Interaction (HCI) is an ongoing interdisciplinary research area that interprets human actions to establish a reliable communication environment between machines and people. In the case of human-human interactions, when interpreting the actions of a human being, one should consider his/her current emotional state. This would result in more accurate responses since the emotions play an important role on human actions (Beale et al., 2008; Picard, 1997). Similarly, performance of the HCI systems can be improved, if the emotional states of the users are accurately detected (Hudlicka, 2003; Picard, Vyzas, & Healey, 2001).

One of the most dominant emotional states observed in human beings is stress. Many scientific findings indicate that stress has a vital effect on human actions, perception and rational decision making mechanism (Collyer & Malecki, 1998; Gerald Matthews, 2000b; Mandler, 1984). Therefore, stress detection using HCI would yield significant enhancements on several different application areas, such as psychological disorder identification, suspicious behavior detection or polygraphy (lie detection) (de Santos Sierra, Sánchez Ávila, Casanova, & Bailador, 2011; Hudlicka, 2002). Moreover, precise detection of stress can also lead to early diagnosis of psychological disorders (Jaimes & Sebe, 2007). For these reasons, rather than trying to identify several emotional states, we, in this study, focus on factors that are involved in the identification of stress.

Researchers have shown that facial expressions and physiological signals are the two of the most effective indicators that can be utilized to detect the stress level (de Santos Sierra et al., 2011; Cohen, Sebe, Chen, Garg, & Huang, 2003; Picard et al., 2001). Results of Dinges, Venkataraman, McGlinchey, and Metaxas (2007) and H. Gao,

Yüce, and Thiran (2014) indicated that specific facial gestures, together with the eye and mouth activities enable accurate detection of stress. However, considering the applications with security concerns, facial expressions are usually not a reliable source, since they can easily be manipulated by people (Ekman & V. Friesen, 2003; Ekman & Friesen, 1982; Surakka & Hietanen, 1998). Unlike face gestures, it is not possible to mimic or fake the physiological responses of the human body, which makes physiological signals more reliable indicators for stress measurement.

Physiological signals are the involuntary responses of the human body when faced with a stressful condition (Zhai, Barreto, Chin, & Li, 2005). The most fundamental physiological changes can be observed over the skin conductance, body temperature, brain signals and the pupillary response (Picard et al., 2001; Scheirer, Fernandez, Klein, & Picard, 2002; Rani, Sarkar, Smith, & Adams, 2003; Bradley, Miccoli, Escrig, & Lang, 2008). Among these indicators, pupillary response provides various considerable advantages. The most important superiority is that measuring the pupillary response does not require sensors to be attached to the user. Hence, this is an unobtrusive method and the measurements can be gathered without notifying the user. Another advantage is that it is not possible to inhibit, fake or exaggerate since pupillary reflex is a response of the autonomic nervous system (ANS) (Partala & Surakka, 2003). Moreover, unlike the body temperature measurements, pupillary response is not affected by variations on the environment temperature. Despite these advantages, lighting conditions have a significant effect on the pupillary response, which may cause misleading interpretation of the measurements (Bradley et al., 2008). Therefore, environment conditions should be adjusted accordingly to prevent any possible lighting interference.

It has been shown that pupil diameter increases more when faced with a positive or negative arousal compared to a neutral one (Bradley et al., 2008; Partala & Surakka, 2003). There are various forms of stimuli that would yield this positive/negative arousal, and hence, a notable change in the pupil diameter. The most effective and the practical ways to affect the pupil diameter are to use images (Bradley & Lang, 1994) and sounds (Partala, Jokiniemi, & Surakka, 2000; Pedrotti et al., 2014) as the stimuli. In this study, we employ visual stimuli to create an arousal on the subjects using the images obtained from International Affective Picture System (IAPS).

Stress causing stimuli generally originate from two different factors, which can be classified as cognitive and emotional stressors. There exist several studies that establish a strong connection between the pupillary response and the stress level. However, these studies make use of either cognitive or emotional stressors to achieve an arousal on the subject. In other words, there is no such study that uses both cognitive and emotional factors and compare the individual effects of them. Therefore, in order to eliminate this gap, we, in this thesis, investigate the marginal effects of cognitive and emotional factors which might trigger stress response in the pupils.

The Aim of the Thesis and Hypotheses

In this thesis, we aim to analyze the effects of cognitive and emotional factors on pupil dilation. By manipulating visual stimuli and the intensity of cognitive load, we wanted to detect the emotional state change, i.e., from relaxation to stress, via eye tracking technology. We will use IAPS images as stimuli while collecting the pupillary responses of the participants. Our research question and hypotheses are as follows.

Research Question: Do visual stimuli triggering cognitive load have a significant impact on subjects' physiological changes especially pupillary response?

Hypothesis 1: The intense cognitive load causes subjects' pupil diameter to increase. With the change of pupil size, subjects' emotional state can be detected.
Hypothesis 2: Color as an emotional load has an effect on pupil dilation and this affect can be detected by measuring pupillary response.

The remainder of the current thesis is organized as follows. Next chapter presents a background regarding the behaviour of the pupil, overview of stress and stress factors, studies of stress and pupillary response and the measurement techniques. Chapter 3 describes the experimental procedure, the whole process of data collection and the analysis of the dataset. This chapter outlines the detailed explanation of data preprocessing, feature extraction and classification algorithms. Experimental and statistical results are presented in Chapter 4. Classification results are discussed completely and future research directions are suggested in Chapter 5. Last chapter summarizes the study and presents a brief conclusion.

CHAPTER 2

LITERATURE REVIEW AND BACKGROUND

This chapter presents the relevant information of the literature regarding pupillary response, stress, emotional and cognitive stress factors, related works about stress - pupillary response relation and the measurement techniques used for eye-tracking. The first part describes the mechanism of pupillary response and indicates the noncognitive and cognitive effects on pupil dilation. The definition of stress and factors cause stress are discussed in the second part and human-computer interaction is mentioned briefly. The third and fourth subsections give information about the related works respectively by emphasizing the relation between stress and pupil dilation and specifying the measurement techniques.

2.1 Pupil Dilation Mechanism

There is a transparent area, called pupil, which allows the light to pass through the outer periphery of the eye and reach the retina. This, typically round, hole is located in the center of the iris. The size of the pupil is not changeless but rather controlled by muscles on the iris such that amount of the light entering the eye is regulated. Two types of muscles are involved in the process of pupil dilation, which can be classified as circular and radial groups. The circular sphincther pupillae and the radial dilator pupillae contract to decrease or dilate the size of the pupil respectively.

Autonomous nervous system regulates these muscles in such a way that rise in the sympathetic activity increases the action of dilator muscles and inhibition of parasympathetic system prevents sphincther muscle to contract (Beatty & Lucero-Wagoner, 2000). Both changes affect the pupillary response and cause dilation. Sympathetic

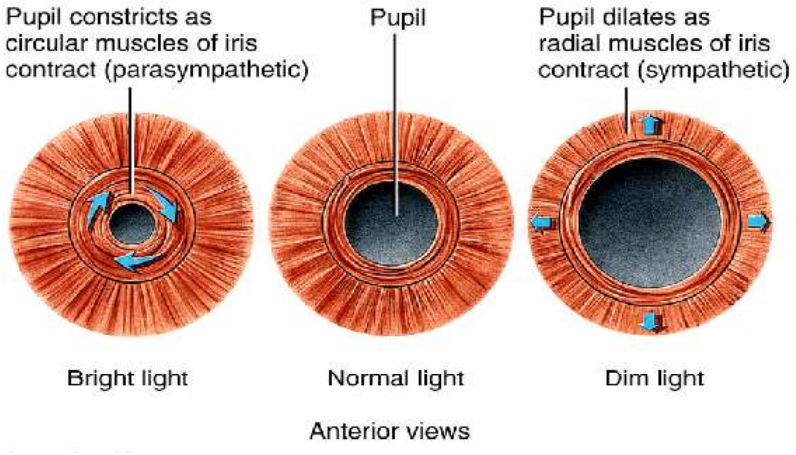


Figure 2.1: Circular and radial muscles of iris, indicating contraction and dilation (Tortora, 1987).

system is effective in fight-or-flight conditions and parasympathetic system is more related to digestion and relaxation. Although neuroendocrine and motor system are also associated with the pupil dilation, the detailed explanation of biological pathways for pupillary response are out of the scope of this study.

The pupillary response is manipulated with respect to the luminance changes. The pupil dilates in dim light and constricts in intense light to accommodate variations and this situation is named as the pupillary light reflex (Andreassi, 2006). The size of the pupil diameter is between 1 mm and 9 mm in normal conditions, the reflex to the intense light occurs in 0.2 sec and peaks around 0.5 to 1.0 sec (Beatty & Lucero-Wagoner, 2000).

In addition to the pupillary light reflex, some pupillary motions have a little impact on pupillary response. Shifting gaze (Klingner, Tversky, & Hanrahan, 2011), changes in accommodation distance (Loewy, 1990), contrast (Ukai, 1985), spatial structure (Cocker, 1996) and the onset of coherent motion (Sahraie & Barbur, 1997) affect the pupil diameter with small dilations or contractions.

Beside these noncognitive effects, the size of the pupil changes with cognitive psychology. H. Hess and M. Polt (1964) conducted a pioneer study where pupil dilations are used as an index of cognitive load. With this motivation, an approach known as task-evoked pupillary response (TEPR) which captures and evaluates the mental

workload changes during different tasks is defined (Beatty, 1982). TEPR emerges at the beginning of the processing with a short latency and stops rapidly following the completion of the event. The amplitude of TEPR is used as a psychophysiological indicator for the cognitive activity (Beatty, 1982; Beatty & Lucero-Wagoner, 2000). The cognitive load depends not only on working memory but also perception and vigilance. In the studies related to memory, mental arithmetic, decision-making and attention, the pupil dilation is exploited to validate cognitive effects (Beatty, 1982). The pupil is also sensitive to the emotional stimuli. The pupil dilates more when viewing emotional pictures compared to one that is emotionally neutral (Hess, 1972; Bradley et al., 2008).

2.2 Definition of Stress

Stress is regarded as a state in which equilibrium between physical, biological and psychological processes is disturbed, and typically cannot be controlled by organisms (Gaillard, 1993). For human beings, some general conditions such as extreme temperatures, loud noises, infectious diseases, sleep deprivation, heavy or prolonged workloads, social pressures, time pressures and negative emotions are considered as stressful. Any environmental, biological or cognitive situation that threaten the well-being of individual and diminish his resources is defined as stressor (Bourne, E, & A. Yaroush, 2003). Stressors arise from exogenous and endogenous sources which are classified as physical stressor and mental (emotional) stressor respectively. The physical stressor can be an environmental condition (cold, heat, noise) or internal physiologic demands of the human body like physical exercise which has a direct impact on the body. On the contrary, the emotional stressor has no direct physical effect on the human body when the stimuli reaches to the brain. This stimuli come from either the cognitive systems (thinking, problem solving) or the emotional system such as fear, disgust, anger (Yuen et al., 2009). The resulting stress states can be acute and time-limited or chronic based on the exposure period and intensity of the stressor. The response of an individual to stressors is generally through mental or physical effort or to present degraded performance (Bourne et al., 2003). The response to the chronic stress over time can result in decreased well-being, sleeping

problems, psychosomatic complaints and increased health risks (Gaillard, 1993).

The negative effects of stress on the quality of human life have obliged to carry out research on the analysis of human stress. The ongoing studies are more related to make machines be able to understand the stress level of the user. For this purpose, the human computer interaction systems that involve two-way exchange system with both participants are aware of each other and respond properly are utilized (Zhai et al., 2005). To measure stress in human computer interaction, eye trackers are generally used in order to detect pupillary responses.

This study measures pupillary responses via eye tracker technology in the human computer environment in order to display the influence of stressful conditions which stimulate mental stressors on pupil dilations. Mental stressors has two bases as cognitive factors and emotional factors. In this study, the main objective is to examine the effects of both cognitive factors and emotional factors on pupil dilation. Different experiments were conducted to stimulate cognitive factors and to increase cognitive load. Emotional factors were evaluated in terms of color. The cognitive factors, emotional factors and the relation between color and emotion are explained in the following subsections.

2.2.1 Cognitive Factors

Cognition is considered as mental processes including extending knowledge and comprehension through thinking, remembering, problem solving and judging. Cognitive processes are affected by attitudes, beliefs and expectations. Therefore, the cognitive understanding of an event as stressful differs from person to person. The main cognitive factors that have an impact on people's cognitive interpretations of stressors are namely appraisals, attributions and self-efficacy (Roesch, Weiner, & Vaughn, 2002).

Cognitive appraisal is the evaluation of a person regarding whether a particular encounter with the environment is convenient with his/her well-being. This term has two linked cognitive processes in order: primary appraisal and secondary appraisal. In the primary appraisal, the person assesses a stressor as either a threat or a challenge. During this decision, values, beliefs, commitments and goals of the person enable

him/her to identify the significance of well-being in stressful transactions (Folkman, S. Lazarus, Schetter, DeLongis, & Gruen, 1986). After determining this interpretation, a secondary appraisal is made in which the person decides the way to deal efficiently with the stressor such as changing the situation, accepting it or restraining from acting on impulse (Folkman et al., 1986). According to studies, negative appraisals like threat or harmful is related to negative psychological and physical modifications while the positive appraisals like challenges are correlated with the positive alterations (Roesch et al., 2002).

The second approach is attributions where the stressful event is reinterpreted or redefined based on thinking and behavior of person. According to Weiner (1985), attribution has three steps including the observation of behavior, decision of behavior and assignment of this behavior to internal or external causes. Attributions has both direct and indirect influence for the positive adjustments as attribution theory is linked to motivation and achievement. However, the uncontrolled or unstable attributions cause stress and affect the psychological and physical situations negatively (Roesch et al., 2002).

Self-efficacy, presented by Bandura (1997), is the confidence in personal ability to effectively perform challenging tasks. Emotions, cognitions, motivation and behavior of a person are affected by self-efficacy. An individual participates in tasks, explains the consequence and develops opinions of his competence within the task domain. During this process, physiological responses and negative emotional states have negative impact on self-efficacy. Thus, individuals with low self-efficacy may avoid activities for which they expect poor performance and attempt tasks that they can achieve. Additionally, low self-efficacy cause sense of failure when obstacles or unsuccessful attempts come across and withdrawal from the situation is more likely in this case (Bandura, 1994).

Various studies show that participants' pupils dilate with increasing cognitive workload being imposed (Kahneman, 1973). This effect is supported with experiments focusing on numerous tasks such as mental arithmetic (Hess, 1965), sentence comprehension (A. Just & A. Carpenter, 1993), and letter matching (Beatty & L Wagoner, 1978). In this study, the cognitive workload is evaluated by counting the number of ar-

rows presented in images. By assigning more difficult tasks such as increasing arrow numbers or destroying background so that arrows wouldn't be distinguished easily, we aim to stimulate cognitive factors of stressors.

2.2.2 Emotional Factors

Every day in human life, people are exposed to emotions. Emotion appears in a large diversity of disciplines and consists of feelings, behaviors, physiological changes. The relation between cognition and emotion is complicated which means the evaluation of an event determines the emotional response. Emotion can be considered as a chain of events while cognition is generally at the beginning of the chain. However, both processes are initialized with an external or internal stimuli (Plutchik, 2001).

Emotion is defined as a behavioral homeostatic process in which a state of equilibrium is restored when unexpected or unusual events create imbalance (Plutchik, 2001). The dynamic interaction between the world and an individual is provided by emotions. Based on this ongoing correlation with the environment, various emotions are presented as anger, envy, jealousy, anxiety, fright, guilt, shame, relief, hope, sadness, happiness, pride, love, gratitude and compassion (Lazarus, 1999). These basic emotions are generally categorized as pleasant like hope, happiness, love and unpleasant such as anger, sadness, guilt. Unpleasant emotions could initiate stress since they evoke threatening and harmful situations.

Knowing the emotion being encountered provides an information about both the individual-environment relationship and the personality trait. The same adjectives are generally used to assess both emotional states and personality traits (Plutchik, 2001). So, personality traits might be comprised of mixture of emotions. The existence of individual differences may trigger emotional stressors, and a stimulus alone is not sufficient to measure stress. To understand the relation between stress and personality characteristics an experiment was performed by Eriksen, Lazarus, and Strange (n.d.). The performance under stress was measured by giving participants false feedback that they are failing. In this case, some participants improve their results or do worse or leave the experiment to protect themselves against failure (Eriksen et al., n.d.). However, the results of the study did not identify any certain personality variable related

to stress.

Another emotional factor is physical sensations in which body reacts to harmful physical conditions. The autonomous nervous system evokes physiological arousal, especially fight-or-flight reaction reveals the emotional response to anger and fear. If the effect is intense or continued, these emotions cause stress and may harm the body (Lazarus, 1999).

The studies investigating the relationship between pupil dilation and emotional stimuli have generally compared neutral and emotional images. Categorizing an image as neutral, pleasant or unpleasant depends on its valence and arousal values. Valence demonstrates how positive or negative an image is and facial expressions are used to predict valence of an emotion. Arousal is more related to physiological features like blood flow and evaluates an image as exciting, calming or disturbing. In this study, rather than comparing neutral and emotional images we only used neutral images with both colored and grayscale versions during our experiments. With this comparison, we aim to observe the effect of color as an emotional load on pupil dilation.

2.2.2.1 The Effect of Color on Emotional Load

Living in a colorful environment can make people to neglect the effect of color on their psychological process. Perception of color can be different based on cultural beliefs, traditions and experiences. Therefore, the emotional reactions of people to color have distinguishable features. Several studies have investigated the effect of color on people's emotional reactions and found that colors have an impact on emotions.

The study of Valdez and Mehrabian (1995) examined the emotional response to color tone, saturation, and brightness by using the Pleasure-Arousal-Dominance emotion model. According to their results, blue, blue-green, green, red purple, purple and purple-blue are the most pleasant colors, while yellow and green-yellow is the least pleasant color. Similarly, green-yellow, blue-green and green are the most arousing, whereas purple-blue and yellow-red are least arousing. The study showed that brighter colors such as whites, light greys, or lighter colors are more pleasant, less arousing, and less dominance inducing than the darker colors like dark greys and

blacks. It is also said that darker colors are likely to elicit feelings that are similar to anger, hostility, or aggression.

In the study of Sroykham, Wongsathikun, and Wongsawat (2014) the measurements of five participants for oxygen saturation (SpO2), pulse rate and quantitative electroencephalography (QEEG) in six colors (white, blue, green, black, yellow and red) are used to evaluate the effect of colors. Based on brain signals, the results show that the brain activity when seeing red or yellow is higher than when perceiving blue, green, white and black respectively. In terms of color, the study reveals that red and green color have a high impact on vitality. Red and yellow stimulate anger and confusion while green color stimulates vigor mostly. Blue color has a moderate effect on confusion, tension and fatigue. Unlike the study of Valdez and Mehrabian (1995), this study claims that white and black colors have a low impact on any mood.

The emotional responses on colors were also investigated in the study of Ou, Luo, Woodcock, and Wright (2004). An experiment was conducted with 31 observers, including 14 British and 17 Chinese, to evaluate 20 colors in 10 color emotion sizes such as hot-cold, heavy-light, active-passive, hard-soft, masculine-feminine and like-dislike. This study maintains the relation between color and emotions. It also demonstrates that emotions and perception of colors may be culture-independent.

Similar to the previous study, X. Gao and Xin (2006) examined the color emotion model. A total of 218 color samples were evaluated by 70 subjects based on 12 basic descriptive variables such as "hot-cold", "weak-strong" and "dynamic-passive". It is found that while defining the emotional meaning of color hue, one of the color perception attributes, has less influence than chroma and lightness. Based on their results, colors with high chroma and high brightness are soft and warm colors while low chroma and low brightness are related to dark colors which brings cool and hard emotional descriptions.

Apart from these studies, Young, Han, and Wu (1993) compared the pupillary responses induced by heterochromatic and achromatic brightness increments in order to investigate whether color and luminance-evoked pupil responses independent of each other. It was found that color and brightness is not inevitably independent on pupillary responses.

Based on these findings, it can be said that the effect of color as an emotional factor can be observed from the pupillary responses. With this motivation, this study compares the pupillary records during the view of colored neutral images and the pupillary records during the view of grayscale neutral images to observe the difference caused by color.

2.3 Studies Regarding Stress Level and Pupillary Response

Finding the factors that cause stress is an ongoing research area in many disciplines. To measure the stress level, one of the easy techniques is to observe pupillary response of subjects. With the help of this correlation, many studies have been investigating the connection between the stress level and various stressors. The following studies mainly focused on pupil dilations to examine the association with the stressors.

In the study of Kinner et al. (2017), the psychophysiological connections of two different cognitive emotion regulation strategies namely reappraisal (increase and decrease) and distraction is investigated via the assessment of pupillary responses, skin conductance reports and subjective emotional responses during an emotional picture viewing task. Neutral and negative pictures are presented in five experimental conditions: view neutral, view negative, decrease, increase and distract. Pupil dilation and SCRs are higher when viewing negative images compared to neutral images. Also, increasing emotional responses to negative pictures cause larger pupil diameter which means pupil diameter is initially related to the extent of mental effort but it is modulated by emotional arousal.

The interaction between emotional stimuli and pupil dilation is studied by Snowden et al. (2016). Fearful and neutral images are compared during three different tasks. The mean luminance, mean contrast, image color and complexity of content are equal for all images to prevent misleading results. The tasks are assessed based on the target duration, repetition of the same stimuli and actively naming the emotion of the stimuli. As a result, the pupil is more dilated after viewing affective pictures, and this effect is independent of the presentation time of the images, not reduced by repeated presentations of the images, and not affected by active processing of the emotional

content of the images compared to passive viewing. For the result of last task, the authors have stated that the emotional modulation of the pupil may be automatic and controlled unconsciously.

Plechawska-Wójcik and Borys (2016) have focused on the hypothesis that additional biomedical data like pupil response and blink features can support the traditional EEG-based cognitive workload analysis in their paper. A case study consisting of three arithmetic tasks is carried out with three participants in order to obtain possible parameters used in machine learning classification systems. Each task has a different difficulty level to analyze the human cognitive process. The EEG data analysis reveals that the change of mental effort manipulates the alpha and theta bands. For the pupillary response side, the increasing complexity of tasks cause pupil dilation to decrease over time. Finally, the authors have found that combining EEG with other biomedical indicators might improve the performance of cognitive load data analysis.

The pupil dilation and periorbital temperature data are used in the study of Baltaci and Gokcay (2016) to detect the stress of a user. An experiment is performed by first showing neutral images and asking participants to count arrows located on the images. Then, negative images are shown to the participant which have more arrows and when the participant gives an answer, the experimenter sometimes responds with a misleading feedback. With these factors, both emotional and cognitive stress factors are manipulated. They extract features from both pupil and thermal data to use in classification algorithms. As a result, this study shows that pupil dilates more after viewing negative images and handling a more difficult task.

The relation between the changes of pupil size and the difficulty levels of a visual motor task is examined in the study of Jiang, Zheng, Bednarik, and Atkins (2015). The subjects perform a simple continuous aiming task while the task requirement is manipulated and measured by Fitts' Index of Difficulty (ID). The ID is a model between environmental stimuli and human response where task requirements are defined as easy or difficult based on the increasing ID. The participants move a surgical tool continuously to point to the circles from bottom to top and then from top to bottom. During the experiment, this continuous aiming movement is calculated and divided as transport where tool leaves for the target circle and landing phases where tool reaches

and touches the target circle. According to results, the pupil constricts in the transport phase and dilates in the landing phase. Additionally, higher task difficulty causes higher pupil dilation and longer peak duration.

Emotion recognition based on pupillary response is achieved with the neural networks system in the study of Aracena, Basterrech, Snáel, and Velásquez (2015). An experiment is conducted on four participants by showing several neutral and emotional images. This study also supports that pupil dilates more after viewing emotional images and it compares pupillary response in terms of neural network model. Based on subject-dependent analyses, it is said that emotional stimuli increase emotional arousal and cause pupil dilation.

The study of Pedrotti et al. (2014) concerns stress detection by recording pupil diameter and electro dermal activity during a simulated driving task. First a baseline run is conducted and then three stress runs are performed together with sound alerts. Wavelet multiresolution decomposition and neural networks are used for feature extraction and classification. The authors have concluded that pupil diameter strongly correlates with stress detection.

The study of Y. Gao, Adjouadi, Ren, and Barreto (2013) investigates the emotional state of a computer user as it transforms from relaxation to stress. They use pupil dilation signals which are pre-processed by Kalman filtering, Wavelet denoising and Walsh transform. From five different classification algorithms, they obtained similar results which indicates that pupil dilation signals are a dominant factor to determine the emotional state of a user.

In the study of Klingner et al. (2011), the effect of auditory versus visual task presentation on pupil dilation is examined. Three tasks are performed to compare visual and auditory presentation: mental multiplication, digit sequence recall and vigilance. Solving mental arithmetic problems depends on working memory and digit sequence recall is used to investigate both short-term memory and long-term memory recall. However, vigilance is less dependent to working memory and it mostly requires attention and motor responses. Based on these experiments, the magnitudes of pupil response are higher for auditory compared to visual tasks, although the patterns of dilation for both types of presentation are similar. So, this study proposes that the

visual task presentation results in lower cognitive load than auditory presentation in all three cases.

Zhai and Barreto (2006)'s study aims to improve human-computer interaction by showing a correlation between emotional states and pupillary response. In this study, the Stroop Color-Word Interference Test is conducted on six participants. The blood volume pulses, galvanic skin responses and pupil data of the subjects are analyzed. The results of this study reveal that the stress state of users are detectable from these physiological data.

2.4 Measurement Techniques for Pupil Dilation

Capturing and measuring the size of the pupil could be challenging in some conditions. To deal with such problems, some studies proposed different measurement techniques for pupillary response. Although diverse methods are used to compute pupils' size, the relation between stressors and pupillary response is coherent in each work.

Marshall (2002) introduces a novel technique called the Index of Cognitive Activity which used as a general psychophysiological measurement of cognitive workload. This method is utilized as a wavelet decomposition to the pupil size signal in order to estimate the average number of abrupt discontinuities in pupil size per second.

In the study of Pomplun and Sunkara (2003), a neural-network based calibration interface technique is presented for video-based eye trackers. During eye tracking experiments, the participant's gaze angle affects the measured pupil size and numerous small factors can interfere with pupil dilation. With this measurement method, the authors have aimed to eliminate the various factors affecting pupil dilation and geometry-based distortion of pupil size. They have also computed the cognitive workload with different display brightness by performing an additional calibration process.

Apart from these novel techniques, some accepted methods such as electro-oculography (EOG), scleral search coils, photo-video oculography (POG-VOG) and dual purkinje method are mostly used for eye tracking.

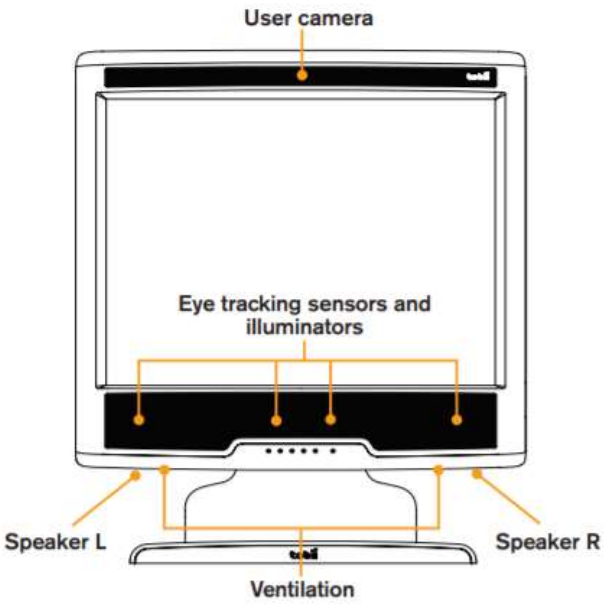


Figure 2.2: Dual purkinje eye tracking method (Tobii T60 & T120 eye tracker front display , (Tobii Technology, 2011))

Tracking of eye movements is achieved in EOG by measuring electrical potential differences between electrodes placed near the eyes. The eye movements are detected in scleral search coils method via wearing contact-lens-like material with a metal coil in it. The fluctuations in an electromagnetic field while a metal coil is moving along with the eyes identify the eye movements (Duchowski, 2007). In photo-video oculography, digital video cameras capture the eye movements. Among these methods, the most recently used one is the dual purkinje method which consists of an infrared camera placed below a monitor of a desktop computer and a special software in this computer (see Figure 2.2). The infrared camera emits infrared light to the eye and when the light enters the retina, it creates the corneal reflection while most of it is reflected back. These reflections cause a bright pupil effect that helps software system for the detection of eye. When the software recognizes the center of the pupil and the corneal reflection, it measures their distance and the point of fixation can be determined. The important issue about finding the point of fixation is to distinguish eye movements from the head movements. For this separation, pupil brightness plays a significant role (Duchowski, 2007).

In addition to fixation measurement and eye movement tracking, eye trackers are able to measure pupillary responses. Diverse methods such as entoptic methods, mir-

ror comparison, scales and callipers, filming, Bellarminow apparatus, Lowenstein pupillograms and infrared photography can be used to measure the size of pupils (Hakerem, 1967). However, current eye trackers calculate the pupils' size with two methods namely pixel-counting and ellipse-fitting. In the pixel counting method, the size of pupils are determined by counting the number of pixels in the pupillary area. In the ellipse-fitting method, the calculation is done by taking the point of reference as the length of the major axis of an ellipse fitted to the pupil (Klingner, Kumar, & Hanrahan, 2008). TOBII eye tracker system measures the pupil size by pixel counting method.

CHAPTER 3

METHODOLOGY

This chapter covers the scale-based material, the stimuli generation process, experimental design, data analysis and classification methods. The experiments are designated to particularly focus on the cognitive and the emotional factors causing stress. There are two phases in both Experiment 1 and Experiment 2, where the pupillary responses of the participants are recorded via eye-tracker system for neutral and stressful states respectively. In the first phases, a baseline is created to achieve a controlled experiment by ensuring that the participants are in neutral condition. The second phases include different tests to observe the marginal effects of different stressors. The datasets are pre-processed and several features are defined to evaluate the results with state-of-the-art classification methods.

3.1 Scale-Based Materials for Data Collection

In the context of this study, two different scale-based materials have been employed, which are Beck Depression Inventory and Positive and Negative Effect Scale.

3.1.1 Beck Depression Inventory

The Beck Depression Inventory (BDI), named by its creator Dr. Aaron T. Beck, is applied to individuals to determine whether they are in a depressive situation (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). The inventory contains a set of 21 multiple choice questions, where each answer is assigned to a value in between 0 to 3. These values are used as indicators to measure how depressive a given answer is, i.e.,

higher values means more depressive responses. With this given structure, the possible total scores are in the range of 0 to 61. The individuals obtaining higher scores are accepted as more prone to show depressive behaviors, while the ones getting lower scores are seen as psychologically healthy. Overall, the content of the questions aims to measure different symptoms of depression like hopelessness or irritability.

This inventory was adapted to Turkish nation and validated by Hisli and Sahin in 1988 and 1992 respectively. For the Turkish version of the inventory, the individuals obtaining scores lower than 17 are assumed to have minimal positive signs of depression. The complete 21 questions for Turkish version are provided in the Appendix A.

In our experiment, the participants were asked to answer each of the questions. At the end of the experiment, results of the participants with BDI scores above 17 were eliminated to ascertain that the participants who are admitted to the experiments are not depressed. Because depression state may interfere with the emotional manipulations of the experiments.

3.1.2 Positive and Negative Affect Scale (PANAS)

The Positive and Negative Affect Scale, developed by Watson, Anna Clark, and Tellegen (1988), is a standardized test for the evaluation of mood or affect. The test consists of 20 items in which 10 items belong to positive impact and the other 10 items are about negative impact. The positive set shows the level of alertness and activeness of a participant while negative set demonstrates the level of encountered anger, guilt and fear. The modification and normalization of this scale in Turkish community was introduced by (Gençöz, 2000).

In the test, participants rate each emotional item on a 5-point Likert type scale where 1 means not at all and 5 is for extreme. For positive set, the values on the items 1, 3, 5, 9, 10, 12, 14, 16, 17 and 19 are summed and for the negative set the remaining items are added together. The total score for each set varies from 10 to 50 where higher values imply higher positive/negative impact. The evaluation of mood is done by comparing these PA and NA values. For positive mood, PA value should be greater than NA

value and for negative mood, the reverse case should be observed. The version used in this study is provided in Appendix B.

In this work, PANAS test was completed by participants between the first phase and the second phase of the experiment. The mood of the participants during the experiment is crucial as the aim is to observe the cognitive and emotional stress factors. Thus, if the PA score of a participant is below 20 or if the NA value is above 20, the data of this subject cannot be valid and should be removed.

3.2 Stimuli Creation

Experiment 1 and 2 used images from IAPS database as stimuli. The adjustments on the size and the modifications on the stimuli are described in the subsequent parts.

3.2.1 IAPS Images

The International Affective Picture System (IAPS) has been a reference for objective emotional evaluation based on visual stimuli. A large database of photos are developed; pictures represent a series of daily experiences to extreme scenes, such as furniture and deformed bodies. Each image is rated according to a subjective score given by a large group of people on a scale which ranges between 1-9. The average of these views is the score of images used in image qualification which means the pictures are classified according to the average of these scores. IAPS provides a way of assessing emotions and is commonly used in emotional experiments (J Lang, M Bradley, & N Cuthbert, 2008). In our experiments, the visual stimuli were taken from the IAPS database and they were chosen to have low arousal and neutral valence values, which are in the ranges of 2.77 ± 1.896 and 4.949 ± 1.185 respectively.

3.2.2 Image Rescaling

For each stimulus, a 3×3 grid consisting of the selected 12 images was created by randomly positioning them. The 3×3 grid structure was chosen due to the exper-

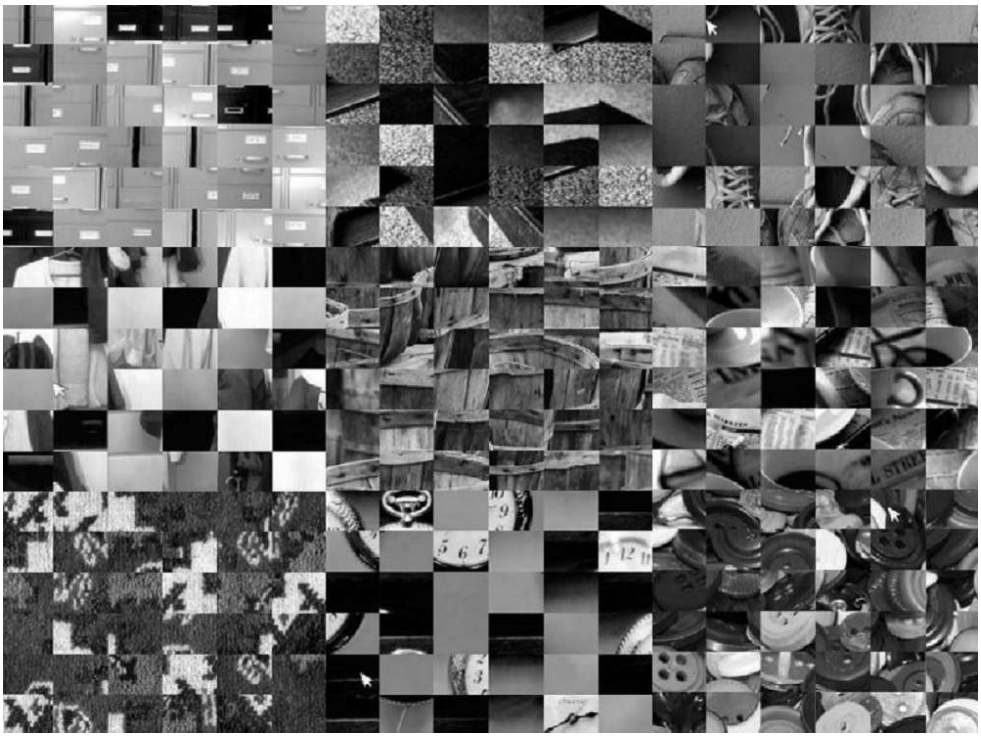
Figure 3.1: Sample stimulus used in Experiment 1

imental results of the study of Baltaci and Gokcay (2016). In order to create this structure, all images were rescaled to smaller sizes and then, combined randomly in a 3×3 grid structure. As a result, from 9 random small images, one large stimulus was formed. 20 different stimuli were obtained using this process. All formed images were rescaled to standardize the size of images. The average width and length of all stimuli were 1000 and 750 pixels, respectively. A sample visual stimulus used in Experiment 1 can be seen in Figure 3.1.

3.2.3 Image Scrambling

The stimuli used in the second phase of Experiment 2 are the scrambled form of the images shown in the first phase of same experiment. To form scrambled images, jigsaw method was executed in MATLAB 2016b and 3×3, 4×4, 5×5 and 6×6 grids were tested separately. The aim is to destroy the image content and to increase participants' alertness. Therefore, 6×6 grid was chosen to design the scrambled images. Both a sample image used in Experiment 2 and the corresponding scrambled version can be seen in Figure 3.2.

(a) Grayscale image



(b) Scrambled form of the same image with 6x6 grid

Figure 3.2: Sample stimuli used in Experiment 2

3.2.4 Intensity Adjustments

The pupils are affected by the brightness and illumination of visual stimuli so the mean intensity values of all images were standardized. The average intensity values of all images were calculated and the outlier values were set to values similar to the average intensity values by using Adobe Photoshop CS6. For Experiment 1, the average intensity of the image set was 96.93 and standard deviation was 0.05. For Experiment 2, the average intensity of the image set was 103.29 and standard deviation was 5.72.

3.2.5 RGB to Grayscale Adjustments

In the Experiment 2, grayscale images were used as the stimuli set. The three colored (RGB) images used in the Experiment 1 were converted into grayscale.

3.3 Experimental Design

The objective of this study is to examine the effect of cognitive and emotional stress factors on pupillary response. To compare the cognitive and emotional factors, two different experiments were conducted. Both experiments had the same apparatus TOBII T120 and followed the similar procedures. However, different stimuli sets were presented to different participants.

3.3.1 Apparatus

A screen-based TOBII T120 was used for the eye tracker system to obtain undistorted pupil data during the experiments. The participants were located in front of a computer screen, where TOBII T120 system is embedded in it. The usage of TOBII technology enabled to not use head restraints. The distance between the participants and the screen was determined as between 0.6-0.7 meters. The pupillary responses were collected at a rate of 60 Hz. The experiments were conducted under the control of Human Computer Interaction (HCI) Laboratory in Middle East Technical Univer-

sity Computer Centre and all continuously recorded pupil values were saved as text files by TOBII Eye Tracking System (TETS) on a 17" TFT monitor with Windows 7 based computer.

3.3.2 Environment

The experiment room was sufficiently illuminated by using a stable light source. In addition, the room temperature was designated to be in between 19-22 ^{o}C. The humidity was also regulated by humidity control and air recycling systems.

3.3.3 Experiment 1

Experiment 1 aims to observe the effect of cognitive load on pupil dilation by administring a task to the subject and increasing his/her attention.

Participants: 16 subjects (6 males, 10 females) between ages of 21-32 (M = 25.4, S.D = 2.96) agreed to participate in the experiment. They had normal or corrected to normal vision, and did not report a neurological or endocrine disease story. All participants read and signed written informed consent (see Appendix C). Since participants had BDI scores of 17 or less, none of the participants were considered as outliers in terms of depressive situation. Since the PANAS scores were acceptable for positive and negative affect scores (for positive influence: higher than 20, for negative influence: below 20), none of the participants were considered to be excluded in terms of mood.

Procedure: Participants were placed in the HCI Laboratory for the experiment. Before a brief explanation of the experimental procedure, participants filled out the BDI for psychiatric evaluation, read and signed written consent and completed demographic information (see Appendix D). It generally took 10 minutes to finish this process.

Prior to the start of the experiment, we applied 9-point calibration of TETS provided by the TOBII T120 Software Development Kit (SDK). On the calibration screen, yellow dots appear one by one and the participant must follow these points. After all the

points are visible, the SDK sends a signal whether the calibration is acceptable or not. For this reason, the user can continue with the experiment or restart the calibration.

There were two phases to assess the effects of stressors in Experiment 1. The first phase of the experiment was designated to create a baseline for the standard pupil size of the participants when they were in neutral state. To this end, we avoided to put the participants in a stressful situation during this phase. We presented the instructions of experiment before starting. 20 sample images on the center of gray background (R: 106 G: 106 B: 106) were displayed to the participants. In addition, a total number of 3-6 arrow figures were added to some random positions in each image. While viewing the images, the participants were asked to count the total number of arrows they saw on the stimulus. Each stimulus was presented for 6 seconds. The order of presentation of the stimuli was randomly selected. A fixation point between images was presented for 12 seconds and during this time, we expected participants to verbally express their responses. We gave them a feedback showing the correctness of the response immediately after the responses. If the answer is correct, the feedback was only a "True" statement. In the other case, we said "False, the number of arrows is x". Fixation durations and pupil diameters were collected throughout the experiment, but only the data for the stimulus presentation was used for the analysis.

After the completion of the first phase, the participants were asked to fill out PANAS for mood assessment. The 9-dot calibration of TETS was reapplied before starting the second phase.

The second phase of the experiment aimed to enhance the effects of cognitive stress sources. In the second phase, the only difference was the number of arrows the participants had to count on each trial. We increased the total number of arrows to the range of 6-9 in each stimulus. Since the images were still displayed for 6 seconds, the counting task became more difficult and stressful for the participants. The corresponding pupillary responses and fixation durations during the experiment were collected but only the pupil data during stimulus presentation was analyzed to evaluate the effect of assigning more difficult tasks to a subject.

After the completion of the entire experiment, participants were asked to complete a questionnaire to receive their views on the experiment and to compare the declared

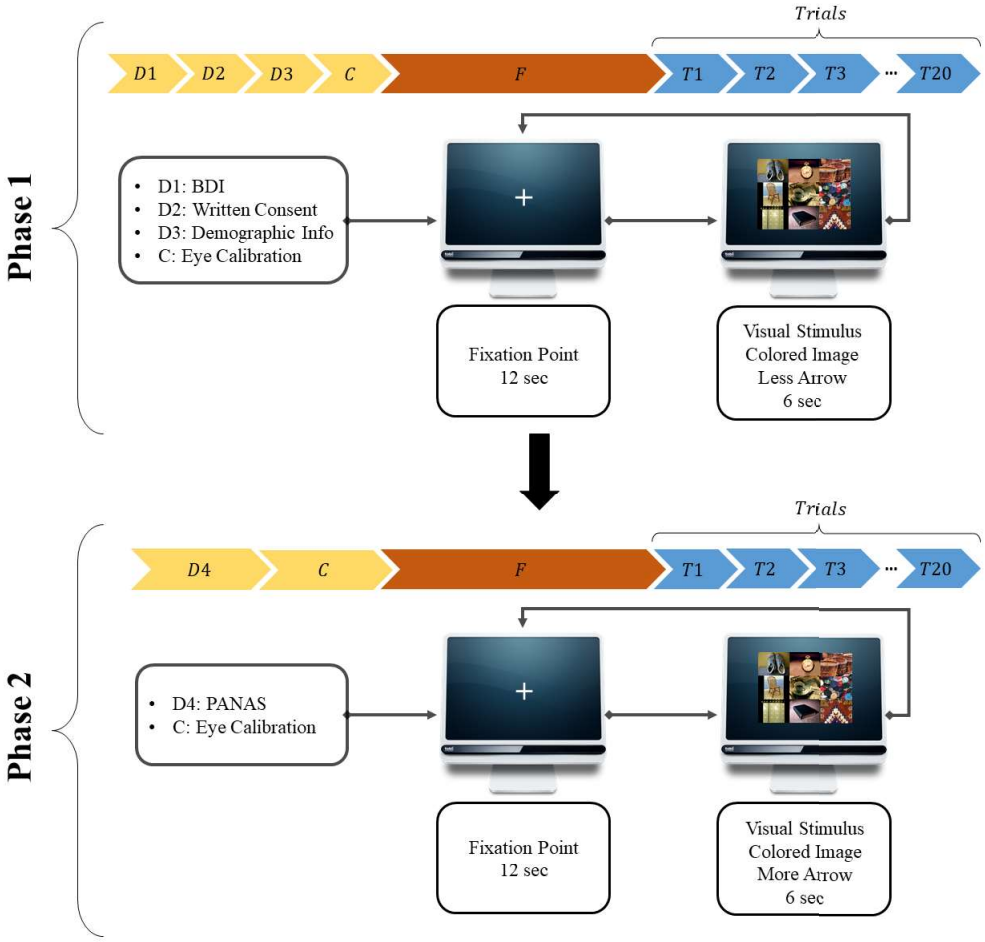


Figure 3.3: The flow of Experiment 1

mental status of the participants with the measured values (see Appendix E). Survey questions were designed to assess participants' mental status after exposure to stress factors. The flow of the Experiment 1 can be seen in Figure 3.3.

3.3.4 Experiment 2

Experiment 2 examines the effect of cognitive load on pupil dilation like the previous experiment. However, a different stimuli set was used in this experiment.

Participants: 15 subjects (6 males, 9 females) between ages of 21-29 (M = 24.7, S.D = 2.23) agreed to participate in the experiment. They had normal or corrected to normal vision, and did not report a neurological or endocrine disease story. All participants read and signed written informed consent. Since the PANAS scores were

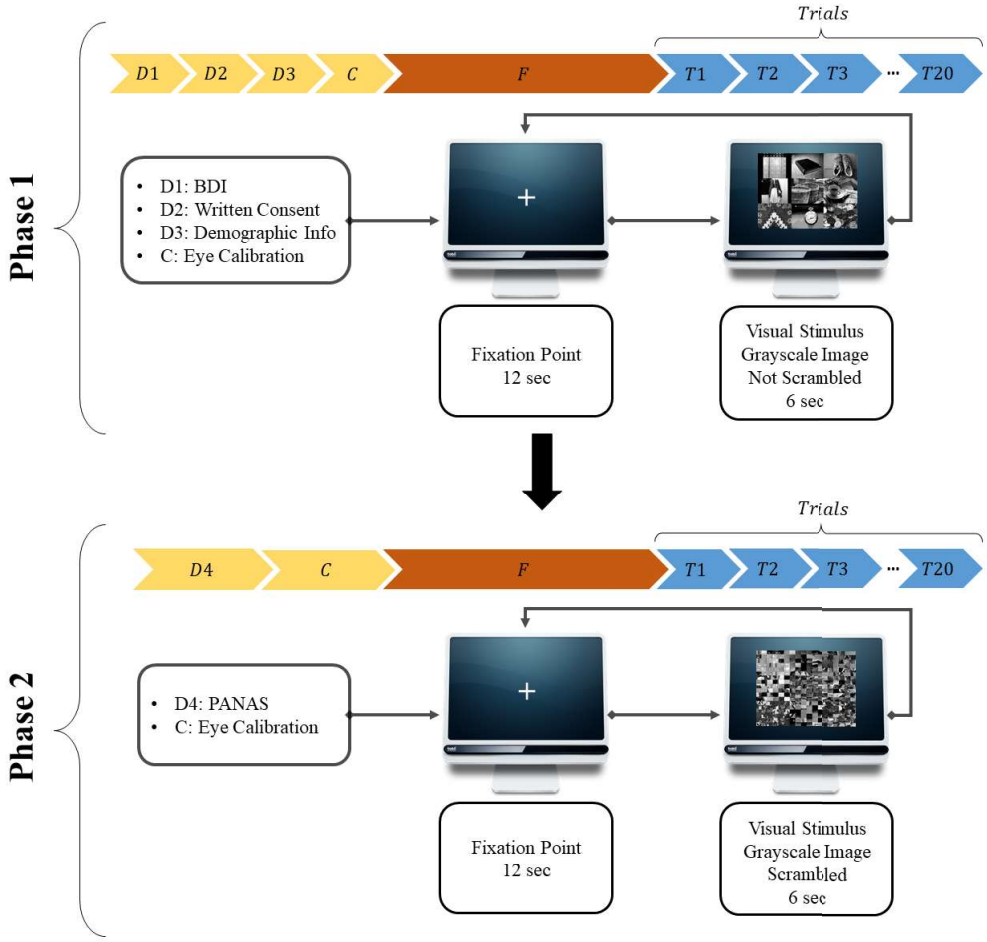


Figure 3.4: The flow of Experiment 2

acceptable for positive and negative affect scores (for positive influence: higher than 20, for negative influence: below 20), none of the participants were considered to be excluded in terms of mood. However, in terms of depressive status, one participant had a BDI score higher than 17 so this subject was eliminated before analysis.

Procedure: Participants were placed in the HCI Laboratory for the experiment. We followed the similar procedures like in the Experiment 1. Participants filled out the BDI, read and signed written consent and completed demographic information before a brief explanation of the experimental procedure. It generally took 10 minutes to finish this process. Prior to the start of the experiment, 9-point calibration of TETS was applied.

The experiment has two phases to assess the effects of stressors. The first phase of

the experiment was similar to the first stage of Experiment 1. The only difference was that RGB images were converted to grayscale images. The instructions were presented by the experimenter before starting. 20 sample grayscale images on the center of gray background (R: 106 G: 106 B: 106) were shown to the participants. In addition, a total number of 3-6 arrow figures were added to some random positions in each image. While viewing the images, the participants were asked to count the total number of arrows they saw on the stimulus. Each stimulus was presented for 6 seconds. The order of presentation of the stimuli was randomly selected. A fixation point between images was presented for 12 seconds and during this time participants were invited to verbally express their responses. A feedback showing the correctness of the response is given immediately after the responses. If the answer is correct, the feedback is only a "True" statement. In the other case, the expression was "False, the number of arrows is x". Fixation durations and pupil diameters were collected throughout the experiment, but only the data for the stimulus presentation was used for the analysis.

After the completion of the first phase, the participants were asked to fill out PANAS for mood assessment. The 9-dot calibration of TETS was reapplied before starting the second phase.

The purpose of the second phase of the experiment is to highlight the marginal effects of cognitive stressor on pupil dilation. In the second phase, the images were scrambled and the number of arrows the participants had to count on each trial wasn't changed. Since the images were still displayed for 6 seconds, the counting task became more difficult and stressful for the participants as the visual background was converted a more complex image. The corresponding pupillary responses and fixation durations during the experiment were collected but only the pupil data during stimulus presentation was used for examination.

After the completion of the entire experiment, participants were asked to complete a questionnaire to receive their views on the experiment and to compare the declared mental status of the participants with the measured values. The process of the Experiment 2 can be seen in Figure 3.4.

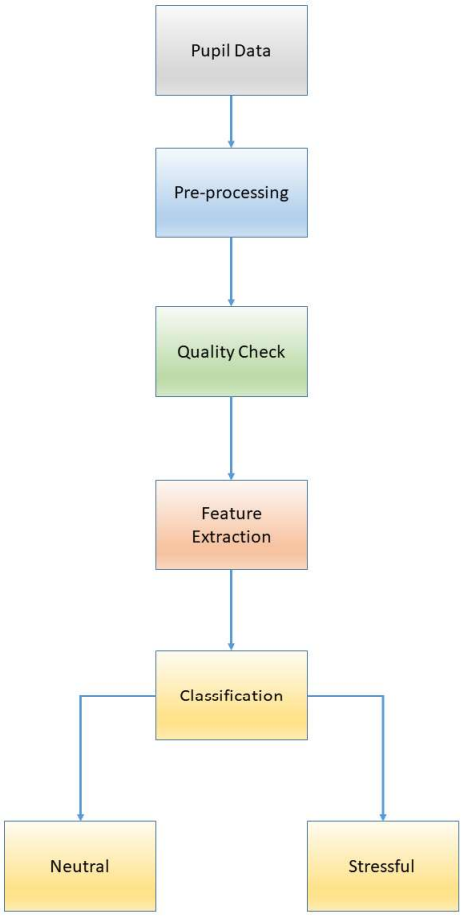


Figure 3.5: General flow of data analysis

3.4 Data Analysis

In this section, we present the analysis process for the data collected from both Experiment 1 and Experiment 2. We formed 4 different datasets from these two experiments. The first dataset, labeled as Dataset 1.1, contains the pupillary responses of the subjects for the Part 1 of Experiment 1. The data in this dataset correspond to a neutral state, where the colored stimuli with less number of arrows were shown to the subjects. The second dataset, i.e., Dataset 1.2, includes the pupillary responses from the Part 2 of Experiment 1, where the same colored images with more number of arrows were shown to the subjects. These two datasets will be used to detect the possible effects of cognitive stress factors, i.e., increasing the difficulty of the task, on pupil dilation. The last two datasets, namely Dataset 2.1 and Dataset 2.2, ob-

tained from the Part 1 and the Part 2 of Experiment 2 respectively. Thus, Dataset 2.1 contains responses from the subjects when the grayscale stimuli with arrows were shown to them. Different from Dataset 2.1, Dataset 2.2 represents the case where the grayscale images were scrambled without changing the number of arrows. Similarly, these two datasets will also be used to observe the effect of cognitive stress factors on pupil dilation. In addition, responses obtained from Dataset 1.1 and Dataset 2.1 will be compared to examine the effect of color as an emotional load on pupil dilation.

It is expected to observe a notable difference between the above mentioned datasets if there is a correlation between the pupillary response and the cognitive stress factors. In other words, the datasets corresponding to the neutral and the stress cases are expected to be separable. Therefore, in this study, we employed several classification algorithms to determine the separability rate of these datasets and proved/disproved our hypotheses based on the classification accuracies.

The datasets obtained from the experiments are composed of raw pupillary data. There exist issues on these raw data such as missing data points, misalignment, noise etc. that prevent them to be directly fed into a classification algorithm. Moreover, we need to form useful features before initiating the classification process. Therefore, we first performed a preprocessing stage followed by a quality check to remove useless data. We then generated specific features to be used in the classification process in the feature extraction stage. Finally, we used several classification algorithms to obtain the results. We worked on WEKA platform to realize the classification process. The flow of data analysis process is represented in Figure 3.5.

3.4.1 Data Preprocessing

A preprocessing stage was employed on the measurements to eliminate the issues on the raw data. Several problems like missing data points, high variations on the data and two observations obtained from left and right pupil can be seen from the Figure 3.6. This stage was performed by implementing the steps given in Baltaci and Gokcay (2016). The flow of the procedure is presented in the Figure 3.7. The detailed procedure is explained as follows:

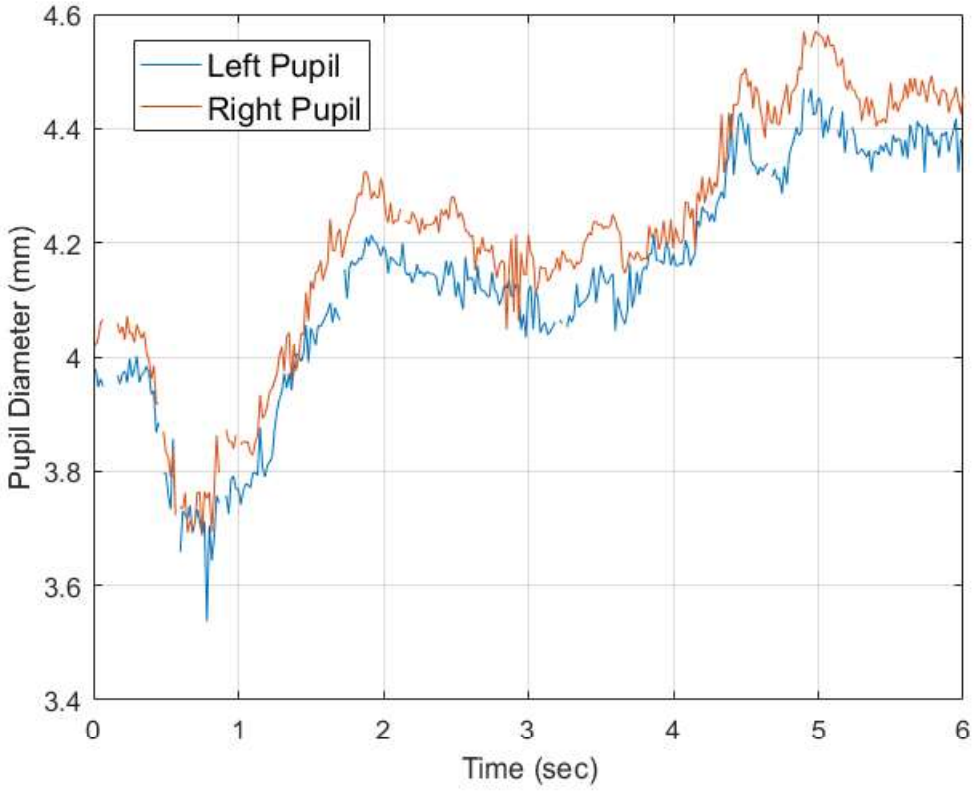


Figure 3.6: Pupillary Response of One Random Trial Before Preprocessing

1. Blink Extraction: The eye-tracking device records the measurement of pupil size as -1 when it detects blink. If we observed -1 in a pupil record of a single eye, we replaced this missing data with the measurement from the other eye to avoid this problem. Otherwise, we performed linear interpolation by using both the average of the last five instances before the missing data and the average of the first three instances after the missing data. Then, during the blinking period, we set the pupil size to the value given by this interpolation function. If more than 30% of the data had to be interpolated, it was discarded since it is considered as unreliable.

2. Merging Left and Right Pupil Data: As a raw data, we had both the left and right pupil diameter measurements. It is expected to have a correlation between left and right pupil data as healthy people have same pupil sizes. Therefore, we first checked the correlation between them and if they were not correlated we omitted that data. We used a threshold value of 0.9 for acceptable correlation. Data that passed the correlation test is merged by averaging pupil values from both left and right eyes.

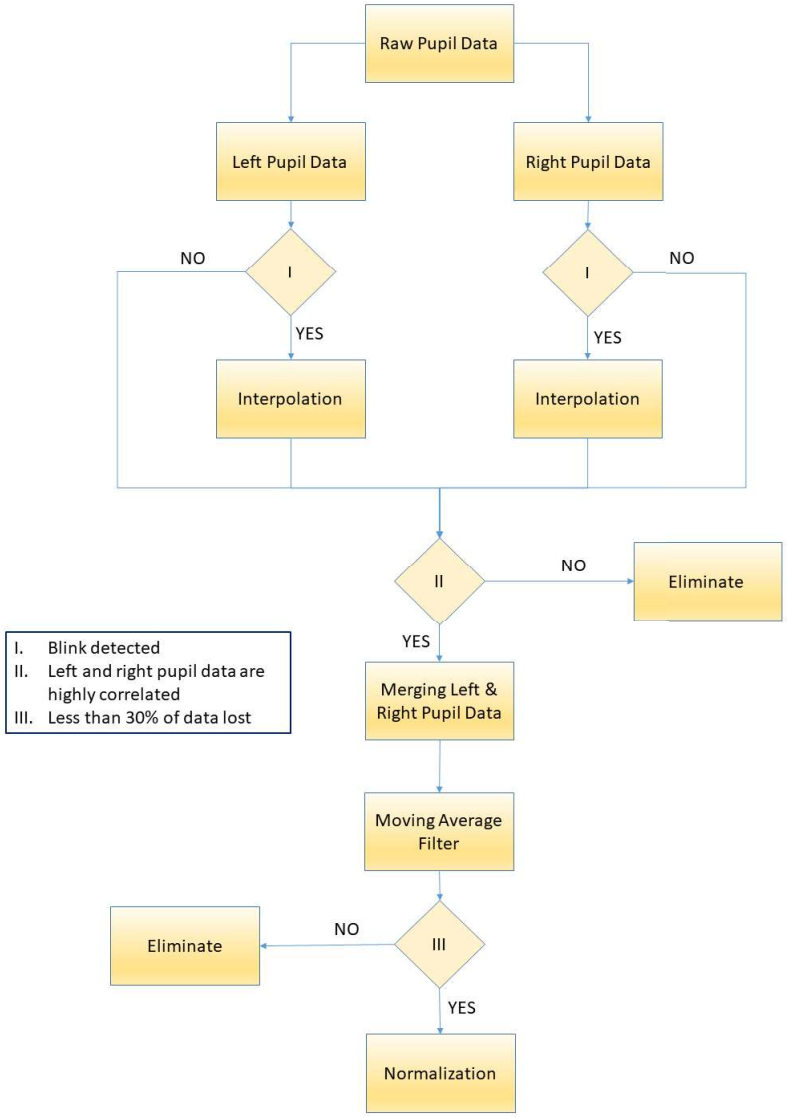


Figure 3.7: The flow of data preprocessing

3. Moving Average Filter: For noise cancellation, we used the moving average filter (Smith, 1997-98) where a small window with 20 time points gave the optimal result for our analyses (see Table 4.8, 4.17, 4.23). Here it is important to compare the data without filter and the filtered data because erasing many points from the data may change the original pattern and cause misinterpretation. Therefore, we assigned an exclusion criteria that if more than 30% of pupil data was lost, this record was eliminated completely.

4. Normalization: The eye tracking device measures the pupil size in millimeters.

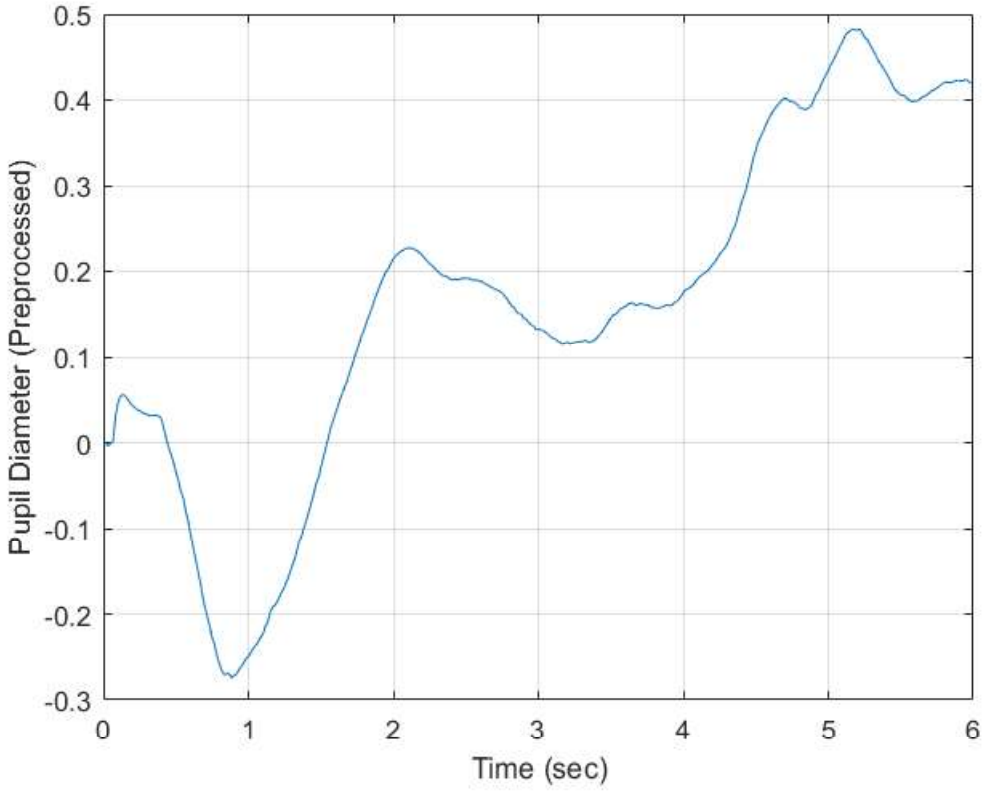


Figure 3.8: Pupillary Response of One Random Trial After Preprocessing

We applied normalization in order to disregard signal variability between subjects since participants can have different pupil sizes. With normalization, we shifted all signals in order to make them have the initial value of 0. Figure 3.8 shows the result of these preprocessing steps implemented on the trial which is given in the Figure 3.6.

Before preprocessing, for all subjects who participated, Dataset 1.1 and Dataset 1.2 had 320 trials (16x20). After preprocessing, Dataset 1.1 has 189 trials where the 40.94% of its data was eliminated. Similarly, 39.38% of Dataset 1.2 was removed after preprocesing and it has 194 trials. For Dataset 2.1 and Dataset 2.2, for all subjects, 300 trial data was gathered at first (15x20). After preprocessing, Dataset 2.1 has 177 trials (41% of its data was removed) and Dataset 2.2 has 205 trials with 31.7% data loss.

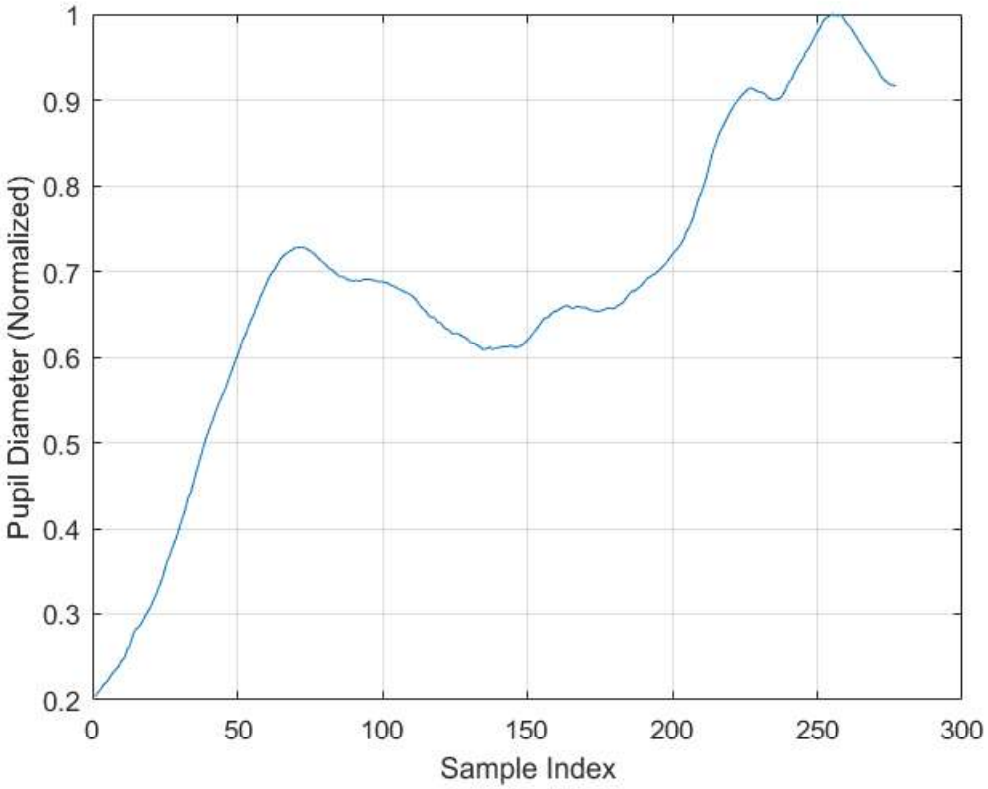


Figure 3.9: Normalized Pupillary Response of One Random Trial

3.4.2 Quality Control

We evaluated the quality of the datasets by first analyzing the constriction point and the slope of each pupil data. It is shown that pupillary response has a pattern, i.e., when viewing an image it constricts during the first 1.5 seconds due to initial light reflex and then dilates (Beatty & Lucero-Wagoner, 2000; Bradley et al., 2008). Based on this pattern, we assumed to see a constriction point between 0-90 time points and after this period, we expected to see an increment in the data. To verify the expected pupillary response in our study, we used polynomial fitting which is one of the most successful methods. We applied 6th order polynomial to each trial of each subject's data to assign a constriction point in the specified time points. If the constriction point wasn't between 0-90 time points, we eliminated that trial. However, if the constriction point was in that interval, the beginning of the data was admitted as this constriction point and the data before this point was discarded. Then, a linear regression was utilized on each trial and the slope of the data was calculated. If the value of the

slope wasn't positive, then the data was omitted. Finally, signals of each subject were rescaled in order to adjust a standard metric in [0 1] range. Rescaling method finds the maximum and minimum pupil points of each subject, makes the minimum value of subject's data as 0 and divides the all data of subject to maximum value to label the maximum point as 1. In Figure 3.9, the trial was normalized such that it has a maximum point although it hasn't minimum point as 0.

During quality control, some data was eliminated from datasets. Dataset 1.1 lost 1 trial in constriction point checkpoint and 23 trials over slope analysis. So, it has 165 trials at the end with 12.7% data loss. In Dataset 1.2, 5 trials were eliminated after constriction point fixation and 21 trials were removed after slope analysis. In total, it lost 13.4% of its data and it has 168 trial data. In Dataset 2.1, no trials were removed during constriction point control but 40 trials were eliminated after slope analysis. Thus, total number of trials reduced from 177 to 137 with 22.6% data loss. Dataset 2.2 lost its one trial over constriction point analysis and 60 trials after positive slope checkpoint. It had 205 trials before quality control, with 29.76% data loss, the number of total trials became 144 at the end.

The quality with slope analysis was done before feature extraction. After features were extracted, the statistical analysis was performed in SPSS Software for quality check of features. We first performed the normality tests called as Kolmogorov-Smirnov Test to observe the distribution of features and to determine what type of statistical methods (parametric or nonparametric) should be used. The normality test showed that some features had normal distributions while some of them had non-normal distributions, so we analyzed the data using non-parametric method with Mann-Whitney U test. We compared the asymptotic significance (2-tailed) p-value between neutral and stress data to interpret test results. Since we chose the significance interval as 95% where the differences with the probability $\alpha \leq 0.05$ were regarded as significant.

3.4.3 Feature Extraction

In this work, entropy based features were extracted by following the same procedure introduced in Baltaci and Gokcay (2016). Different from their work, area under curve,

polynomial fitting and regression slopes of processed data were also used as features.

Before explaining the features, it is important to introduce the method called Shannon entropy which is one of the widely used novel information-theoretic approaches for signal analysis. Shannon proposed a mathematical theory to quantify information and described entropy as a measure of uncertainty in a random variable (Shannon, 1948). Entropy is related to a probability distribution of an event and can be calculated using any kind of data. In our case, we computed Shannon entropy dependent on time by assigning a specific interval (window) that slides along the entire record. We divided this window into a discrete number of signal amplitude levels and calculated the probability density of each discretized signal level. With the known probability values, we calculated entropy value of each window by using the formula of Shannon entropy. As the window was sliding through the data, we added entropy values and formed entropy sequences for each pupil record. The formula of Shannon entropy:

$$H = -\sum_{n=1}^{N} P(\ell_n) \log_2(P(\ell_n))$$

where ℓ_n denotes the n^{th} level and $P(\ell_n)$ is the probability of level n. The discretization level affects the entropy value so it is important to find the optimal level. In Appendix F, the relation between different discretization levels and the entropy were shown. It is also given that as the level increases, the entropy value shows little difference. Thus, it can be said that the convergence of entropy values start when the level was chosen between 100-150. From the probability distribution plots, it is seen that as the level increases from 4 to 64, the distribution became smoother and gave better results. However, after 64-level the distribution became meaningless as many peaks appeared. Therefore, we decided to divide our data into 100 discrete levels for our entropy calculation.

In this method, the choice of window parameter is significant because the interval on the data affects both the quality of data and the computation time. We chose the window size for dataset 1 as 25 (see Table 4.9), for dataset 2 (see Table 4.18) as 10 and for dataset 3 as 50 (see Table 4.24).

Based on these entropy method, we presented main statistical features which are minimum, maximum, mean, median, standard deviation, kurtosis and skewness values of

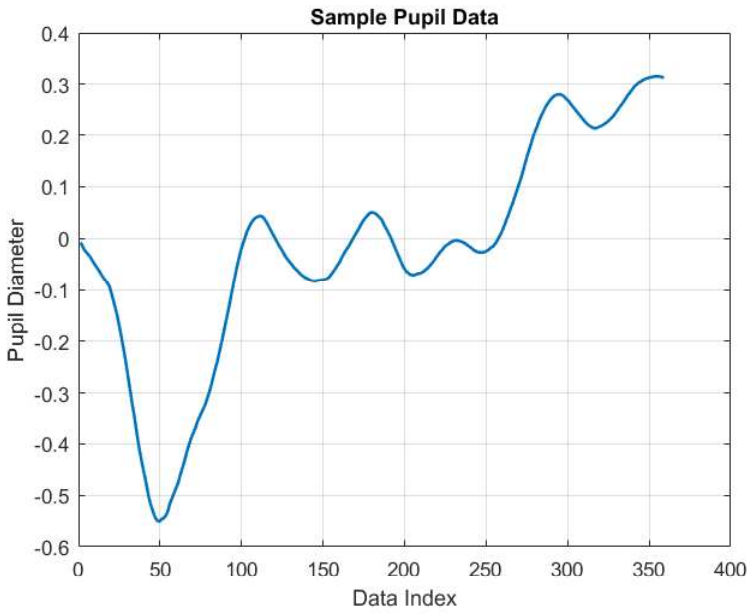


Figure 3.10: A trial data of one subject's pupil data before quality control

entropy series. In addition, we generated these statistical features by using the actual pupil records. However, we also defined different features for actual measurements that are area under curve, regression slope, polynomial coefficients and curve correlation values.

Minimum is the minimum value of the entropy based data or the absolute pupil data.
Maximum is the maximum value of the entropy based data or the absolute pupil data.
Mean is the average value of the entropy based data or the absolute pupil data.
Median is the middle value of the entropy based data or the absolute pupil data.
Standard Deviation is the amount of variation of the data in the entropy window or the absolute pupil data.
Kurtosis is the amount of probability in the tails of the entropy based data or the absolute pupil data.
Skewness is the measure of symmetry in the distribution of the entropy based data or the absolute pupil data.
Area Under Curve (AUC) is the area in between the data points and the x-axis, i.e., the integral value of the pupil data after the constriction point.

Regression Slope

The slope values of each record were used as a feature. As we mentioned in the qual-

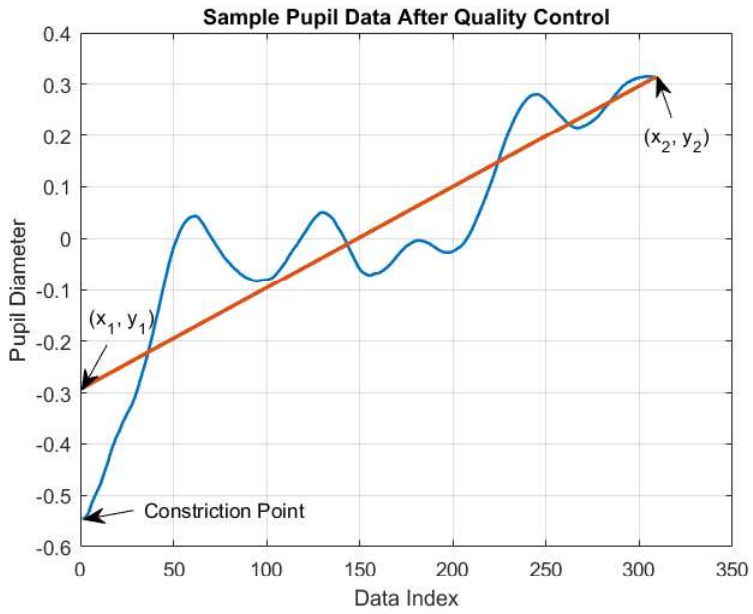


Figure 3.11: The data after quality control where linear regression line is fitted

ity control section, we used regression to fit the pupil data into a 6th order polynomial and found a constriction point as a beginning point of the data. The slope was computed in this piece where the beginning is the constriction point and the end is the last point of data with the help of linear regression. We eliminated the records which had non-positive slope value. In Figure 3.10, a sample pupillary response of one subject is given. It is clear that there is a constriction point during first 90 time points. In Figure 3.11, the principle of regression slope feature is illustrated. After finding constriction point, a linear regression line is fitted on the pupil data and the slope was calculated as

$$Slope = \frac{y_2 - y_1}{x_2 - x_1}, \tag{3.1}$$

where (x_1, y_1) denotes the beginning point of the line and (x_2, y_2) represents the end of point of the line.

Polynomial Coefficients

The computation of the coefficients of polynomial curve fitting is another method used in signal analysis. In this study, we tested both 3rd order and 6th order polynomial regression in our data. 6th order polynomials resulted in overfitting, so we computed the coefficients of 3rd order polynomial curve and used these values as our features. The sample data and its 3rd order polynomial are given in the Figure 3.12.

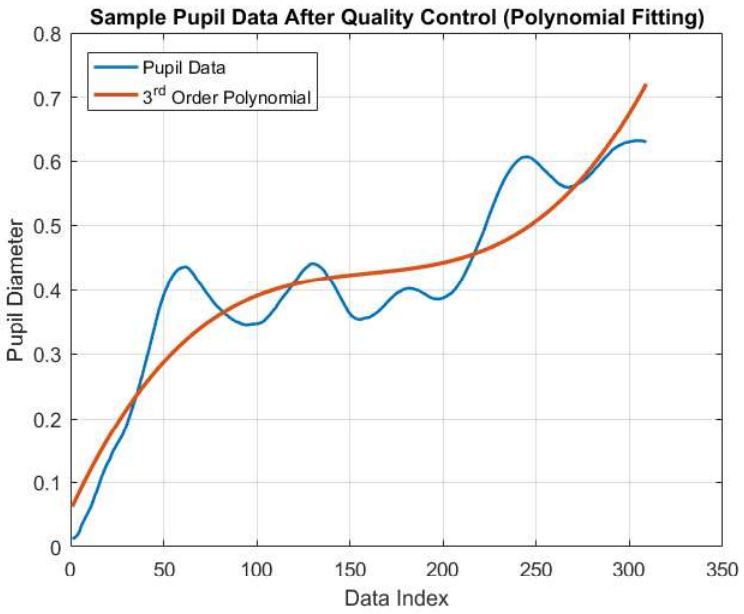


Figure 3.12: Illustration of polynomial fitting to a sample pupil data

The polynomial has an expression where the coefficients are in descending powers $p(x) = a_1x^3 + a_2x^2 + a_3x + a_0$. Therefore, we labeled our coefficient features as a1, a2, a3 and a0 similar to this expression and calculated their values for each trial.

Curve Correlation

The last feature in our study is related to the curve correlation of records. We had both phase 1 and phase 2 pupil records for each subject. In this feature, we generated the average of both phase 1 and 2 data (see Appendix G) and compared a single random data with both of them as given in the Figure 3.13. The distance between the observed record and the averaged record summed up and gave the correlation result. The formula of this feature can be expressed as

$$CurveCorr = \sum_{n=1}^{N} D_n \tag{3.2}$$

where D_n is the Euclidean distance between observed data and averaged data. Because we compared the monitored data both with the averaged phase 1 and phase 2 records, we labeled this feature as curvecorrelation1 and curvecorrelation2 respectively.

Curve Correlation Difference

This feature is obtained by subtracting curvecorrelation2 from curvecorrelation1.

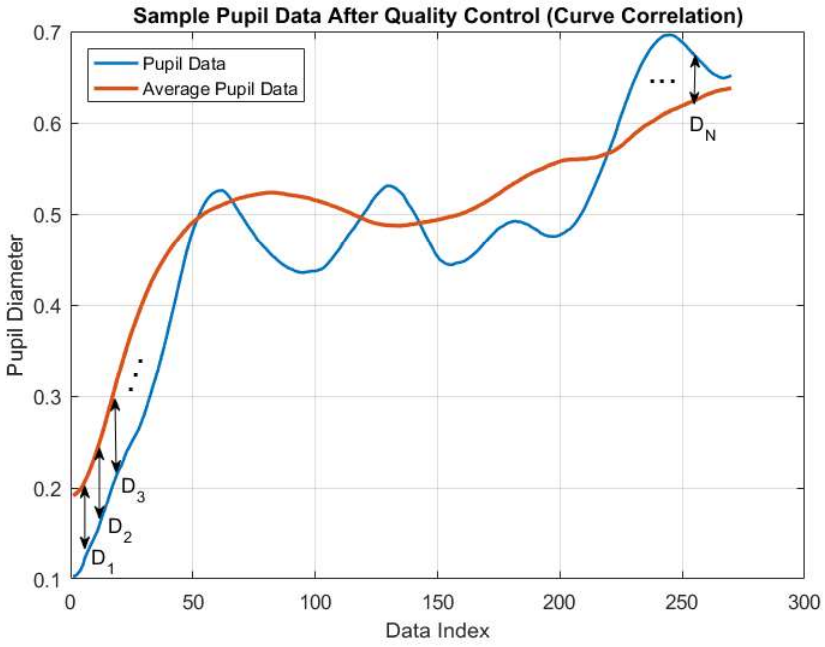


Figure 3.13: Illustration of curve correlation feature

All of the features can be categorised into 3 groups as entropy-based, absolute, and global. These are listed in Appendix K.

3.4.4 WEKA

The classification process was performed by using WEKA that was developed at the University of Waikato in New Zealand. WEKA platform includes state-of-the-art machine learning algorithms and numerous pre-processing tools. It enables users to implement various filters for pre-processing of inputs and feature selection, many classification algorithms, visual interpretations of data and evaluation of results (Witten & Frank, 2005).

In this study, as an input, we had 22 features extracted from the processed pupil data and described them as class 1 for neutral group and class 2 for stressed group. So, in binary classification model, 0 refers to the class 1 as our neutral data and 1 refers to the class 2 as our stressful data. In order to determine the most reliable classification algorithms, we first examined the accuracies achieved by several state-of-the-art algorithms, including Naive Bayes, Logistic Regression, SVM, K-nearest neighbor, Adaboost, Bagging, Decision Tree and Random Forest. We then determined the algo-

rithms providing the highest accuracies for each analysis and selected the successful algorithms in common, which are SVM, Adaboost and Random Forest (see Appendix H).

SVM is a popular supervised machine learning algorithm for classification problems. It classifies the dataset by selecting a hyperplane which separates the data by their class. The margin that is the distance between the hyperplane and the nearest data point should be large for better classification. Thus, SVM algorithm finds the coefficients that maximize the margin. In this study, the default parameters in WEKA were used and as the kernel type puk kernel was chosen.

AdaBoost is a successful ensemble method developed for binary classification. Ensemble method is basically designing a model from the training data and then building a second model which corrects the errors from the first model. This process continues until the training set is perfectly predicted. Adaboost makes use of any machine learning algorithms but it is best performed with decision trees, i.e., weak learners, as these trees only make one decision for classification. The training weights are updated in this algorithm by giving more weight to the incorrect predictions while giving less weight to the correct predictions. In this study, Adaboost was used with default parameters and as a weak learner, random forest was chosen to obtain less error rate on classification model.

Random Forest is one of the most powerful ensemble machine learning methods. In this algorithm, a bunch of decision trees are created with random subset of the data. Over training, features in decision trees are selected from random samples and models are designed for each data sample. The output of each decision tree in the forest is combined and averaged to make a final prediction. In the study, the default parameters of the classifier were used.

Except for finding the most reliable classification algorithm, it is also important to select the test option on WEKA. During classification process, the given dataset is divided into two sets as training set and test set where the labels of test set are unknown and the compatibility of test set into training set determines the classification accuracy (Witten & Frank, 2005). This division in the given dataset can be done with several methods like cross validation or percentage split. We used the cross validation

method (Efron & Tibshirani, 1993) in our study and found that 10-fold gave the best results compared to 5-fold and 15-fold (see Table 4.7, 4.16, 4.22). The 10-fold cross validation method divided our dataset into 10 equal subsets where nine subsets were training sets and one subset was test set. The selected classification algorithm trained these nine subsets and based on them, the remaining set was tested. This cycle was repeated until every 10 subsets were used as a test set. The result was calculated based on the average of these classification cycles.

CHAPTER 4

RESULTS

This chapter presents the statistical test results, the classification accuracies and the effects of parameters on classification results for three different analyses. There are three main subsections where the results of each analysis are illustrated separately.

We performed 3 different analyses based on 4 different datasets obtained from the experiments in order to detect any possible correlation between the pupillary response and the stress factors. In the first two analyses, we focused on investigating the marginal effect of cognitive factors. We worked on two different cases that increase the cognitive load. In one case, the cognitive load, i.e., attention, was raised by increasing the number of arrows in the stimuli. In the other case, the background of the stimuli were destroyed by scrambling the images while keeping the number of arrows the same, which led to increased cognitive load. For the first case, we performed several classification algorithms on Dataset 1.1 and Dataset 1.2 to determine the separability. For the second case, we used the same classification algorithms on Dataset 2.1 and Dataset 2.2. The third and the final analysis aimed to reveal the effect of emotional load on the pupillary response. For this case, we worked on Dataset 1.1 and Dataset 2.1 since the only difference in the stimuli was color for these datasets.

This study aims to verify the following hypotheses:

Hypothesis 1: The intense cognitive load causes subjects' pupil diameter to increase. With the change of pupil size, subjects' mental state can be detected.
Hypothesis 2: Color as an emotional load has an effect on pupil dilation and this affect can be detected by measuring pupillary response.
The results of the first two analyses were utilized to test the Hypothesis 1, while the

Table4.1: Normality test for the first analysis

Features	Statistic	Sig. (p)
min_ent	**.070**	**.200**
max_ent	.474	.000
mean_ent	**.053**	**.200**
std_ent	**.042**	**.200**
median_ent	**.047**	**.200**
kurt_ent	.189	.000
skew_ent	**.040**	**.200**
max_abs	**.060**	**.200**
mean_abs	**.043**	**.200**
std_abs	.075	.000
median_abs	**.035**	**.200**
kurt_abs	.127	.000
skew_abs	**.068**	**.200**
slope_abs	**.046**	**.200**
curve_1_abs	.089	.013
curve_2_abs	.109	.000
curve_diff_abs	.112	.000
a3_coeff	.386	.000
a2_coeff	.200	.000
a1_coeff	**.047**	**.200**
a0_coeff	**.057**	**.200**
AUC	**.042**	**.200**

results of the third analysis were used to test Hypothesis 2.

4.1 First Analysis

In this section, we first represent the statistical test results on Dataset 1.1 and Dataset 1.2. We then illustrate the classification accuracies obtained by three state-of-the-art algorithms, namely SVM, Adaboost and Random Forest.

4.1.1 Statistical Test Results

Statistical analyses provide useful information that enable us to detect any relation between the data samples belonging to different classes. One of the most powerful tools to extract statistical information from a given dataset is the T-Test analysis. There exist parametric and non-paramteric T-Test analyses which are specialized to

Table4.2: Mann-Whitney U Test Results for the first analysis

Features	Mann-Whitney U	Z	2-tailed Asymp. Sig. (p)
min_ent	**7985.000**	**-2.019**	**.045**
max_ent	8886.500	-.042	.966
mean_ent	7864.000	-1.651	.099
std_ent	8356.000	-.870	.384
median_ent	8142.500	-1.209	.227
kurt_ent	**7944.000**	**-1.964**	**.028**
skew_ent	8728.000	-.281	.779
max_abs	8217.000	-1.091	.275
mean_abs	**7505.000**	**-2.220**	**.026**
std_abs	8756.000	-.236	.813
median_abs	**7272.000**	**-2.589**	**.010**
kurt_abs	**7411.000**	**-2.369**	**.018**
skew_abs	**7299.000**	**-2.546**	**.011**
slope_abs	8600.000	-.484	.629
curve_1_abs	8555.000	-.555	.579
curve_2_abs	8495.000	-.650	.516
curve_diff_abs	**7997.500**	**-2.205**	**.016**
a3_coeff	8513.000	-.622	.534
a2_coeff	8016.000	-1.410	.159
a1_coeff	**7601.000**	**-2.068**	**.039**
a0_coeff	**8454.000**	**-2.132**	**.013**
AUC	**7500.000**	**-2.228**	**.026**

normal and non-normal distributions respectively. Therefore, to be able to perform the appropriate T-Test analysis, we need to know about the distribution of our features extracted from the datasets. To this end, we first applied a normality test on the features obtained from both Dataset 1.1 and Dataset 1.2 using SPSS Statistics v.24 to check whether they have a normal distribution or not.

We performed Kolmogorov-Smirnov normality test consisting of two hypotheses, where the null hypothesis was associated with the normal distribution. We fed our features into this test and obtained the corresponding significance values, denoted as p. Here, p represents the probability of Type 1 error, i.e., rejecting the null hypothesis when it is true. Therefore, higher values of p indicate that null hypothesis should not be rejected. We chose the confidence interval as 95%, which means we assigned features with $p > 0.05$ as normally distributed. In Table 4.1, the normally distributed features are indicated in bold letters. Since not all the features are normally distributed, we decided to use non-parametric T-Test analysis.

Table4.3: Behavioral Response Accuracies

	Accuracy (%)	
Subjects	Part 1	Part 2
S1	95	90
S2	95	90
S3	95	95
S4	100	90
S5	85	85
S6	75	85
S7	70	45
*S8	50	45
S9	90	75
S10	70	65
*S11	45	65
*S12	60	55
S13	70	60
S14	75	65

Similar to the structure of the normality test, non-parametric T-Test also has the form of binary hypothesis testing, where the null hypothesis is associated with no difference case, i.e., the data samples belonging to different classes are not separable. We used Mann-Whitney U Test to obtain the 2-tailed asymptotic significance p, where p denotes the probability that there is no observable difference exists between the data samples of different classes. The confidence interval for this test, same as the normality test, was chosen as 95%. This suggests that features with $p < 0.05$ imply significant difference. In Table 4.2, we illustrate the p values of the features, where the features in bold letters have p values less than 0.05. Based on these results, we observed that it is possible to separate the pupillary responses of neutral and stress states. This indicates that the cognitive load achieved by increased number of arrows has an impact on the pupillary response, which supports our Hypothesis 1.

4.1.2 Classification Results

In this section, we present the classification accuracies obtained by 3 different state-of-the-art algorithms, i.e., SVM, Adaboost and Random Forest. We denoted Dataset

Table4.4: Classification Accuracies of Subject Based Analysis

SVM		Adaboost		Random Forest	
Subjects	Accuracy	Subjects	Accuracy	Subjects	Accuracy
1	80	1	**93.33**	1	80
2	57.15	2	**71.43**	2	61.91
3	**65**	3	60	3	55
4	**65**	4	60	4	60
5	**66.67**	5	**66.67**	5	62.5
6	60	6	**64**	6	56
7	75	7	71.87	7	**81.25**
8	66.67	8	**79.17**	8	68.3
9	66.67	9	**75**	9	55.56
10	61.91	10	60.61	10	**62.38**
11	60	11	60	11	**70**
12	**72.73**	12	59.09	12	59.09
13	**66.67**	13	**66.67**	13	61.90
14	69.23	14	64.84	14	**78.06**

Table4.5: Classification Accuracies of Collective Analysis

	SVM	Adaboost	Random Forest
Cumulative Accuracy	60.36	**61.96**	58.04

Table4.6: Effects of Different Feature Sets on Classification Accuracies

Analysis 1			
Classification Algorithm	Entropy	Absolute	Global
SVM	55.6	53.73	54.12
Adaboost	54.9	55.3	52.16
Random Forest	50.58	56.08	52.18

1.1 as the neutral class that refers to 0 in classification analysis and Dataset 1.2 as the stress class that refers to 1 during classification. We obtained 22 features and used all these features during analyses.

At first, we realized subject by subject classification, where the pupillary response of each subject was evaluated individually to determine the neutral versus stressed state of that subject. In Table 4.3, behavioral performance of the subjects are shown, based on their accuracies in counting the arrows. Except for 4 subjects, all subjects achived a performance above chance level (55%) in both parts of the task. In Table 4.4, we represent the classification accuracies for the individual cases. Considering the subject by subject classification accuracies in Table 4.4, we observed that it is possible to achieve a classification accuracy between 62.38% and 93.33% depending on the subject, regardless of the classifier used (best classifier performance for each subject is shown with boldface). These results indicate that the pupillary response is certainly affected by cognitive load, which is achieved by increasing the number of arrows.

In collective analysis, we eliminated some of the subjects with low performance (these are shown by asterisk in Table 4.3). During Part 1 of the experiment, subjects gave correct answers with 77% accuracy on average. In Part 2, this average behavioral response decreases to 72% since the task was harder. Therefore, when we evaluated subjects' performance, the response accuracy on the first part was the criteria for elimination. In this case, the data of 3 subjects were removed before the collective classification. Table 4.5 shows the results of collective analysis.

We found out that the algorithms SVM, Adaboost and Random Forest obtain accuracies of 60.36%, 61.96% and 58.04% respectively for the collective classifications.

Table4.7: Effect of Different Folds on Classification Accuracies

Analysis 1			
Classification Algorithm	5 folds	10 folds	15 folds
SVM	54.65	56.16	51.65
Adaboost	54.05	60.66	50.75
Random Forest	57.36	58.86	60.36

Although collective classification accuracies are not as high as the individual results, they still show that the pupillary responses of the neutral and the stress cases are separable. Overall, the classification accuracies support the Hypothesis 1.

The 22 features used in classification can be divided into three categories as entropy based features, absolute data features and global features as described in Appendix K. When we analyzed our data by exlusively using features from a single category, the accuracy results given in Table 4.6 are obtained. Overall, the classification accuracy of features in the three categories are similar, and much lower than the combined set of 22 features.

4.1.3 Effect of Algorithm Parameters

We further evaluated the classification results by examining the effects of different parameter selections. We focused on 3 different parameters, where the first one is the number of folds used for cross validation, the second one is the size of the window used for moving average filter in the preprocessing stage and the last parameter is the size of the window that we used to extract entropy related features.

In Table 4.7, we illustrate the collective classification accuracies of each algorithm for different fold numbers of 5, 10 and 15. Based on these results, we decided to use 10 fold cross validation for each algorithm. In this comparison, the window sizes for the moving average filter and the entropy based features were selected as 20 and 100 respectively.

In Table 4.8, we represent the collective classification accuracies of each algorithm for different window sizes that we used for moving average filtering. We deduced from

Table4.8: Effect of Window Size (Moving Average Filter)

Analysis 1					
Classification Algorithm	WS=5	WS=10	WS=20	WS=30	WS=40
SVM	54.35	55.25	**56.16**	53.15	50.46
Adaboost	54.95	57.96	**60.66**	56.46	54.96
Random Forest	54.65	54.65	**58.86**	51.95	52.05

Table4.9: Effect of Window Size (Entropy Based Features)

Analysis 1					
Classification Algorithm	WS=10	WS=25	WS=50	WS=100	WS=150
SVM	56.16	**60.36**	56.16	56.16	59.76
Adaboost	59.16	**61.96**	56.46	60.66	56.76
Random Forest	58.56	**58.04**	56.76	58.86	55.86

these results that a moving average with window size 20 gives the best accuracies. Thus, for the first analysis, we chose the window size as 20. The number of folds and the window size for entropy based features were selected as 10 and 100 respectively for this comparison.

In Table 4.9, we provide the collective classification accuracies for different window sizes that we used for extracting entropy based features. We observed that the best accuracies are obtained when the window size is 25. In this comparison, the number of folds and the window size of the moving average filter were chosen as 10 and 20 respectively.

The classification accuracies given in Table 4.4 and 4.5 were obtained by using the optimal parameters found in this section.

4.2 Second Analysis

In this section, we first represent the statistical test results on Dataset 2.1 and Dataset 2.2. We then illustrate the classification accuracies obtained by three state-of-the-art algorithms, namely SVM, Adaboost and Random Forest.

Table4.10: Normality test for the second analysis

Features	Statistic	Sig. (p)
min_ent	.319	.000
max_ent	.395	.000
mean_ent	**.069**	**.200**
std_ent	**.098**	**.200**
median_ent	.132	.000
kurt_ent	.175	.000
skew_ent	**.103**	**.200**
max_abs	**.086**	**.200**
mean_abs	.118	.122
std_abs	**.062**	**.200**
median_abs	**.101**	**.200**
kurt_abs	.199	.000
skew_abs	.138	.029
slope_abs	**.083**	**.200**
curve_1_abs	**.090**	**.200**
curve_2_abs	**.097**	**.200**
curve_diff_abs	**.086**	**.200**
a3_coeff	.410	.000
a2_coeff	.236	.000
a1_coeff	**.097**	**.200**
a0_coeff	**.085**	**.200**
AUC	.118	.113

4.2.1 Statistical Test Results

Similar to the previous analysis, we first performed a normality test on the features obtained from Dataset 2.1 and Dataset 2.2 in order to decide whether the distribution is normal or not. The results of the Kolmogorov-Smirnov normality test is given in Table 4.10. It is clear there are features that have a p value of less than 0.05. Hence, not all the features are normally distributed. As a result, we continued to use the non-parametric Mann-Whitney U Test in order to detect any significant difference between the feature sets obtained from Dataset 2.1 and Dataset 2.2.

In Table 4.11, we represent the results of Mann-Whitney U Test. We observed that there exist features with p value less than 0.05. As discussed earlier, this shows that some features have a significant difference, which imply that it is possible to separate the datasets. With these observations, we can say that increasing the cognitive load by destroying the background has an effect on pupil dilation, which supports the

Table4.11: Mann-Whitney U Test Results for the second analysis

Features	Mann-Whitney U	Z	2-tailed Asymp. Sig. (p)
min_ent	**998.000**	**-2.041**	**.028**
max_ent	1173.000	-.564	.598
mean_ent	**1100.000**	**-2.127**	**.038**
std_ent	948.000	-1.626	.104
median_ent	1150.000	-.167	.867
kurt_ent	1087.000	-.621	.534
skew_ent	1101.000	-.520	.603
max_abs	**1140.000**	**-2.238**	**.012**
mean_abs	**1163.000**	**-2.072**	**.032**
std_abs	1164.000	-.065	.948
median_abs	1167.000	-.043	.965
kurt_abs	1112.000	-.441	.659
skew_abs	1162.000	-.079	.937
slope_abs	1029.000	-1.040	.298
curve_1_abs	1081.000	-.665	.506
curve_2_abs	1053.000	-.867	.368
curve_diff_abs	**1003.000**	**-2.229**	**.014**
a3_coeff	1138.000	-.253	.800
a2_coeff	1163.000	-.072	.942
a1_coeff	1165.000	-.058	.954
a0_coeff	1164.000	-.065	.948
AUC	1165.000	-.058	.954

Hypothesis 1.

4.2.2 Classification Results

In this section, we present the classification accuracies obtained by the same state-of-the-art algorithms used in previous analysis, i.e., SVM, Adaboost and Random Forest. In this case, we labeled Dataset 2.1 as the neutral class and Dataset 2.2 as the stress class. Similar to the previous analysis, we performed both collective and subject by subject classification. Similar to the former collective analysis, the data of some subjects were eliminated based on their responses on the given task during the experiment. We asked subjects the number of arrows on images and if the subject failed to achieve 70% accuracy on the given task, we removed its data from the collective analysis. During Part 1 of the experiment, subjects gave correct answers with 67% accuracy on average. In Part 2, this average behavioral response decreases sharply to 47% which indicates that subjects had difficulty to complete the task. Addi-

Table4.12: Behavioral Response Accuracies

	Accuracy (%)	
Subjects	Part 1	Part 2
*S1	50	50
*S2	60	55
*S3	60	40
S4	70	40
S5	85	45
S6	85	50
*S7	60	50
*S8	65	65
S9	70	15
*S10	60	50
*S11	65	45
*S12	55	30
S13	80	60

tionally, we also analyzed the responses of the subjects to the debriefing form. Some subjects said that they were excited from the beginning (see Appendix I). These subjects data were also excluded from the collective analysis. In this case, the data of 8 subjects were removed before the collective classification (see Table 4.12, (*) mark is for the eliminated subjects). We illustrate the classification accuracies in Table 4.13 and 4.14.

In Table 4.13, for the classification of individual subject data, classification accuracies that are highest for each subject are shown in bold face. For each subject, the highest classification accuracy of cognitive state varied between 63.16% and 87.5%. With these results, we can claim that the pupillary response is affected by cognitive load created by destroying background of the stimuli.

For the collective classification, results are presented in Table 4.14. Even though the accuracies, i.e., 61.85% for SVM, 58.72% for Adaboost and 57.73% for Random Forest, are lower than the individual cases, they still illustrate the effect of cognitive load on pupillary response. Similar to the previous analysis, we used 22 features during classification. These features can be categorised into three groups and the evaluation based on these categories can be seen in Table 4.15.

Table4.13: Classification Accuracies of Subject Based Analysis

SVM		Adaboost		Random Forest	
Subjects	Accuracy	Subjects	Accuracy	Subjects	Accuracy
1	**69.57**	1	65.22	1	60.87
2	**69.23**	2	61.54	2	61.54
3	68.75	3	**87.5**	3	62.5
4	61.90	4	**76.19**	4	66.67
5	**75**	5	69.16	5	58.33
6	**75**	6	**75**	6	60.37
7	57.63	7	**65.22**	7	54.17
8	**63.16**	8	57.89	8	62.63
9	**66.67**	9	61.90	9	56.38
10	**63.64**	10	50	10	54.64
11	60	11	64	11	**68**
12	75	12	**75**	12	**75**
13	71.05	13	**76.32**	13	63.16

Table4.14: Classification Accuracies of Collective Analysis

	SVM	Adaboost	Random Forest
Cumulative Accuracy	**61.85**	58.72	57.73

4.2.3 Effect of Algorithm Parameters

We again evaluate the classification results by examining the effect of parameters given in the previous section. In Table 4.16, we illustrate the collective classification accuracies of each algorithm for different fold numbers of 5, 10 and 15. Based on these results, we decided to use 10 fold cross validation for each algorithm, similar to the analysis 1. In this comparison, the window sizes for the moving average filter and the entropy based features were selected as 20 and 100 respectively.

In Table 4.17, we represent the collective classification accuracies of each algorithm for different window sizes that we used for moving average filtering. We deduced from these results that a moving average with window size 20 gives the best accuracies. Thus, for this analysis, we again chose the window size as 20. The number of folds and the window size for entropy based features were selected as 10 and 100 respectively for this comparison.

Table4.15: Effects of Different Feature Sets on Classification Accuracies

Analysis 2			
Classification Algorithm	Entropy	Absolute	Global
SVM	56.7	58.7	57.8
Adaboost	52.53	50.52	56.7
Random Forest	50.52	54.6	51.5

Table4.16: Effects of Different Folds on Classification Accuracies

Analysis 2			
Classification Algorithm	5 folds	10 folds	15 folds
SVM	50.89	**55.87**	54.80
Adaboost	51.96	**56.23**	56.23
Random Forest	51.60	**57.29**	56.23

In Table 4.18, we provide the collective classification accuracies for different window sizes that we used for extracting entropy based features. We observed, different from analysis 1, that the best accuracies are obtained when the window size is 10. In this comparison, the number of folds and the window size of the moving average filter were chosen as 10 and 20 respectively.

The classification accuracies given in Table 4.13 and 4.14 were obtained by using the optimal parameters found in this section.

4.3 Third Analysis

In this section, different from the previous two analyses, we investigate the effect of emotional load in term of color on the pupillary response. Therefore, we compared Dataset 1.1 with Dataset 2.1. Similar to the previous analyses, we first represent the statistical test results on Dataset 1.1 and Dataset 2.1. We then illustrate the classification accuracies obtained by three state-of-the-art algorithms, namely SVM, Adaboost and Random Forest.

Table4.17: Effect of Window Size (Moving Average Filter)

Analysis 2					
Classification Algorithm	WS=5	WS=10	WS=20	WS=30	WS=40
SVM	51.96	53.02	**55.87**	55.16	52.67
Adaboost	52.45	54.65	**56.23**	55.87	54.05
Random Forest	52.67	51.89	**57.29**	56.95	53.19

Table4.18: Effects of Window Size (Entropy Based Features)

Analysis 2					
Classification Algorithm	WS=10	WS=25	WS=50	WS=100	WS=150
SVM	**61.85**	53.74	53.74	55.87	56.23
Adaboost	**58.72**	53.38	55.52	56.23	54.45
Random Forest	**57.73**	56.58	54.80	57.29	55.16

Table4.19: Normality test for the third analysis

Features	Statistic	Sig. (p)
min_ent	**.052**	**.200**
max_ent	**.103**	**.200**
mean_ent	**.051**	**.200**
std_ent	**.080**	**.200**
median_ent	**.062**	**.200**
kurt_ent	**.105**	**.200**
skew_ent	**.071**	**.200**
max_abs	**.083**	**.200**
mean_abs	**.069**	**.081**
std_abs	**.082**	**.200**
median_abs	**.062**	**.200**
kurt_abs	.176	.000
skew_abs	.136	.200
slope_abs	**.075**	**.200**
curve_1_abs	.190	.000
curve_2_abs	.164	.000
curve_diff_abs	.132	.019
a3_coeff	.274	.000
a2_coeff	.148	.005
a1_coeff	**.072**	**.200**
a0_coeff	**.077**	**.200**
AUC	**.069**	**.200**

Table4.20: Mann-Whitney U Test Results for the third analysis

Features	Mann-Whitney U	Z	2-tailed Asymp. Sig. (p)
min_ent	**1290.000**	**-2.720**	**.012**
max_ent	1105.000	-1.891	.059
mean_ent	**1031.000**	**-2.357**	**.018**
std_ent	**1334.000**	**-2.142**	**.036**
median_ent	**1049.000**	**-2.244**	**.025**
kurt_ent	1303.000	-.638	.523
skew_ent	1228.000	-1.112	.266
max_abs	**1081.500**	**-2.039**	**.041**
mean_abs	1316.000	-.556	.578
std_abs	**884.000**	**-3.286**	**.001**
median_abs	1312.000	-.581	.561
kurt_abs	1240.000	-1.036	.300
skew_abs	1372.000	-.202	.840
slope_abs	**1377.000**	**-2.192**	**.035**
curve_1_abs	1100.000	-.171	.865
curve_2_abs	1137.000	-1.687	.092
curve_diff_abs	**1172.500**	**-2.163**	**.042**
a3_coeff	**1268.000**	**-2.059**	**.046**
a2_coeff	**1223.000**	**-2.144**	**.039**
a1_coeff	**1161.000**	**-2.236**	**.025**
a0_coeff	1288.000	-.733	.464
AUC	1311.000	-.588	.557

4.3.1 Statistical Test Results

Similar to the previous analyses, we again performed a normality test on the features obtained from Dataset 1.1 and Dataset 2.1. The results of the Kolmogorov-Smirnov normality test is given in Table 4.19. Based on these results, we observed that not all the features are normally distributed. Thus, we performed the non-parametric Mann-Whitney U Test in order to detect any significant difference between the feature sets obtained from Dataset 1.1 and Dataset 2.1.

We represent the results of Mann-Whitney U Test in Table 4.20. As discussed earlier, we again observed some features with p value less than 0.05 which shows a significant difference. So, with these results we conclude that these datasets are separable that means color as an emotional load has an effect on pupil dilation, which supports the Hypothesis 2.

Table4.21: Effects of Different Feature Sets on Classification Accuracies

Analysis 3			
Classification Algorithm	Entropy	Absolute	Global
SVM	56.6	66.04	64.15
Adaboost	53.78	64.15	63.21
Random Forest	52.6	62.26	64.15

Table4.22: Effects of Different Folds on Classification Accuracies

Analysis 3			
Classification Algorithm	5 folds	10 folds	15 folds
SVM	61.03	**64.83**	62.41
Adaboost	57.93	**61.72**	58.62
Random Forest	56.90	**60.69**	58.62

4.3.2 Classification Results

In this section, we present the classification accuracies obtained by the same state-of-the-art algorithms used in previous analysis, i.e., SVM, Adaboost and Random Forest. In this case, we labeled Dataset 1.1 as the stress class and Dataset 2.1 as the neutral class as color has an effect on emotions. All features were used during analysis and the effect of each feature set on the classification accuracy is given in Table 4.21. Different from the previous analyses, we only performed collective classification since the subjects in each dataset were different. In Table 4.25, cumulative classification accuracies are represented. Based on these results, we obtained a classification accuracy of 69.81% by SVM, 64.15% by Adaboost and 66.04% by Random Forest. In comparison, these accuracies are larger than those of analysis 1 and analysis 2. With these results, we can claim that the pupillary response is affected by emotional load in terms of the presence of color.

Table4.23: Effects of Window Size (Moving Average Filter)

Analysis 3					
Classification Algorithm	WS=5	WS=10	WS=20	WS=30	WS=40
SVM	55.86	61.03	**64.83**	59.31	57.93
Adaboost	60.34	58.72	**61.72**	60.69	57.38
Random Forest	58.62	59.65	**60.69**	58.62	57.16

Table4.24: Effects of Window Size (Entropy Based Features)

Analysis 3					
Classification Algorithm	WS=10	WS=25	WS=50	WS=100	WS=150
SVM	63.10	63.45	**69.81**	64.83	61.72
Adaboost	61.72	60.69	**64.15**	61.72	63.45
Random Forest	63.10	62.07	**66.04**	60.69	59.65

4.3.3 Effect of Algorithm Parameters

We similarly evaluate the classification results by analyzing the effect of parameters given in the previous section. In Table 4.22, we illustrate the collective classification accuracies of each algorithm for different fold numbers of 5, 10 and 15. Based on these results, we decided to use 10 fold cross validation for each algorithm, similar to the analysis 1 & 2. In this comparison, the window sizes for the moving average filter and the entropy based features were selected as 20 and 100 respectively.

In Table 4.23, we represent the collective classification accuracies of each algorithm for different window sizes that we used for moving average filtering. We observed that a moving average with window size 20 gives the best accuracies and chose the window size as 20. The number of folds and the window size for entropy based features were selected as 10 and 100 respectively for this comparison.

In Table 4.24, we provide the collective classification accuracies for different window sizes that we used for extracting entropy based features. We obtained that the best accuracies are gained when the window size is 50 -unlike previous analyses. In this comparison, the number of folds and the window size of the moving average filter were chosen as 10 and 20 respectively.

Table4.25: Classification Accuracies of Collective Analysis

	SVM	Adaboost	Random Forest
Cumulative Accuracy	**69.81**	64.15	66.04

The classification accuracies mentioned in the previous subsection were obtained by using the optimal parameters found in this section.

CHAPTER 5

DISCUSSION

The aim of this thesis was to examine the marginal effects of cognitive and emotional factors on pupil dilation. The pupillary responses were recorded during two phases of the experiments in which cognitive load was manipulated. The four different datasets were utilized in three different analyses to verify our hypotheses. In line with the aims of the study and literature review provided on Chapter 2, we hypothesized that intense pupil dilation is observed with increased cognitive or emotional stressors and can be used to detect stressful state. Our results showed that higher dilations occur when assigning difficult tasks or showing colored images.

The effects of viewing negative images and regulating emotions by changing these images on pupil dilation and skin conductance were investigated in the study of Kinner et al. (2017). Although pupil dilation is modulated by emotion regulation strategies, skin conductance is not affected. This study shows that pupil dilation is not only affected by emotional responses but also emotional regulation strategies. Based on the authors, these strategies provide evidence for increased cognitive effort: pupil dilates more with active cognitive load.

Plechawska-Wójcik and Borys (2016) proposed an approach where EEG signals are combined with pupillary response to detect cognitive load. Three participants performed experiments with an increased difficulty level. The subject-based analyses showed that pupil data enable detection of cognitive load similar to EEG results.

In the study of Snowden et al. (2016), the emotional response on pupil dilation was investigated by showing both neutral and fearful images. During one of the experiments, they examined passive and active viewing where the cognitive load increases

in active viewing. Their results showed no difference between passive and active viewing. This might be interpreted as emotional load represses the effect of cognitive load on pupil dilation or it can be said that emotional load has more intense effect on pupil dilation than cognitive load similar to our findings.

The proposed method used in Baltaci and Gokcay (2016) combines both pupillary response and thermal data. During emotional image viewing, they recorded both pupil and thermal data. They detected stress states of a user by implementing Adaboost and Decision Tree algorithms. Their accuracy rate ranged between 80-84% by Adaboost with Random Forest algorithm.

In the study of Jiang et al. (2015) the effect of visual motor task as an increased cognitive load on pupil dilation was investigated. A one-handed grasper was used by subjects to point to circles displayed on a monitor. According to the results of this study, the continuous movement task increased the cognitive load and this lead to greater pupil dilation.

The study of Aracena et al. (2015) proposed a classification approach for emotion detection. During image viewing, they recorded pupillary responses and used them in their neural network model and decision tree algorithm. Their study revealed a classification accuracy in the range of 70-78% on subject based analyses.

Pedrotti et al. (2014) also studied neural networks for stress detection. They collected their data from both pupillary response and electrodermal activity. They found that electrodermal activity signals are less capable of separating emotional states than pupil dilation. Their model achieved an accuracy of 79.2% which indicates that pupil dilates more with stressful conditions.

The study of Y. Gao et al. (2013) evaluated pupillary responses to identify the computer user's emotional states. Their signal processing depended on Kalman filtering, Wavelet denoising and Walsh transform. They used five classification algorithms and obtained accuracies in the range of 80-85%. Although they obtained high accuracy on collective analysis, their method requires many pre-processing steps which obstruct real-time analysis.

Klingner et al. (2011) studied the effects of visual and auditorial presentation on cog-

nitive load by performing three different tasks. They recorded pupil data during experiments and found that active cognitive load evokes greater pupil dilation which confirms our results.

In the study of Zhai and Barreto (2006), the identification of stress is performed by using SVM algorithm. They used not only pupil dilation signals but also blood volume pulse and galvanic skin response signals. Their accuracy reached up to 80% yet their system had remote sensors and this makes it unsuitable for real-time environment.

Overall our results are consistent with the literature (see Table 5.1). Most studies in the literature have been identifying stress by showing emotional pictures. Different from that, in this thesis, we mainly focused on different tasks that aimed to increase cognitive load to detect stress. Moreover, although there are several studies regarding the effects of color on emotions by creating an color-emotion model, this study investigates emotional effect of colors on a pupillary responses.

With respect to our hypotheses, stress detection by using pupil diameter is achievable on the subject based evaluation. In the cognitive load factor, we found that the neutral and stress classes of a subject can be classified with an accuracy between 87.75%-93.33%. However, on a collective based analysis, the best accuracy we obtained was approximately 61% which is not suitable to built applications for. This result can be due to subjective differences and inter-subject response variability so that in a collective analysis, the characteristical differences between subjects might decrease the overall accuracy. The previous experiences, different backgrounds and the habits of participants can increase the heterogeneity in the group and this would affect the group performance in a negative way. This difference between collective and subjective analyses can also be seen in Appendix J. In these analyses, the entropy values of the averaged and individual pupil data were compared. It is expected that the entropy values of the averaged data should be lower than the individual trials and the outcomes verify the situation where the subject based analyses have higher accuracies.

In the emotional load factor, we used the effect of color and compared datasets which corresponds the colored neutral images and grayscale versions of it. In this case, we had to make a collective analysis and we could separate responses for colored and

Table5.1: Summary of Studies Performing Classification Analysis

Reference	**Measurement Techniques**	**Classifiers**	**Classification Accuracy**
Zhai & Barreto (2006)	Pupil Data, Blood Volume Pulse, Galvanic Skin Response, Skin Temperature	Naive Bayes, SVM, Decision Tree	Up to 80% (collective)
Gao et al. (2013)	Pupil Data, Galvanic Skin Response	Naive Bayes, RF, JRip, KNN, Multilayer Perceptron	80-85% (collective)
Pedrotti et al. (2014)	Pupil Data, Electrodermal Activity	Neural network	79.2% (collective)
Aracena et al. (2015)	Pupil Data	Neural network, Decision Tree	70-78% (subject based)
Baltaci & Gokcay (2016)	Pupil Data, Facial Temperature	Adaboost, RF, Bagging, Decision Tree	80-84% (collective)
Our Study	Pupil Data	Adaboost, RF, SVM	87.75-93.33% (best subject based) 61.96% (analysis 1 - collective) 61.85% (analysis 2 - collective) 69.81% (analysis 3 - collective)

grayscale datasets with 69.81% accuracy at best. This indicates there is a significant difference between the pupil responses. We assumed that this difference arises from the emotional load due to the presence of color.

5.1 Limitations and Suggestions for Future Work

There are some important limitations in our study regarding emotions, selected stimuli and data analyses. First of all, although statistically significant differences were observed between groups, the number of participants are not enough to generalize these results to entire society. Besides, even though high accuracy results are obtained on subject-based analyses, the results of some subjects reveal no stress identification. Therefore, during participant selection, different questionnaires such as Emotion Quotient Test might be used to increase the performance of emotion detec-

tion experiments.

The human-computer interaction is an active environment with unpredictable shifts between emotions due to dynamic real time stimuli. Although we conducted our experiments in near real time, our stimuli did not evolve dynamically. Therefore, it is needed to test the detection of stress in real-life emotional environment. In addition, some subjects stated that they felt excitement while starting the experiment. Although we designed such an experiment that focused on increasing stress level on second phase, some subjects might feel such positive stress during first and second phases and this might effect the accuracy of results. For the third analysis, where we examined the effect of emotional factors, we assumed that colorless (grayscale) images do not cause emotional load, however, because of personal experiences and association of objects with specific colors, these images might result in unwanted emotional arousal – perhaps related to nostalgy in grayscale images. This is basically due to the difficulty of creating an emotionless stimuli.

Another limitation of this study is related to the gender distribution of the participants. The gathered data might be biased since the majority of the participants are female in both experiments. There might be a physiological difference between the stress levels of different genders. Hence, low classification accuracies for the collective analysis might result from this issue.

We classified our data by using some of the state-of-the-art machine learning algorithms. To achieve a higher classification accuracy, an intelligent fine-tuning of the classification parameters is required. In addition, other classification algorithms should be utilized and especially, deep learning method should be explored. To use deep learning technique and to improve the performance of classifiers, larger amount of experimental data should be gathered. Another approach to improve the classification accuracies might be to perform feature selection methods, such as PCA or information gain. We divided our feature set into three categories and evaluated our data by using each feature set separately. The results show that using each feature set did not improve the accuracies. So, instead of categorising features it would be better to use feature selection methods for future research.

We obtained better results in subject-based analyses compared to collective analy-

ses due to within group heterogeneity. It means that the stress detection in human-computer interaction might require an active intervention by the user's side which is not suitable for an automatic emotion detection system. With this perspective, future research for automatic emotion detection system might be investigated with particular attention toward facial expressions and facial temperature.

Another suggestion for future work is to combine classifiers with decision level fusion to obtain higher accuracy results. In our classification accuracies, we observed that for one subject, one classifier gave higher accuracy while other classifier underperformed. So, with the fusion of classifiers, it would be possible to classify data better.

The experimental setup can be designed differently for the future studies. One suggestion is to ask participants to count the number of arrows on a given image stimuli. Different from this study, different grid sizes can be used for the background image in each phase, i.e., arrows with no image (blank background) for the first phase (baseline), arrows on top of a single image for the second phase, arrows on top of a 2×2 image grid for the third phase and arrows on top of a $i \times i$ image grid for the i^{th} phase. In this scenario, since the background is further distorted in each phase, the cognitive load on the participants is expected to increase. Another suggestion is to add one more experiment by scrambling the colored stimuli used in this study. By this way, the effect of background distortion with the existence of color can be examined. One final suggestion is to rotate the images used in the stimuli. In this study, the objects in the stimuli are represented with their natural orientation in daily life. The cognitive load can be increased by representing the objects with unnatural orientations. For example, illustrating an upside down coffee cup might distract the attention of participants. Apart from that, during experiment the eye movements can be used to adjust and analyse the data. Since TOBII eye tracker assumes that pupils focus on the middle of the screen, but participants move their eyes throughout screen to perform the task, the position of pupils and the angle of the view can be included for data analysis to improve the quality of the pupil diameter measurements. It is also possible to evaluate the peak points of the data and whether these coincide with the time of noticing arrows on the stimuli in order to find a relevant relationship between the gaze and pupil diameter.

CHAPTER 6

CONCLUSION

In this thesis, we investigated the effects of cognitive and emotional factors on pupil dilation in a human-computer interaction environment. For this purpose, we performed experiments with eye-tracker system and used IAPS neutral images as stimuli. We recorded pupillary responses during experiments in order to analyze them in supervised classifiers. The reason for using machine learning classifiers is to verify that cognitive and emotional factors induce distinguishable changes on the pupillary responses. Our results indicate that cognitive and emotional factors cause greater pupillary response compared to neutral state.

We had two different experiments where both of them focused on increasing cognitive load. We also examined the effect of color as an emotional factor by analyzing normalized pupillary responses on a collective basis. With this perspective, this thesis achieved detection rates comparable to the existing literature with respect to both cognitive and emotional loads. This work also shows that the emotional response for perceiving colors can be detectable from pupillary response.

The stress detection method of this study consists of featured-based classification. Beside statistical features, exploring the use of entropy, regression slope and polynomial functions can be considered as the main contributions of this thesis. The pre-processing, feature extraction and classification methods presented in this thesis are appropriate for emotion detection because they are practical and fast.

Our thesis obtained the accuracy results by using three machine learning algorithms, i.e., SVM, Adaboost and Random Forest. Based on our classification results, we observed that subject-based detection of cognitive load from pupillary responses is

feasible. Our best result was 93.33% in the first analysis and 87.5% in the second analysis. This indicates that cognitive factors generate larger pupil dilations. It also shows the impact of subjective differences on a group accuracy. Additionally, based on collective analyses, we noticed that the accuracy result of emotional load is higher than the results of cognitive effect.

To conclude, we simulated a human-computer interaction environment and performed stress detection techniques unobtrusively. Our results verify our current knowledge concerning the impact of cognitive and emotional factors on pupil dilation. With other classification techniques and dynamic stimuli, the presented method can be used for real-time emotion classification in a customized manner for individual subjects.

REFERENCES

A. Just, M., & A. Carpenter, P. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian journal of experimental psychology = Revue canadienne de psychologie expérimentale*, *47*, 310-39.

Andreassi, J. (2006). Pupillary response and behavior. *Psychophysiology: Human Behavior and Physiological Response*, 350-371.

Aracena, C., Basterrech, S., Snáel, V., & Velásquez, J. (2015). Neural networks for emotion recognition based on eye tracking data. In (p. 2632-2637). doi: 10.1109/SMC.2015.460

Baltaci, S., & Gokcay, D. (2016). Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features. *International Journal of Human–Computer Interaction*, *32*(12), 956-966. Retrieved from `https://doi.org/10.1080/10447318.2016.1220069` doi: 10.1080/10447318.2016.1220069

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behavior* (Vol. 4, p. 71-81). New York: Academic Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY, US: W H Freeman/Times Books/ Henry Holt & Co.

Beale, R., Peter, C., Palen, L., Bødker, S., Bainbridge, W., Lichtenstein, A., ... Jonsson, I.-M. (2008). *Affect and emotion in human-computer interaction.*

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276-292.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In G. Berntson & L. G. Tassinar (Eds.), *Handbook of psychophysiology* (p. 142-162). Hillsdale, NJ: Cambridge University Press.

Beatty, J., & L Wagoner, B. (1978). Pupillometric signs of brain activation vary with level of cognitive processing. *Science (New York, N.Y.)*, *199*, 1216-8.

Beck, A., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An in-

ventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561-571. Retrieved from http://dx.doi.org/10.1001/archpsyc.1961.01710120031004 doi: 10.1001/archpsyc.1961.01710120031004

Bourne, E, L., & A. Yaroush, R. (2003). Stress and cognition: A cognitive psychological perspective. *Unpublished manuscript, NASA grant NAG2-1561*.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49 - 59. Retrieved from http://www.sciencedirect.com/science/article/pii/0005791694900639 doi: https://doi.org/10.1016/0005-7916(94)90063-9

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45 4*, 602-7.

Cocker, K. (1996). Development of pupillary responses to grating stimuli. *Ophthalmic and Physiological Optics*, *16*, 64-67.

Cohen, I., Sebe, N., Chen, L., Garg, A., & Huang, T. S. (2003). Facial expression recognition from video sequences: Temporal and static modelling. In (pp. 160–187).

Collyer, S., & Malecki, G. (1998). Tactical decision making under stress: History and overview. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (p. 3-15). Washington, DC, US: American Psychological Association.

de Santos Sierra, A., Sánchez Ávila, C., Casanova, J., & Bailador, G. (2011). Real-time stress detection by means of physiological signals. In (p. 23-44).

Dinges, D. F., Venkataraman, S., McGlinchey, E. L., & Metaxas, D. N. (2007). Monitoring of facial stress during space flight: Optical computer recognition combining discriminative and generative methods. *Acta Astronautica*, *60*(4), 341 - 350. Retrieved from http://www.sciencedirect.com/science/article/pii/S0094576506003079 (Benefits of human presence in space - historical, scientific, medical, cultural and political aspects. A selection of papers presented at the 15th IAA Humans in Space Symposium, Graz, Austria, 2005) doi: https://doi.org/10.1016/j.actaastro.2006.09.003

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice*. Berlin,

Heidelberg: Springer-Verlag.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (No. 57). Boca Raton, Florida, USA: Chapman & Hall/CRC.

Ekman, P., & Friesen, W. V. (1982, 01). Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, *6*(4), 238–252. Retrieved from https://doi.org/10.1007/BF00987191 doi: 10.1007/BF00987191

Ekman, P., & V. Friesen, W. (2003). Unmasking the face: A guide to recognizing emotions from facial clues.

Eriksen, C. W., Lazarus, R. S., & Strange, J. R. (n.d.). Psychological stress and its personality correlates. *Journal of Personality*, *20*(3), 277-286. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.1952.tb01110.x doi: 10.1111/j.1467-6494.1952.tb01110.x

Folkman, S., S. Lazarus, R., Schetter, C., DeLongis, A., & Gruen, R. (1986). Dynamics of a stressful encounter: Cognitive appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology*, *50*, 992-1003.

Gaillard, A. W. K. (1993). Comparing the concepts of mental load and stress. *Ergonomics*, *36*(9), 991-1005. Retrieved from https://doi.org/10.1080/00140139308967972 (PMID: 8404841) doi: 10.1080/00140139308967972

Gao, H., Yüce, A., & Thiran, J. P. (2014). Detecting emotional stress from facial expressions for driving safety. In (p. 5961-5965). doi: 10.1109/ICIP.2014.7026203

Gao, X., & Xin, J. (2006). Investigation of human's emotional responses on colors. *Color Research & Application*, *31*, 411 - 417.

Gao, Y., Adjouadi, M., Ren, P., & Barreto, A. (2013). Affective assessment by digital processing of the pupil diameter. *IEEE Transactions on Affective Computing*, *4*, 2-14. Retrieved from doi.ieeecomputersociety.org/10.1109/T-AFFC.2012.25 doi: 10.1109/T-AFFC.2012.25

Gençöz, T. (2000). Pozitif ve negatif duygu ölçeği: Geçerlik ve güvenirlik çalışması. *Türk Psikoloji Derneği*, *15*(46), 19-26.

Gerald Matthews, S. J. W. . R. B. S., D. Roy Davies. (2000b). *Human performance: Cognition, stress and individual differences*. Philadelphia, PA: Taylor and Fran-

cis.

Hakerem, G. (1967). Pupillography. In P. Venables & I. Martin (Eds.), *A manual of psychophysiological methods* (p. 335-349). Amsterdam: North Holland Publishing Co.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, *212*, 46-54. doi: 10.1038/scientificamerican0465-46

Hess, E. H. (1972). Pupillometrics: A method of studying mental, emotional and sensory processes. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 491–531). New York: Holt, Rinehart & Winston.

H. Hess, E., & M. Polt, J. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science (New York, N.Y.)*, *143*, 1190-2.

Hudlicka, E. (2002). This time with feeling: Integrated model of trait and state effects on cognition and behavior. *Applied Artificial Intelligence*, *16*(7-8), 611-641. Retrieved from `https://doi.org/10.1080/08339510290030417` doi: 10.1080/08339510290030417

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies*, *59*, 1-32.

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, *108*(1), 116 - 134. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1077314206002335` (Special Issue on Vision for Human-Computer Interaction) doi: https://doi.org/10.1016/j.cviu.2006.10.019

Jiang, X., Zheng, B., Bednarik, R., & Atkins, M. (2015). Pupil responses to continuous aiming movements. *International Journal of Human-Computer Studies*, *83*.

J Lang, P., M Bradley, M., & N Cuthbert, B. (2008). International affective picture system (iaps): Affective ratings of pictures and instruction manual (rep. no. a-8). *Technical Report A-8*.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kinner, V., Kuchinke, L., Dierolf, A., Merz, C., Otto, T., & T. Wolf, O. (2017). What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes: Pupillary responses during emotion regulation. *Psy-*

chophysiology, *54*.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In (pp. 69–72). New York, NY, USA: ACM. doi: 10.1145/1344471.1344489

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*, 323-32.

Lazarus, R. S. (1999). *Stress and emotion: A new synthesis*. New York, NY, US: Springer Publishing Co.

Loewy, A. D. (1990). Autonomic control of the eye. In A. Loewy & K. Spyer (Eds.), *Central regulation of autonomic functions* (p. 268-285). New York: Oxford University Press.

Mandler, G. (1984). *Mind and body: The psychology of emotion and stress*. New York: Norton.

Marshall, S. P. (2002). The index of cognitive activity: measuring cognitive workload. In (p. 7-5-7-9). doi: 10.1109/HFPP.2002.1042860

Ou, L.-C., Luo, M., Woodcock, A., & Wright, A. (2004). A study of colour emotion and colour preference. part i: Colour emotions for single colours. *Color Research & Application*, *29*, 232 - 240.

Partala, T., Jokiniemi, M., & Surakka, V. (2000). *Pupillary responses to emotionally provocative stimuli.*

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1), 185 - 198. Retrieved from `http://www.sciencedirect.com/science/article/pii/S107158190300017X` (Applications of Affective Computing in Human-Computer Interaction) doi: https://doi.org/10.1016/S1071-5819(03)00017-X

Pedrotti, M., Mirzaei, M. A., Tedesco, A., Chardonnet, J.-R., Mérienne, F., Benedetto, S., & Baccino, T. (2014). Automatic stress classification with pupil diameter analysis. *International Journal of Human–Computer Interaction*, *30*(3), 220-236. Retrieved from `https://doi.org/10.1080/10447318.2013.848320` doi: 10.1080/10447318.2013.848320

Picard, R. W. (1997). *Affective computing*. Cambridge, MA, USA: MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175-1191. doi: 10.1109/34.954607

Plechawska-Wójcik, M., & Borys, M. (2016). An analysis of eeg signal combined with pupillary response in the dynamics of human cognitive processing. In (p. 378-385). doi: 10.1109/HSI.2016.7529661

Plutchik, R. (2001). The nature of emotions. *American Scientist*, *89*, 344.

Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. *Proceedings of the International Conference on HCI*.

Rani, P., Sarkar, N., Smith, C. A., & Adams, J. A. (2003). Affective communication for implicit human-machine interaction. In (Vol. 5, p. 4896-4903 vol.5). doi: 10.1109/ICSMC.2003.1245758

Roesch, S., Weiner, B., & Vaughn, A. (2002). Cognitive approaches to stress and coping. *Current Opinion in Psychiatry*, *15*, 627-632.

Sahraie, A., & Barbur, J. L. (1997). Pupil response triggered by the onset of coherent motion. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *235*(8), 494–500. Retrieved from `https://doi.org/10.1007/BF00947006` doi: 10.1007/BF00947006

Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers*, *14*(2), 93-118. Retrieved from `http://dx.doi.org/10.1016/S0953-5438(01)00059-5` doi: 10.1016/S0953-5438(01)00059-5

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Smith, S. W. (1997-98). *The scientist and engineer's guide to digital signal processing*. California Technical Publishing.

Snowden, R., R O'Farrell, K., Burley, D., Erichsen, J., V Newton, N., & Gray, N. (2016). The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology*, *53*.

Sroykham, W., Wongsathikun, J., & Wongsawat, Y. (2014). The effects of perceiving color in living environment on qeeg, oxygen saturation, pulse rate, and emotion

regulation in humans. In (p. 6226-6229). doi: 10.1109/EMBC.2014.6945051

Surakka, V., & Hietanen, J. (1998). Facial and emotional reactions to duchenne and non-duchenne smiles. In (Vol. 29, p. 23-33).

Tobii Technology, I. (2011). Tobii t60 & t120 eye tracker user manual [Computer software manual].

Tortora, G. J. (1987). *Principles of anatomy and physiology*. Harper and Row.

Ukai, K. (1985). Spatial pattern as a stimulus to the pupillary system. *J. Opt. Soc. Am. A*, *2*(7), 1094–1100. Retrieved from `http://josaa.osa.org/abstract.cfm?URI=josaa-2-7-1094` doi: 10.1364/JOSAA.2.001094

Valdez, P., & Mehrabian, A. (1995). Effects of color on emotions. *Journal of experimental psychology. General*, *123*, 394-409.

Watson, D., Anna Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review*, *92*, 548-73.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques (third edition)*.

Young, R. S., Han, B.-C., & Wu, P.-Y. (1993). Transient and sustained components of the pupillary responses evoked by luminance and color. *Vision Research*, *33*(4), 437 - 446. Retrieved from `http://www.sciencedirect.com/science/article/pii/004269899390251Q` doi: https://doi.org/10.1016/0042-6989(93)90251-Q

Yuen, P., Hong, K., Chen, T., Tsitiridis, A., Kam, F., Jackman, J., ... Lightman, S. (2009). Emotional amp; physical stress detection and classification using thermal imaging technique. In (p. 1-6). doi: 10.1049/ic.2009.0241

Zhai, J., & Barreto, A. (2006). Stress detection in computer users through non-invasive monitoring of physiological signals. *Biomedical sciences instrumentation*, *42*, 495-500.

Zhai, J., Barreto, A. B., Chin, C., & Li, C. (2005). Realization of stress detection using psychophysiological signals for improvement of human-computer interactions. In *Proceedings. ieee southeastcon, 2005.* (p. 415-420). doi:

10.1109/SECON.2005.1423280

APPENDIX A

BECK DEPRESSION INVENTORY

Aşağıda, kişilerin ruh durumlarını ifade ederken kullandıkları bazı cümleler verilmiştir. Her madde, bir çeşit ruh durumunu anlatmaktadır. Her maddede o ruh durumunun derecesini belirleyen 4 seçenek vardır. Lütfen bu seçenekleri dikkatle okuyunuz. Son bir hafta içindeki (şu an dahil) kendi durumunuzu göz önünde bulundurarak, size en uygun ifadeyi bulunuz. Daha sonra o maddenin yanındaki harfin üzerine (X) işareti koyunuz.

1. (a) Kendimi üzgün hissetmiyorum.
   (b) Kendimi üzgün hissediyorum.
   (c) Her zaman için üzgünüm ve kendimi bu duygudan kurtaramıyorum.
   (d) Öylesine üzgün ve mutsuzum ki dayanamıyorum.

2. (a) Gelecekten umutsuz değilim.
   (b) Geleceğe biraz umutsuz bakıyorum.
   (c) Gelecekten beklediğim hiçbir şey yok.
   (d) Benim için bir gelecek yok ve bu durum düzelmeyecek.

3. (a) Kendimi başarısız görmüyorum.
   (b) Çevremdeki birçok kişiden daha fazla başarısızlıklarım oldu sayılır.
   (c) Geriye dönüp baktığımda, çok fazla başarısızlığımın olduğunu görüyorum.
   (d) Kendimi tümüyle başarısız bir insan olarak görüyorum.

4. (a) Her şeyden eskisi kadar zevk alabiliyorum.

(b) Her şeyden eskisi kadar zevk alamıyorum.

(c) Artık hiçbir şeyden gerçek bir zevk alamıyorum.

(d) Bana zevk veren hiçbir şey yok. Her şey çok sıkıcı.

5. (a) Kendimi suçlu hissetmiyorum.

(b) Arada bir kendimi suçlu hissettiğim oluyor.

(c) Kendimi çoğunlukla suçlu hissediyorum.

(d) Kendimi her an için suçlu hissediyorum.

6. (a) Cezalandırıldığımı düşünmüyorum.

(b) Bazı şeyler için cezalandırılabileceğimi hissediyorum.

(c) Cezalandırılmayı bekliyorum.

(d) Cezalandırıldığımı hissediyorum.

7. (a) Kendimden hoşnutum.

(b) Kendimden pek hoşnut değilim.

(c) Kendimden hiç hoşlanmıyorum.

(d) Kendimden nefret ediyorum.

8. (a) Kendimi diğer insanlardan daha kötü görmüyorum.

(b) Kendimi zayıflıklarım ve hatalarım için eleştiriyorum.

(c) Kendimi hatalarım için çoğu zaman suçluyorum.

(d) Her kötü olayda kendimi suçluyorum.

9. (a) Kendimi öldürmek gibi düşüncelerim yok.

(b) Bazen kendimi öldürmeyi düşünüyorum, fakat bunu yapmam.

(c) Kendimi öldürebilmeyi isterdim.

(d) Bir fırsatını bulsam kendimi öldürürdüm.

10. (a) Her zamankinden daha fazla ağladığımı sanmıyorum.

(b) Eskisine göre şu sıralarda daha fazla ağlıyorum.

(c) Şu sıralarda her an ağlıyorum.

(d) Eskiden ağlayabilirdim, ama su sıralarda istesem de ağlayamıyorum.

11. (a) Her zamankinden daha sinirli değilim.

    (b) Her zamankinden daha kolayca sinirleniyor ve kızıyorum.

    (c) Çoğu zaman sinirliyim.

    (d) Eskiden sinirlendiğim şeylere bile artık sinirlenemiyorum.

12. (a) Diğer insanlara karşı ilgimi kaybetmedim.

    (b) Eskisine göre insanlarla daha az ilgiliyim.

    (c) Diğer insanlara karşı ilgimin çoğunu kaybettim.

    (d) Diğer insanlara karşı hiç ilgim kalmadı.

13. (a) Kararlarımı eskisi kadar kolay ve rahat verebiliyorum.

    (b) Şu sıralarda kararlarımı vermeyi erteliyorum.

    (c) Kararlarımı vermekte oldukça güçlük çekiyorum.

    (d) Artık hiç karar veremiyorum.

14. (a) Dış görünüşümün eskisinden daha kötü olduğunu sanmıyorum.

    (b) Yaslandığımı ve çekiciliğimi kaybettiğimi düşünüyor ve üzülüyorum.

    (c) Dış görünüşümde artık değiştirilmesi mümkün olmayan olumsuz değişiklikler olduğunu hissediyorum.

    (d) Çok çirkin olduğumu düşünüyorum.

15. (a) Eskisi kadar iyi çalışabiliyorum.

    (b) Bir işe başlayabilmek için eskisine göre kendimi daha fazla zorlamam gerekiyor.

    (c) Hangi iş olursa olsun, yapabilmek için kendimi çok zorluyorum.

    (d) Hiçbir iş yapamıyorum.

16. (a) Eskisi kadar rahat uyuyabiliyorum.

    (b) Şu sıralarda eskisi kadar rahat uyuyamıyorum.

    (c) Eskisine göre 1 veya 2 saat erken uyanıyor ve tekrar uyumakta zorluk çekiyorum.

    (d) Eskisine göre çok erken uyanıyor ve tekrar uyuyamıyorum.

17. (a) Eskisine kıyasla daha çabuk yorulduğumu sanmıyorum.

    (b) Eskisinden daha çabuk yoruluyorum.

    (c) Şu sıralarda neredeyse her şey beni yoruyor.

    (d) Öyle yorgunum ki hiçbir şey yapamıyorum.

18. (a) İştahım eskisinden pek farklı değil.

    (b) İştahım eskisi kadar iyi değil.

    (c) Şu sıralarda iştahım epey kötü.

    (d) Artık hiç iştahım yok.

19. (a) Son zamanlarda pek fazla kilo kaybettiğimi sanmıyorum.

    (b) Son zamanlarda istemediğim halde üç kilodan fazla kaybettim.

    (c) Son zamanlarda istemediğim halde beş kilodan fazla kaybettim.

    (d) Son zamanlarda istemediğim halde yedi kilodan fazla kaybettim.
    Daha az yemeye çalışarak kilo kaybetmeye çalışıyorum.
    Evet ( ) Hayır( )

20. (a) Sağlığım beni pek endişelendirmiyor.

    (b) Son zamanlarda ağrı, sızı, mide bozukluğu, kabızlık gibi sorunlarım var.

    (c) Ağrı, sızı gibi bu sıkıntılarım beni epey endişelendirdiği için başka şeyleri düşünmek zor geliyor.

    (d) Bu tür sıkıntılarım beni öylesine endişelendiriyor ki, artık başka hiçbir şey düşünemiyorum.

21. (a) Son zamanlarda cinsel yaşantımda dikkatimi çeken bir şey yok.

    (b) Eskisine oranla cinsel konularla daha az ilgileniyorum.

    (c) Şu sıralarda cinsellikle pek ilgili değilim.

    (d) Artık cinsellikle hiçbir ilgim kalmadı.

APPENDIX B

POSITIVE AND NEGATIVE AFFECT SCALE (PANAS)

Bu ölçek farklı duyguları tanımlayan bir takım sözcükler içermektedir. Şu anda nasıl hissettiğinizi düşünüp her maddeyi okuyun. Uygun cevabı her maddenin yanında ayrılan yere (puanları daire içine alarak) işaretleyin. Cevaplarınızı verirken aşağıdaki puanları kullanın.

1. Çok az veya hiç 2. Biraz 3. Ortalama 4. Oldukça 5. Çok fazla

1. İlgili	1	2	3	4	5
2. Sıkıntılı	1	2	3	4	5
3. Heyecanlı	1	2	3	4	5
4. Mutsuz	1	2	3	4	5
5. Güçlü	1	2	3	4	5
6. Suçlu	1	2	3	4	5
7. Ürkmüş	1	2	3	4	5
8. Düşmanca	1	2	3	4	5
9. Hevesli	1	2	3	4	5
10. Gururlu	1	2	3	4	5
11. Asabi	1	2	3	4	5
12. Uyanık (dikkati açık)	1	2	3	4	5
13. Utanmış	1	2	3	4	5
14. İlhamlı (yaratıcı düşüncelerle dolu)	1	2	3	4	5
15. Sinirli	1	2	3	4	5
16. Kararlı	1	2	3	4	5
17. Dikkatli	1	2	3	4	5
18. Tedirgin	1	2	3	4	5
19. Aktif	1	2	3	4	5
20. Korkmuş	1	2	3	4	5

APPENDIX C

INFORMED CONSENT FORM

Orta Doğu Teknik Üniversitesi İnsan Araştırmaları Etik Kurulu Gönüllü Katılım (Bilgilendirilmiş Onay) Formu

Orta Doğu Teknik Üniversitesi Enformatik Enstitüsü Sağlık Bilişimi Bölümü öğretim üyelerinden Y. Doç. Dr. Didem Gökçay tarafından yürütülmekte olan "Göz bebeği kayıtlarından bilişsel ve afektif süreç kestirimi" adlı araştırmaya katılmak için seçildiniz. Çalışmaya katılım gönüllülük esasına dayalıdır. Kararınızdan önce araştırma hakkında sizi bilgilendirmek istiyoruz. Bilgileri okuyup anladıktan sonra araştırmaya katılmak isterseniz lütfen bu formu imzalayınız.

Günlük hayatımızda olumluluk açısından farklı içeriklere sahip birçok görsel uyaranla karşılaşmaktayız. Bu uyaranların aynı zamanda heyecan verici olma, stres yaratma gibi özellikleri de bulunabilir. Bu araştırmayı yapma nedenimiz tüm bu maruz kaldığımız uyaranların duygusal ve bilişsel süreçlerde davranışlarımız üzerindeki etkilerini incelemektir.

Çalışma sırasında sizden 2 deneye katılmanız ve her birinde 20'şer resmi değerlendirmeniz istenmektedir. Değerlendirme sırasında görsel uyaranın üzerindeki okları saymanız ve odaklanma noktasını gördüğünüz sırada ok sayısını sesli bir şekilde belirtmeniz beklenmektedir. Daha sonra size söylediğiniz sayı üzerine doğru ya da yanlış olduğuna dair bir geri bildirme yapılacaktır. Çalışmaya katılmayı kabul ettiğiniz takdirde, deneyin işleyişi hakkında bilgilendirileceksiniz. Çalışma süresi yaklaşık yarım saat olarak planlanmıştır.

Bu çalışmada gözbebeği büyümesini ve hareketlerini takip edip kayıt altına almak için bir göz izleme cihazı kullanılmaktadır. Bu cihazlar insan sağlığı ya da ruhsal

durumu açısından en ufak bir risk teşkil etmemektedir.

Bu formu imzalayarak araştırmaya katılım için onay vermiş olacaksınız. Çalışmayı tamamladığınız takdirde, kimlik bilgileriniz çalışmanın herhangi bir aşamasında açıkça kullanılmayacaktır. Doldurduğunuz anketlere verdiğiniz cevaplar ve araştırma süresince görsel cihaz kullanılarak edinilen her türlü bilgi yalnızca bilimsel amaçlar için kullanılacaktır. Bilgileriniz hiçbir kimse ile ya da ticari bir amaç için paylaşılmayacaktır.

Çalışma hakkında daha fazla bilgi edinmek için aşağıda belirtilen araştırmacılarla iletişime geçebilirsiniz.

Y. Doç. Dr. Didem Gökçay, ODTÜ Enformatik Enstitüsü, A-216, (312) 210 3750, dgokcay@metu.edu.tr

---

"Göz bebeği kayıtlarından bilişsel ve afektif süreç kestirimi" çalışması hakkında bilgilendirildim. Çalışmayı istediğim zaman terk edebileceğimi ve bana ait kişisel bilgilerle beraber benden toplanan kişisel değerlendirmelerin hiçbir zaman açıkça kullanılmayacağını biliyorum. Bu çalışmaya gönüllü olarak katılıyorum. Kişisel bilgilerimin hiçbir şekilde paylaşılmayacağını, sadece deneydeki performans faktörlerini etkileyebileceği için sorulduğunu biliyorum.

Ad, Soyad:
Tarih:
İmza:

APPENDIX D

DEMOGRAPHIC INFORMATION FORM

Kişisel Bilgiler:

Uygulama Tarihi: . . . / . . . / . . .

Adı Soyadı:

Cinsiyeti: Kadın ( ) Erkek ( )

Doğum Tarihi: . . . / . . . / . . .

Yaşı: . . .

Medeni Hali: Evli ( ) Bekar ( ) Dul ( ) Boşanmış ( )

Mesleği: . . . . . .

El Tercihi: Sağ ( ) Sol ( )

Eğitim Durumu:

İlkokul (0-5 yıl) ( )

Ortaokul (6-8 yıl) ( )

Lise (9-12 yıl) ( )

Üniversite (12+) ( )

Sağlık Durumuna İlişkin Bilgiler:

İşitme Bozukluğu: Var ( ) Yok ( )

Varsa düzeltilmiş mi? . . . . . .

Görme Bozukluğu var mı? Var ( ) Yok ( )

Varsa hangisi? Miyop ( ) Astigmat ( ) Hipermetrop ( )

Varsa düzeltilmiş mi? .....

Renk Körlüğü: Var ( ) Yok ( )

Fiziksel Özür: Var ( ) Yok ( )

Varsa türü: .....

Geçirdiği Önemli Rahatsızlıklar (Psikiyatrik, Nörolojik veya Psikolojik):

Halen Kullanmakta Olduğu İlaç: Var ( ) Yok ( )

Varsa ilacın/ilaçların adı: .....

Uzun Süre Kullanıp Bıraktığı İlaç: Var ( ) Yok ( )

Varsa ilacın/ilaçların adı: .....

Varsa kullanım süresi: .....

Kadın ise son adet günü: .....

APPENDIX E

DEBRIEFING FORM

1. Deneyimizi nasıl buldunuz?

2. Kendi performansınızı nasıl degerlendiriyorsunuz?

3. Okları tutarlı olarak sayabildiğinizi düşünüyor musunuz?

4. Birinci kısım ile ikinci kısım arasında ne gibi bir fark hissettiniz?

5. Birinci kısım ile ikinci kısım arasında heyacanlanmanızda degişiklik oldu mu?

6. Resimlerin içerikleri hakkında ne düşünüyorsunuz?

7. Eklemek istediğiniz başka şeyler var mı?

APPENDIX F

PLOTS OF ENTROPY LEVEL AND PROBABILITY DISTRIBUTION

The analysis between the entropy levels and the probability of the data was performed on a random trial. The window size was selected as the length of the whole trial. For different discretization levels, the probability of the trial was calculated and evaluated. The lowest and the highest levels give little information about the probability distribution while the middle values have meaningful results. In Figure F.7, the convergence limit was provided such that as the level increases, the entropy value approaches 6.5.

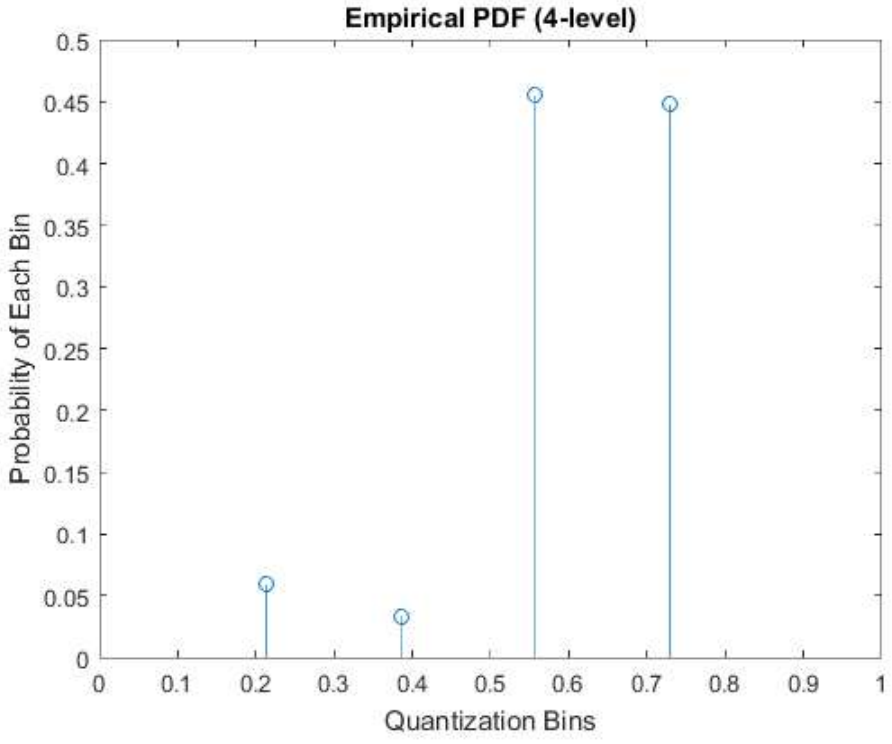


Figure F.1: Probability Distribution of a Random Trial for 4-level discretization

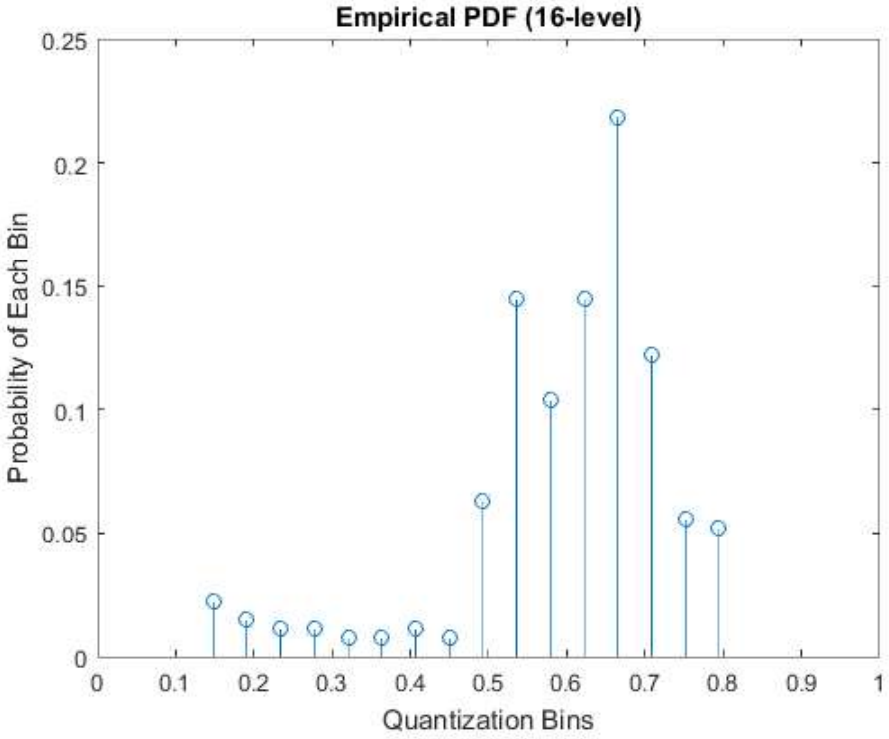


Figure F.2: Probability Distribution of a Random Trial for 16-level discretization

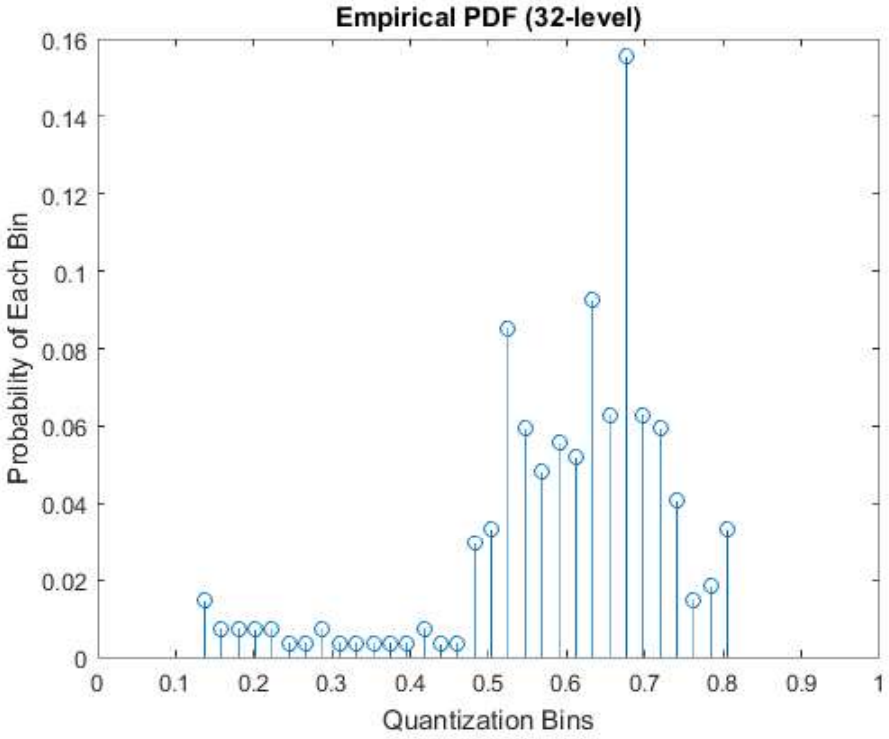


Figure F.3: Probability Distribution of a Random Trial for 32-level discretization

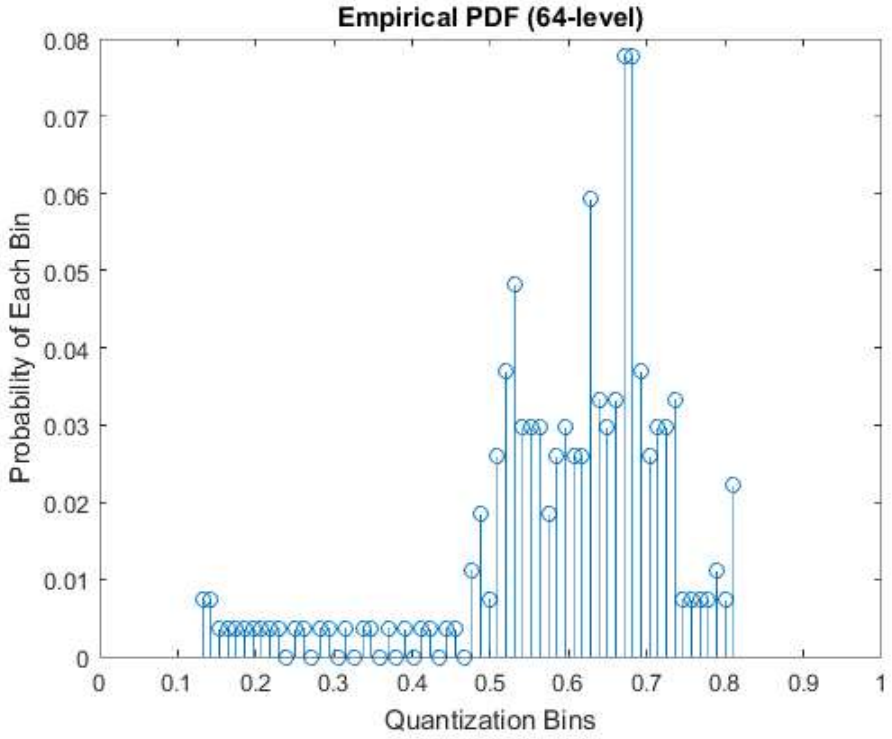


Figure F.4: Probability Distribution of a Random Trial for 64-level discretization

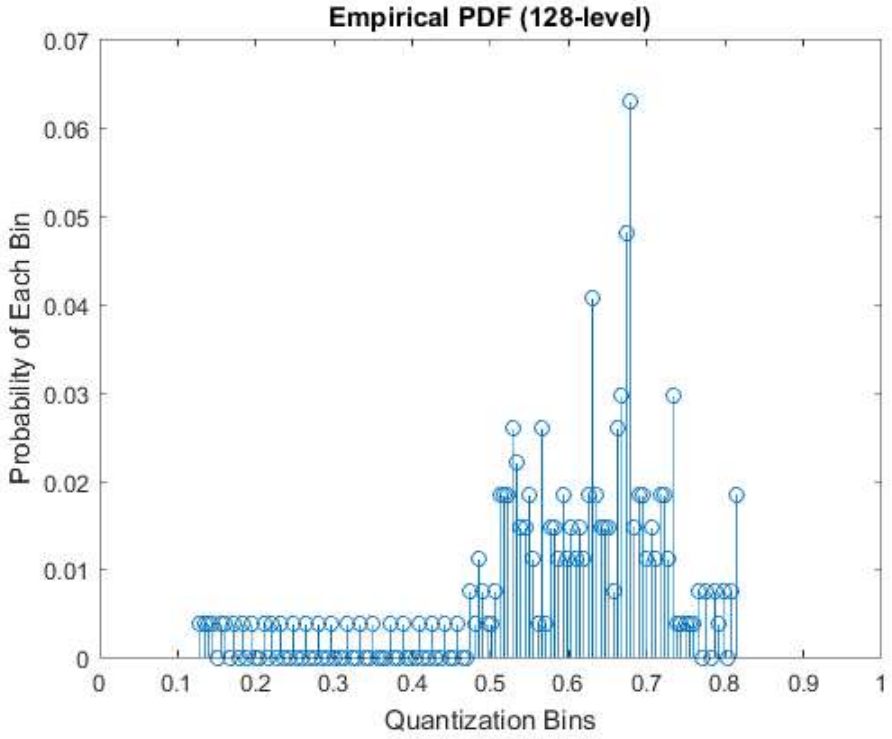


Figure F.5: Probability Distribution of a Random Trial for 128-level discretization

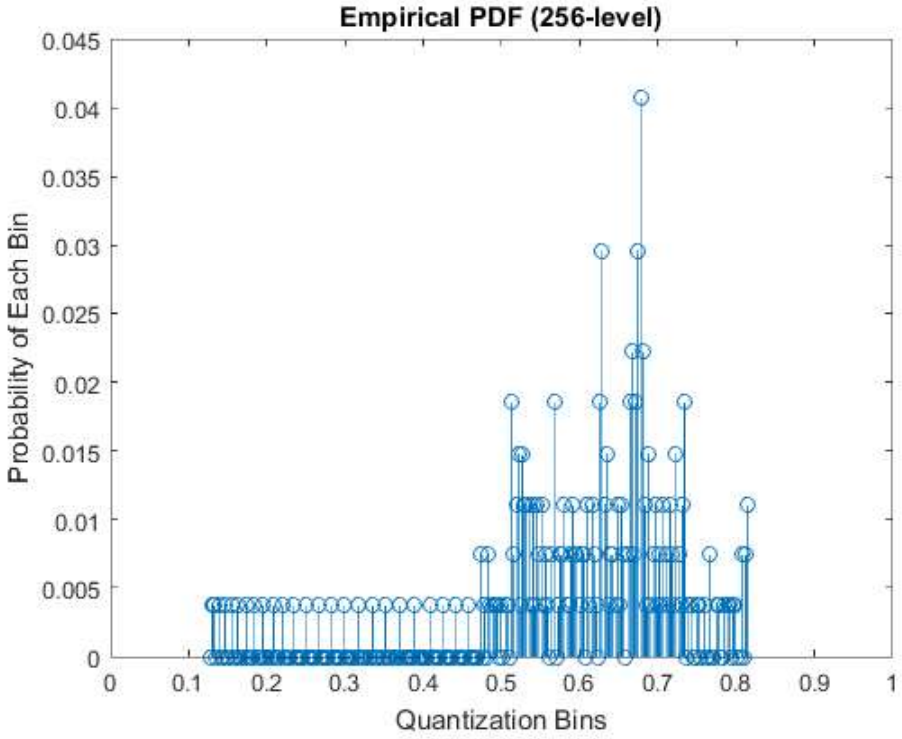


Figure F.6: Probability Distribution of a Random Trial for 256-level discretization

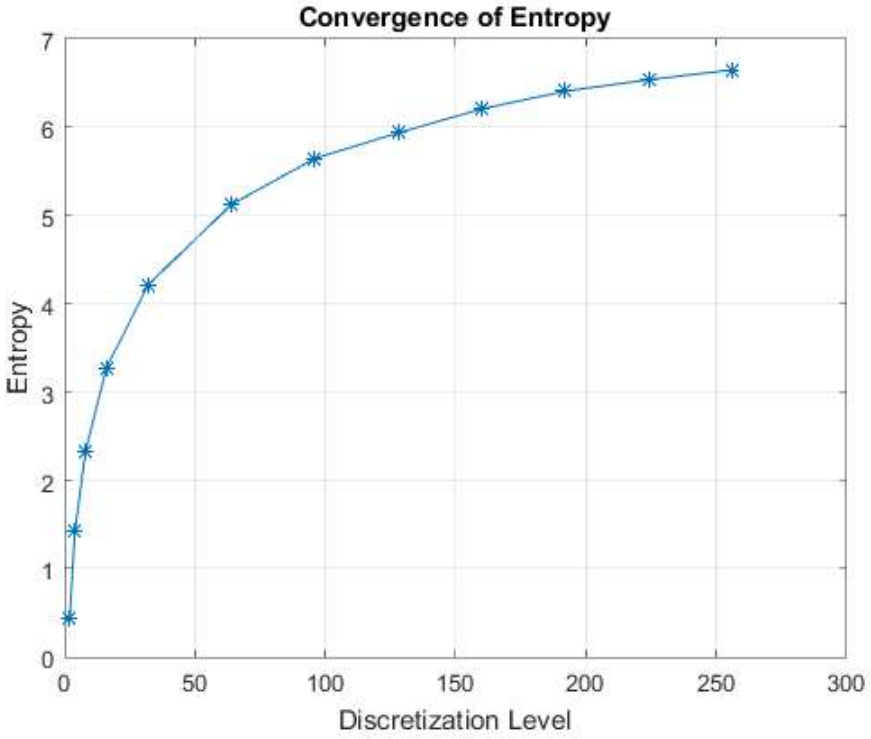


Figure F.7: Entropy Value for Different Discretization Levels

APPENDIX G

PLOTS OF AVERAGED PUPIL DATA

All trials from all subjects were used to calculate averaged pupil data. As expected, Part 2 datasets in Figure G.1 and Figure G.2 have higher pupil diameters compared to the Part 1 datasets. Similarly, in Figure G.3, colored data is higher than grayscale data which supports our Hypothesis.

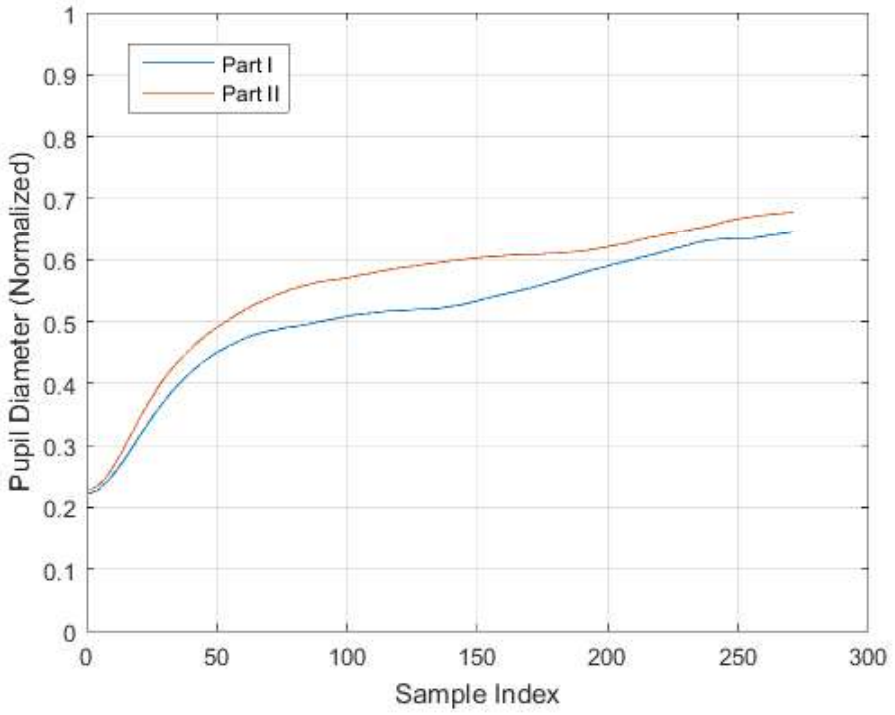


Figure G.1: Averaged Pupil Data for Dataset 1.1 (part 1) & 1.2 (part 2)

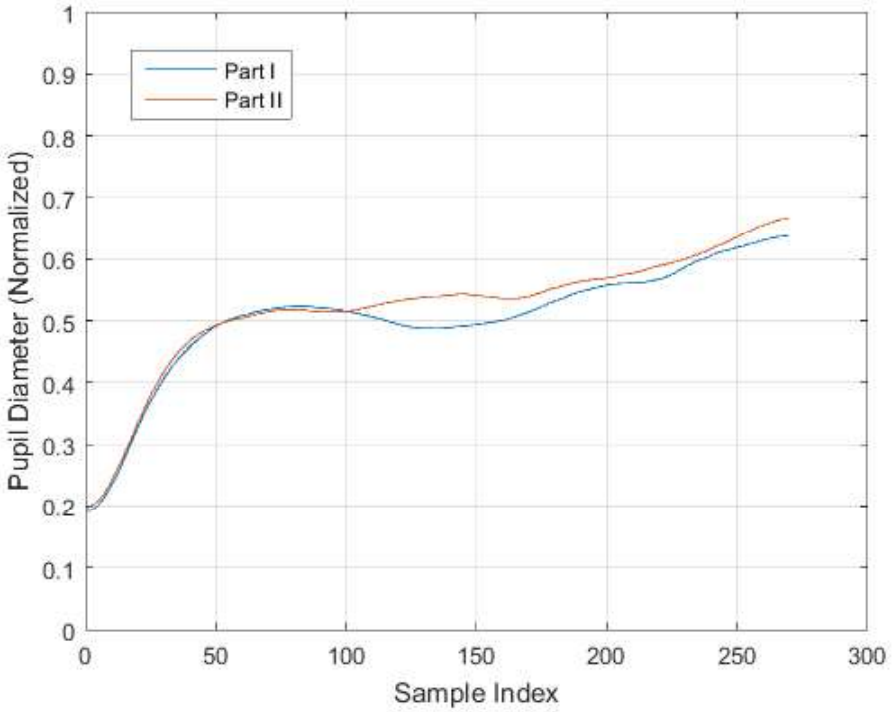


Figure G.2: Averaged Pupil Data for Dataset 2.1 (part 1) & 2.2 (part 2)

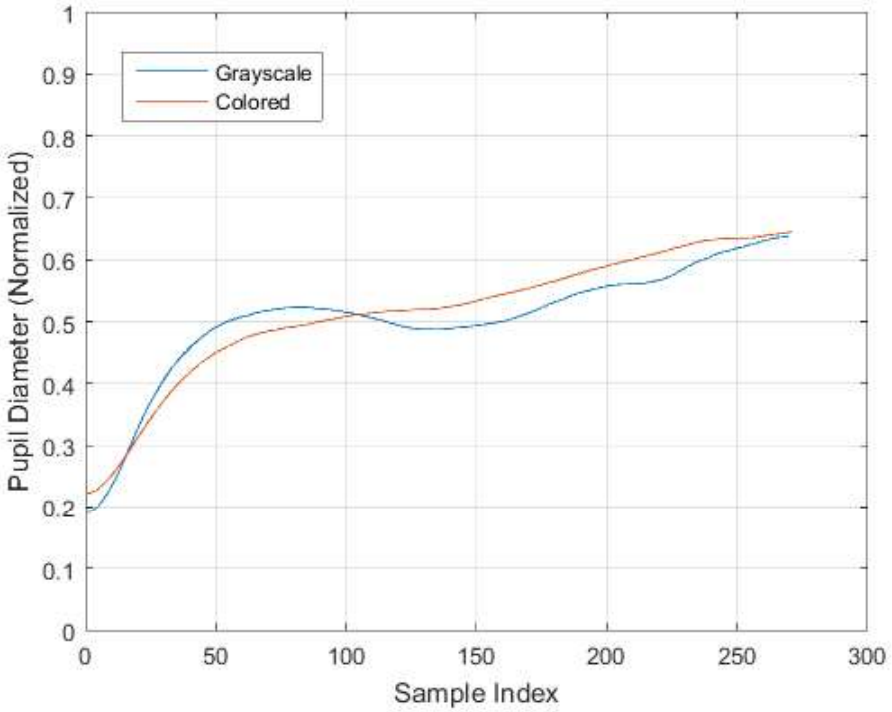


Figure G.3: Averaged Pupil Data for Dataset 1.1 (colored) & 2.1 (graycale)

APPENDIX H

CLASSIFICATION ACCURACIES OF SELECTED ALGORITHMS

Each dataset was analyzed in WEKA with these given classification algorithms. The feature set with 22 features was used without any feature selection methods. All these classifiers were implemented with their default parameters in WEKA. The number of folds was 10, the window size of moving average filter was 20 and the window size of entropy based features was 100 during these analyses. The following accuracy results belong to the collective analyses where all trials from all subjects were put together in a pool to be classified as class 1 or 2. In our case, class 1 refers to neutral state while class 2 is for the emotional state. Accuracy values above 55% were written in bold. From these results, it is obtained that three classifiers which are SVM, Adaboost and Random Forest have higher performance for each dataset. Therefore, these three algorithms were selected for classification analysis.

Classification Algorithm	Dataset 1	Dataset 2	Dataset 3
Naive Bayes	53.45	52.31	54.48
Logistic Regression	51.35	51.25	**63.79**
SVM	**56.16**	**55.87**	**64.83**
K-nearest	51.35	51.60	**57.24**
Adaboost	**60.66**	**56.23**	**61.72**
Bagging	54.95	53.74	**60.34**
Decision Tree	52.55	**56.23**	53.8
Random Forest	**58.86**	**57.29**	**60.69**

APPENDIX I

RESPONSE REPORT OF DEBRIEFING FORM

At the end of our experiments, participants fill out debriefing form to express their opinions about the process. The evaluation of the participants of Experiment 2 is as follows.

1. Participants thought that the experiment was interesting, easy and fun.

2. About their performance, 6 subjects said they were moderate; 4 subjects said that their performance in the first phase was good but the second part was difficult. 5 subjects answered that their performance was bad.

3. 4 subjects thought that they counted the arrows consistently while 6 subjects thought the reverse. 2 subjects thought that they counted arrows during the first part coherently but they could not in the second phase. The rest of the participants answered this question as moderate.

4. Most of the participants (11 subjects) said that the second phase was harder than the first phase. One subject responded this question by saying that the content of the second part was the destroyed version of the images in the first part. 2 subjects felt no differences and the last subject thought that the first phase was harder.

5. About excitement, most of the participants (11 subjects) declared that they felt excitement in the second phase. However, the rest of the participants said they felt excitement at the beginning and also there was no difference regarding their emotional state between the two phases.

6. 7 subjects did not notice the image content while 3 subjects found them as neutral images and 2 subjects noticed the carpet and book images. One subject said the images were like indoor objects and another subject thought that they seemed like old and nostalgic.

APPENDIX J

THE COMPARISON OF ENTROPY FOR AVERAGED AND INDIVIDUAL DATA

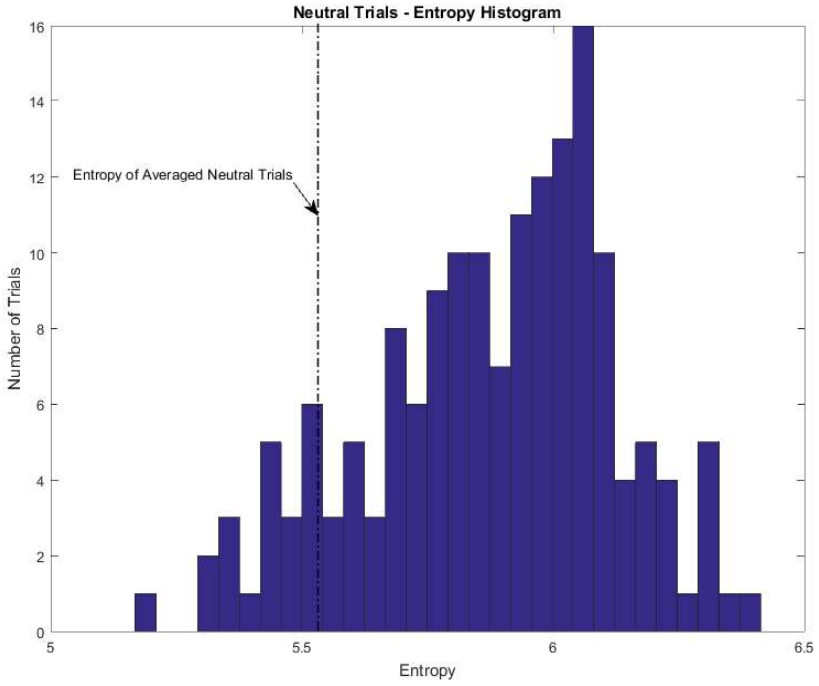


Figure J.1: Entropy Histogram of Averaged Pupil Data and Individual Trials of Dataset 1.1

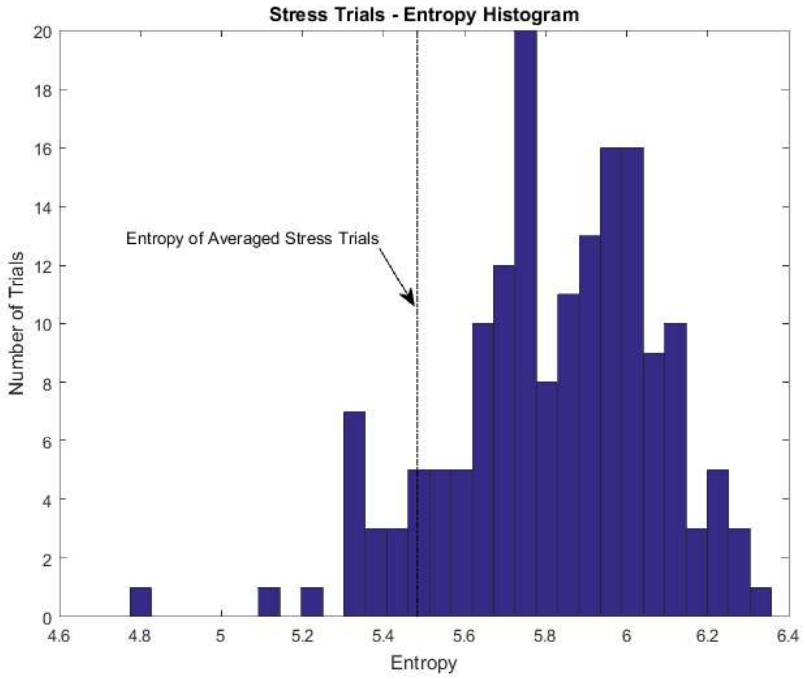


Figure J.2: Entropy Histogram of Averaged Pupil Data and Individual Trials of Dataset 1.2

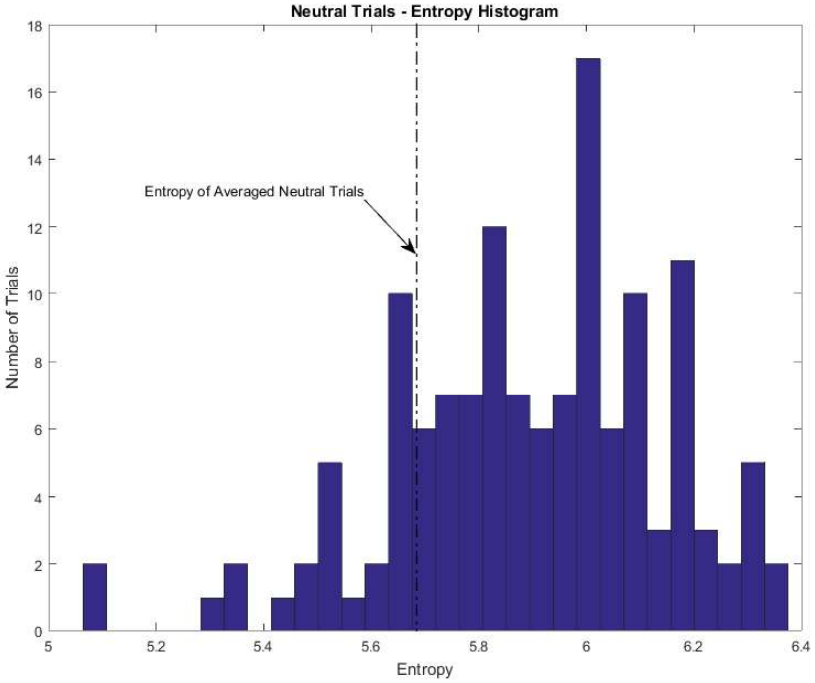


Figure J.3: Entropy Histogram of Averaged Pupil Data and Individual Trials of Dataset 2.1

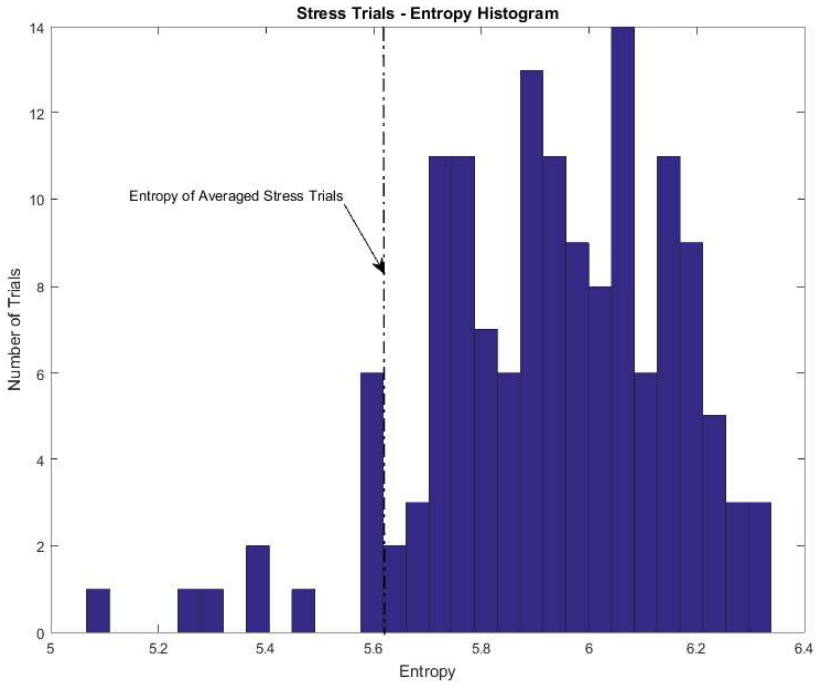


Figure J.4: Entropy Histogram of Averaged Pupil Data and Individual Trials of Dataset 2.2

APPENDIX K

CATEGORIES OF FEATURE SET

TableK.1: Types of features

Features	**Type**
min_ent	Entropy
max_ent	Entropy
mean_ent	Entropy
std_ent	Entropy
median_ent	Entropy
kurt_ent	Entropy
skew_ent	Entropy
max_abs	Absolute
mean_abs	Absolute
std_abs	Absolute
median_abs	Absolute
kurt_abs	Absolute
skew_abs	Absolute
slope_abs	Absolute
curve_1_abs	Absolute
curve_2_abs	Absolute
curve_diff_abs	Absolute
a3_coeff	Global
a2_coeff	Global
a1_coeff	Global
a0_coeff	Global
AUC	Global