

THE ROLE OF VISUAL FEATURES IN TEXT-BASED CAPTCHAS: AN FNIRS
STUDY FOR USABLE SECURITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

EMRE MÜLAZİMOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF CYBER SECURITY

JANUARY 2020

Approval of the thesis:

**THE ROLE OF VISUAL FEATURES IN TEXT-BASED CAPTCHAS: AN FNIRS STUDY
FOR USABLE SECURITY**

Submitted by EMRE MÜLAZİMOĞLU in partial fulfillment of the requirements for the degree of
Master of Science in Cyber Security Department, Middle East Technical University by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Aysu Betin Can
Head of Department, **Cyber Security**

Assoc. Prof. Dr. Cengiz Acartürk
Supervisor, **Cognitive Science Dept., METU**

Asst. Prof. Dr. Murat Perit Çakır
Co-Supervisor, **Cognitive Science Dept., METU**

Examining Committee Members:

Assoc. Prof. Dr. Aysu Betin Can
Information Systems Dept., METU

Assoc. Prof. Dr. Cengiz Acartürk
Cognitive Science Dept., METU

Asst. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU

Asst. Prof. Dr. Murat Perit Çakır
Cognitive Science Dept., METU

Asst. Prof. Dr. Hulusi Kafalıgönül
UMRAM & Aysel Sabuncu Brain Research Center,
Bilkent University

Date:

15.01.2020

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: EMRE MÜLAZİMOĞLU

Signature : _____

ABSTRACT

THE ROLE OF VISUAL FEATURES IN TEXT-BASED CAPTCHAS: AN fNIRS STUDY FOR USABLE SECURITY

MÜLAZİMOĞLU, EMRE

MSc., Department of Cyber Security

Supervisor: Assoc. Prof. Dr. Cengiz Acartürk

January 2020, 61 pages

In order to mitigate dictionary attacks or similar undesirable automated attacks to information systems, developers mostly prefer using CAPTCHA challenges as Human Interactive Proofs (HIPs) to distinguish between human users and scripts. An appropriate use of CAPTCHA requires a setup balance between robustness and usability during the design of a challenge. The previous research reveals that most of the usability studies have used accuracy and response time as measurement criteria for quantitative analysis. The present study aims at applying optical neuroimaging techniques for the analysis of CAPTCHA design. In particular, fNIRS (Functional Near Infrared Spectroscopy) is a neuroimaging technique used for mental workload analysis by means of analyzing hemodynamic responses on brain. The present study reports an experimental investigation in which 25 participants solved a group of text-based CAPTCHA with various visual characteristics.

Keywords: fNIRS (Functional Near Infrared Spectroscopy), captcha, human factors, usability, cyber security

ÖZ

METİN TABANLI CAPTCHA ARAÇLARINDA GÖRSEL ÖZELLİKLERİN ROLÜ: KULLANILABİLİRLİLİK İÇİN FNIRS ÇALIŞMASI

MÜLAZİMOĞLU, EMRE

Yüksek Lisans, Siber Güvenlik Bölümü

Tez Yöneticisi: Doç. Dr. Cengiz Acartürk

Ocak 2020, 61 sayfa

Bilgi istemlerine yapılan sözlük veya benzer istenmeyen otomatik saldırıları önlemek için geliştiriciler tarafından genellikle güvenlik kodları (CAPTCHA) kullanılmaktadır. Bu güvenlik kodları sistemi kullananın otomatik yazılım veya gerçek kişi olup olmadığını anlamaya yaramaktadır. Güvenlik kodunun uygun şekilde uygulanması için kodun zorluğu ve kullanılabilirliği arasında bir denge kurmak gerekmektedir. Yapılan önceki çalışmalar kullanılabilirliği nicel olarak ölçümlemek için güvenli kodunun doğruluğu ve tepki süresini ölçümlemişlerdir. Bu çalışma güvenlik kodunun ölçülenmesi için optik nöro görüntüleme tekniklerinin kullanılmasını amaçlamaktadır. Özellikle fNIRS (Functional Near Infrared Spectroscopy) beyindeki hemodinamik yanıtları analizi ile zihinsel iş yükünün ölçülenmesi için kullanılan bir nöro görüntüleme tekniğidir. Bu çalışma 25 katılımcının çeşitli metin tabanlı güvenli kodunun çözmesi ve ilgili katılımcıların zihinsel iş yükünün ölçülünerek raporlanmasına yönelik deneysel uygulamalar içermektedir.

Anahtar Sözcükler: fNIRS (Functional Near Infrared Spectroscopy), captcha, insan faktörü, kullanılabilirlik, siber güvenlik

To My Wife and Daughters

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Assoc. Prof. Dr. Cengiz Acartürk for his guidance and support at every stage of my thesis and my co-supervisor Asst. Prof. Dr. Murat Perit Çakır for his contributions and great support to complete this study.

Besides my supervisors, I would like to thank my wife Özgül Yeğin for her support during my studies in the period of birth and grow of our little daughters. Without her patience, I would have had a hard time completing this work. I also would like to ask for forgiveness from my girls Ece and İpek who are not fully aware that I could not spend more fun time with them.

I would like to thank to my mother Nagihan, my father İbrahim Ethem and my brother Selçuk for their support through every moment of my life.

I would like to express my gratitude to my managers for their encouragement and patience for my studies with work life.

Lastly, I would like to thank to Furkan Yücebaş, Efe Yılmaz, Gamze Eşdur and Elif Esmer for their great help during experiments.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZ.....	v
DEDICATION	vi
ACKNOWLEDGMENTS	vii
TABLE OF CONTENTS.....	viii
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS	xii
CHAPTERS	1
1. INTRODUCTION	1
1.1. The Purpose of The Study	3
1.2. Research Questions	5
1.3. The Outline of The Thesis	5
2. BACKGROUND AND LITERATURE REVIEW	7
2.1. Captchas.....	7
2.1.1. Technical background for captchas	7
2.1.2. Visual features in the text-based captchas	9
2.2. Usability for Captchas	11
2.2.1. Technical background for usability of captchas	11
2.2.2. Literature review for usability studies on text-based captchas	13
2.3. fNIRS (Functional Near Infrared Spectroscopy)	17
2.3.1. Technical background for fNIRS	17
2.3.2. Literature review for usability studies using fNIRS.....	21
3. METHODOLOGY	25
3.1. Research Design.....	25
3.2. Hypotheses.....	30

3.3.	The Procedure.....	31
3.4.	Data Collection and Materials	31
3.5.	The Participants	33
4.	RESULTS	35
4.1.	Behavioral Data (Accuracy and Response Time) Results	35
4.2.	fNIRS Data Results.....	40
4.3.	Summary of Results.....	46
5.	CONCLUSION.....	47
	REFERENCES	50
	APPENDICES	54
	APPENDIX A	54
	APPENDIX B.....	57
	APPENDIX C.....	59
	APPENDIX D	60

LIST OF TABLES

Table 1: Security Features for text-based captchas defined by Bursztein, Martin, & Mitchell (2011).....	11
Table 2: Selected features to be analyzed with difficulty levels	26
Table 3: Final security features of eighteen types of text-based captcha used in our experiment.....	27
Table 4: Descriptive Statistics for collected accuracy in percentage	36
Table 5: Descriptive Statistics for collected response time in 1000*seconds	36
Table 6: Detected significant changes in mean values (Tests of Within-Subjects Effects)	41
Table 7: Mauchly's test of Sphericity of optodes for each detected significant changes in mean values.....	42

LIST OF FIGURES

Figure 1: Examples of text-based and image-based type reCaptcha (v1, v2)	1
Figure 2: Example of an fNIRS device from Biopac Company	2
Figure 3: Original and new design of reCaptcha v1 after the experiment conducted by Bursztein and co-workers.....	4
Figure 4: Example framework for breaking text-based captchas	10
Figure 5: Regions of feasibility as a function of HIP difficulty for humans and computer algorithms.....	12
Figure 6: Absorption spectrum in NIR window	18
Figure 7: Spatial distribution of NIR Light	19
Figure 8: Illustration of three types for NIRS technique	19
Figure 9: Functional Division of the human prefrontal cortex.....	21
Figure 10: Used different web layouts during fNIRS Analysis	22
Figure 11: Example images generated for each single captcha feature	28
Figure 12: Sequence of captcha types shown in one main block design	29
Figure 13: Sequence of captcha types shown in test-run block design	29
Figure 14: Overall block design used in the experiment	29
Figure 15: Screenshots for user interface of PHP application.....	30
Figure 16: Experiment Setup.....	32
Figure 17: Marker codes and actual positions in a block design.....	32
Figure 18: The picture of experiment environment in the laboratory with the permission of one participant	33
Figure 19: Box-plot for collected accuracy in percentage	37
Figure 20: Box-plot for collected response time in 1000*seconds	37
Figure 21: Estimated marginal means for accuracy performance data.....	38
Figure 22: Estimated marginal means for response time performance data	39
Figure 23: The preparation process of collected fNIRS data	40
Figure 24: Estimated marginal means for HbO measurements.....	43
Figure 25: Optode layout on the headband and affected optodes.....	46

LIST OF ABBREVIATIONS

AMT	Amazon's Mechanical Turk
CAPTCHA	Completely Automated Public Turing Test To Tell Computers and Humans Apart
CW	Continuous Wave
DLPFC	Dorso-Lateral Pre-Frontal Cortex
DMPFC	Dorso-medial Pre-Frontal Cortex
EEG	Electro-Encephalography
FD	Frequency Domain
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared Spectroscopy
HbO	Oxygenated Hemoglobin
HbR	Deoxygenated Hemoglobin
HbT	Total Hemoglobin
HIPs	Human Interactive Proofs
HTTP	Hypertext Transfer Protocol
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
MBLL	Modified of the Lambert-Beer's Law
MEG	Magnetoencephalography
NIRS	Near-Infrared Spectroscopy
OWASP	Open Web Application Security Project
PET	Positron Emission Tomography
PFC	Prefrontal Cortex
TD	Time Domain
UI	User Interface
VLPCF	Ventro-Lateral Pre-Frontal Cortex
VMPFC	Ventro-medial Pre-Frontal Cortex

CHAPTER 1

INTRODUCTION

CAPTCHA challenges have an important role as a countermeasure of automated web application attacks. According to OWASP's Top 10 web application security risk list, brute force and dictionary attacks are classified as the second most critical attack by cyber security experts in 2017 (The OWASP Foundation, 2017, p.8). In order to mitigate dictionary attacks or similar undesirable automated attacks accomplished by scripts, developers usually prefer using CAPTCHA challenges in the form of Human Interactive Proofs shortly HIPs to distinguish human users from automated scripts. CAPTCHA is an acronym for "Completely Automated Public Turing Test To Tell Computers and Humans Apart" which was firstly introduced by researchers from Carnegie Mellon University in 2000 (Von Ahn, Blum, Hopper, & Langford, 2003, p. 298). Within a test, users are expected to solve a given challenge to prove that they are, indeed, human. These challenging tasks come in a variety of different forms. For instance, in its text-based form, it may be typing text characters that appear in a distorted image. Similarly, in its image-based form, it may consist of choosing certain requested objects (see Figure 1 for examples of two different forms of captcha).



Figure 1: Examples of text-based and image-based type reCaptcha (v1, v2)

Chew and Tygar (2004) outlined three properties that must be satisfied by the presented challenge as: "1) Easy for humans to solve; 2) Hard for scripts to solve; and 3) Easy for

tester software to generate and grade" (p.268). The first two properties address a balance between usability and robustness for captchas during the design of a challenge; it must be usable for humans and robust against automated scripts. To assess the robustness, researchers have developed and applied various attack methodologies that employ methodologies from multiple disciplines such as image processing, pattern recognition, machine learning, computer vision among others (Chen, Luo, Guo, Zhang, & Gong, 2017, p.1). For the last two decades, once a new challenge design became solvable by programming, or compromised by an automated means, researchers developed stronger security features within same type or novel type of the challenge (pp.1-2).

However, hard or new challenges make users less satisfied due to reduced usability, i.e. a higher probability of failure or low familiarity. This has led researchers to study the usability aspects of captchas. The surveys reveal the importance of visual security features including distortion, content and presentation elements such as colors, fonts, character types, lengths, text content, alongside specific captcha types (Yan & El Ahmad, 2008, pp.44-45; Beheshti & Liatsis, 2015, pp.131-132). A review of the literature shows that most of the usability studies on captchas have used accuracy and response (solving) time as measurement criteria for quantitative analysis so far.



Figure 2: Example of an fNIRS device from Biopac Company (Retrieved from <https://www.biopac.com/knowledge-base/fnir-faq/>)

More recently, optical neuroimaging techniques are used for usability and mental workload analysis in order to understand hemodynamic responses in the brain (Hill & Bohil, 2016, p.4). Recently, neuroimaging has very limited applications in cyber security research. In particular, as of our knowledge, no research has been found that focus on using neuroimaging techniques during the course of captcha solving. Functional Near Infrared Spectroscopy (fNIRS) is a relatively non-intrusive neuroimaging techniques that fits to the goal of the present research. An fNIRS device is an instrument widely used in neuroscience research since the past decade, due to its low-cost, portability, and non-

invasiveness (see Figure 2 for an example of a device from Biopac Company). In the present study, an fNIRS experiment is reported, which was carried out with 25 participants while they were solving a group of text-based captcha with different security features. The following section presents the purpose of the study in more detail.

1.1. The Purpose of The Study

In 2014, Bursztein, Aigrain, Moscicki, & Mitchell (2014) introduced a methodology that was able to solve text-based captchas automatically. It was a novel, single step approach that used machine learning algorithms. After then, application developers started to prefer using more challenging captcha types to bypass automated systems, such as image-based captchas. However, due to their complexity of deployment for implementation and the dependency of third-party cloud services for API (Application Programming Interface) implementation, text-based captcha tests are still widely accepted and used in recent ICT (Information and Communication Technologies) systems. Today, enterprise professionals usually deploy traditional challenges that accompany other countermeasures, such as rate-limiting or interaction detection, to provide secure authentication in sensitive applications. The broad use of text-based captcha has been the underlying motivation for the present study.

Bursztein, Moscicki, Fabry, Bethard, Mitchell, & Jurafsky (2014) reports an experiment conducted on Amazon's Mechanical Turk with over 27,000 respondents. They asked the respondents to solve nearly a total of a million captchas. That study has been the largest-scale usability evaluation of text-based captcha, based on behavioral measures (accuracy and response time). The main purpose of the study was to redesign Google's reCaptcha v1. The focus was not only to analyze a set of visual, anti-segmentation, and anti-recognition features in isolation (approximately 20 features including content types) but also to analyze the interaction of the features, which eventually led to a large stimuli set. In particular, they used mostly 6-8 digit, overlapping (with line for backup), up to 20-degree random rotation, length and font size randomization and sinusoidal based waving features with specific configuration. They report nearly 95% human accuracy (or recognizing the text correctly), which was 7% improvement compared the previous versions of the captcha-generation tool. The authors proposed a new design as an outcome of the experimental study. Figure 3 shows an excerpt that compared an instance of the previous design and the new design.



Figure 3: Original and new design of reCaptcha v1 after the experiment conducted by Bursztein and co-workers (Retrieved from Bursztein, Moscicki, Fabry, Bethard, Mitchell, & Jurafsky, 2014)

In the present study, we selected the major features from Bursztein et al. (2014). We used 6-digit captchas for baseline and three visual features (Line, Waving, Rotation). On the other hand, we employed a different methodology for the usability evaluation of captchas. In particular, we employed a brain imaging technique, namely fNIRS (Functional Near Infrared Spectroscopy) to measure mental workload of the participants during the course of solving captchas.

Recently, fNIRS is used for mental workload research. For instance, Pinti, Tachtsidis, Hamilton, Hirsch, Aichelburg, Gilbert, & Burgess (2018) describe the use of fNIRS device for measuring hemodynamic responses in order to quantify mental workload. Some specific cognitive tasks like mental arithmetic, reading, listening, and emotion induction activate specific areas of brain in prefrontal cortex (PFC) (p.5). When one part of the brain is activated, amount of cerebral blood flow to that area also increase, which may be used as an indicator of mental workload (p.5). This is called hemodynamic response and increase in the blood flow produces an increase in the ratio of oxygenated hemoglobin relative to deoxygenated hemoglobin in that certain area, which can be measured by the fNIRS device (p.5).

The specific purpose of the present study is to investigate whether a text-based captcha solving task produces a differentiable hemodynamic response in brain besides behavioral variables (accuracy and response time) for the selected visual features. If it is produced, secondly our next aim is to examine the effects of the selected visual features (Line, Wave, Rotation) with respect to the PFC regions. In the present study, the visual features are taken as the independent variables, whereas response time, accuracy and changes in oxy-hemoglobin and deoxy-hemoglobin are taken as the dependent variables.

We expect that the result of the current study may contribute methodologically to the field of cyber security, neuroscience, as well as usability designers in captcha designs and the design of similar challenges for authentication.

1.2. Research Questions

The present study, namely the investigation of the effects of three visual features in text-based captchas by employing behavioral measures and brain-imaging, addresses the following research questions:

Q1: Can fNIRS device be used for finding brain-imaging signatures of text-based captcha solving tasks that employ specific visual features, namely line, wave and rotation?

Q2: Which regions of the prefrontal cortex (PFC) are affected by the visual features in text-based captchas?

1.3. The Outline of The Thesis

The overall structure of the thesis takes the form of five chapters, including this introduction chapter. The following chapter (Chapter 2) contains a technical background and literature review related to our research study. Chapter 3 gives information on the design and methodology in conducted experiment. In Chapter 4, data collected from experiment is presented. Finally, Chapter 5 contains discussion of the results, limitation of the study and suggestion for future studies.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

Technical background for captcha challenges and visual features in text-based captcha challenges are presented in Section 2.1. Later technical background for usability studies on text-based captchas and literature review for usability studies on text-based captchas are given in Section 2.2. Lastly technical background for fNIRS and literature review for usability studies using fNIRS are presented in Section 2.3

2.1. Captchas

2.1.1. Technical background for captchas

The term CAPTCHA is an acronym for “Completely Automated Public Turing Test To Tell Computers and Humans Apart”. The acronym was first introduced in 2000 (Von Ahn, Blum, Hopper, & Langford, 2000), then by the same researchers from Carnegie Mellon University in 2003 (Von Ahn, Blum, Hopper, & Langford, 2003, p.298). In order to distinguish between humans and automated scripts, a captcha software generates challenging tests that humans can solve effortlessly but automated scripts cannot (p.298). Captchas are mostly deployed on websites for automatically blocking requests coming from scripts especially known as bots.

There are many practical applications of captchas used for web security. Protection against brute force attacks or dictionary attacks towards authentication mechanism, protection against online poll frauds, protection against spam attacks on blog pages or on any page allowing comments, protection against fake accounts registered by bots, protecting email addresses from bots and protection against search engine bots are given as examples for practical use of captchas (Von Ahn, Blum, Hopper, & Langford, 2000). In addition, protection against HTTP based distributed denial of service attacks was introduced as another major application of the captcha tests (Morein, et al, 2003, pp. 8-19).

Over the years, a large variety of captcha challenges has been proposed and used for web application security, as a result of a need for more robust, as well as usable captchas.

In a review article, Singh & Pal (2014) categorized captchas into five basic types based on what is used for the challenges (p. 2242):

- CAPTCHAs based on text
- CAPTCHAs based on image
- CAPTCHAs based on audio
- CAPTCHAs based on video
- CAPTCHAs based on puzzle

Text-based captchas, which emerged at time it was first introduced and became the de-facto standard style, are still the most popular one due to its easy design and implementation. In this form of captcha, users are expected to type text characters that appear in an image in a visually distorted way. The image usually consists of a randomly selected sequence of letters, numbers or it may be meaningful words. Although the aim of the captcha challenge is to discriminate between human users and automated systems, these characters can be identified automatically by optical character recognition techniques taking advantage of the advances in image processing, pattern recognition and machine learning (Chen, Luo, Guo, Zhang, & Gong, 2017, p.1). There exist certain distortion effects applied on the captcha image in order to increase resistance against this kind of techniques. These effects include adding lines randomly, adding noise on characters, etc. Another example implementation includes two word where one of them is shown in synthetically created image, other one is selected among unrecognized words scanned from books and later distorted with similar effects (see Von Ahn, Maurer, McMillen, Abraham, & Blum, 2008). In addition, some implementations include arithmetic operations to prove that they are human which needs some additional intelligence. This kind of challenges are also categorized as text-based type in the survey conducted by Singh & Pal (2014).

In the image-based type captchas, users are required to identify and choose objects or categorize them from displayed pictures or picture frames (Moradi & Keyvanpour, 2015, p.2139). This second most used type of captchas is stronger against automated, artificial intelligence (AI) attacks. As an early implementation of this type of captchas, in ASSIRA captcha, users are supposed to select only cat images from a set of 12 images of combined cats and dogs (Elson, Douceur, Howell, & Saul, 2007). However, as in text-based captchas, research showed that, image-based types also could be compromised with considerable success rates especially using deep learning techniques since it was introduced (Tang, Gao, Zhang, Liu, Zhang, & Wang, 2018, pp.2531-2532). During this period, many different techniques have been proposed, such as asking users to select from

an extended set of object categories, or from combined artificial and real objects like faces (p.2532).

Audio-based and video-based forms are another type of captchas. Audio-based captchas are essentially designed for visually impaired users and implemented either alone or with other types with optional features. In this form, after users download the audio file, they are expected to enter the spoken words. In order to resist against speech recognition techniques, some noise and sound distortion effects are added in the downloaded audio file. Shirali-Shahreza, Ganjali, Y & Balakrishnan (2011) proposed another form of audio-based type captcha which is expecting from users to speak the shown sentences by microphone. Likewise, in the video-based form, after users download the video file, they are expected to select a tag related to the topic watched in the video. However, both audio-based and video-based form of captcha have limited application as a result of high bandwidth consumption, need of hardware like a speaker or a microphone, localization problems for language and considerations on usability.

In captchas based on puzzle (or games), users are supposed to conduct harder and challenging cognitive tasks in order to prove they are human. Solving a puzzle game while sliding images, dragging and dropping objects as a response to a given task, changing the position of some objects accordingly are example implementations for this last type of captcha mentioned in the survey conducted by Singh & Pal (2014). In addition to these five captcha types, which are based on the challenge type presented to the user, there is another type which doesn't ask any hard challenge from the user but just expect them to click the checkbox. This technique was named by Google Inc. as "No Captcha reCaptcha" (Sivakorn, Polakis, & Keromytis, 2016, p.2). In this form of captchas users are classified as human or bots after a conducted risk analysis where criteria of this analysis may be based on mouse movements, directions, IP address of requester, browser cookie, etc. According to the result of this risk analysis, if there is a suspicious case, one of other captcha form is presented to the user (p.2).

In our study, we focused on only text-based form among these main five types of captcha. In the following section, more information about visual captcha features will be presented.

2.1.2. Visual features in the text-based captchas

According to a survey on breaking techniques of text-based captchas conducted by Chen, Luo, Guo, Zhang, & Gong, (2017), to resist against character recognition attacks, security of text-based captchas is mostly provided by visual features. Those visual features may include rotation, collapsing characters, background color and complexity, twisting or waving effects, addition of random dots or lines, among others (p.2). Moreover, security also depends on the character set, variance in location, length, size, font of texts, broken contours and many other visual features (p.2). Although the breaking methodology depends on the features used in captchas, there is a common approach to solve captchas automatically which consists of a sequential process. Bursztein, Aigrain, Moscicki, & Mitchell (2014) summarized commonly applied standard approaches for solving a text-

based captcha in sequential six steps as pre-processing, segmentation, post-segmentation, recognition and finally post-processing operations (see Figure 4).

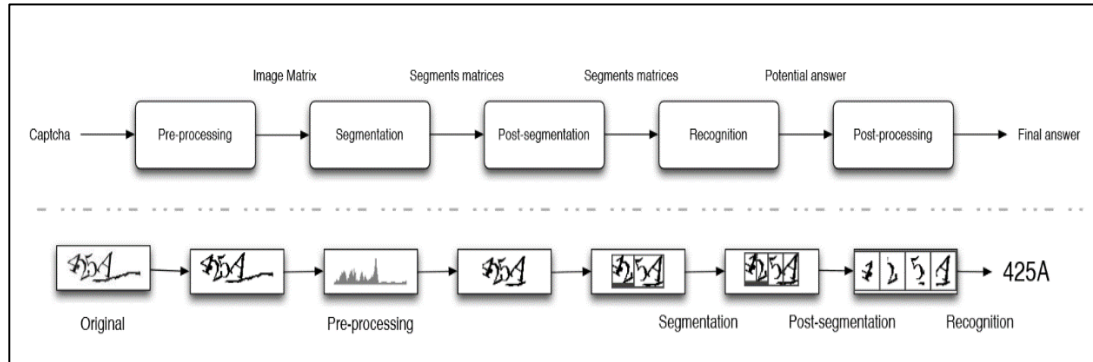


Figure 4: Example framework for breaking text-based captchas (Retrieved from Bursztein, Aigrain, Moscicki, & Mitchell, 2014, pp.2-3)

In the pre-processing step, several operations are performed before segmentation including image binarization, image thinning, and denoising. Image binarization operation highlights focused characters and cleans background using local or global threshold filters; image thinning operation highlights contour as skeleton using thinning algorithms such as Hilditch and Zhang & Suen algorithm; finally, image denoising operation removes noises and lines that intersects characters (Chen, Luo, Guo, Zhang, & Gong, 2017, p.5). In the segmentation step, in order to be more easily recognized in the following steps, characters are separated using various techniques by checking feature, width, contour or projection characteristics of characters (p.5-7). After segmentation completed successfully, minor position corrections are performed at the post-segmentation step. In the recognition step, characters are actually recognized using pattern recognition based on features or structures on characters or machine learning algorithms. Lastly, according to captcha features such as character set and length, recognized characters are rejected or accepted at the post-processing step (p.11).

Bursztein, Martin, & Mitchell (2011) defined ten security features after analyzing real world captcha implementations used by common web services, such as Google, Reddit, Wikipedia and eBay (p.125). These features are grouped into two groups based on security defense techniques; anti-recognition and anti-segmentation (p.125). These features and their actual objective are listed in the Table 1.

Table 1: Security Features for text-based captchas defined by Bursztein, Martin, & Mitchell (2011)

Anti-recognition		Anti-segmentation	
Multi-fonts	Using multiple fonts or font-faces.	Complex background	Try to hide the text in a complex background to "confuse" the solver
Charset	Which charset the scheme uses		
Font Size	Using variable font size	Lines	Add extra lines to prevent the solver from knowing what are the real character segments
Distortion	Distorting the captcha globally using attractor fields		
Blurring	Blurring letters		
Tilting	Rotating characters with various angles	Collapsing	Remove the space between characters to prevent segmentation
Waving	Rotating the characters in a wave fashion.		

Among these features tilting (rotation), lines and waving are selected as independent variables in our experiment design.

A challenge in the domain of captcha research is the contradiction between security and usability. On the one hand, captchas have to be difficult to recognize to strong against automated attacks. On the other hand, too much difficult in character recognition may result in non-readability by humans. The following section presents usability aspects in captcha design and use.

2.2. Usability for Captchas

2.2.1. Technical background for usability of captchas

The standard ISO/IEC 25010 defines the usability and security characteristics, that information systems must have:

Usability: Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Security: Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.

Gordieiev, Kharchenko, & Vereshchak (2017) emphasized the importance of the balance between usability and security characteristics during the design of captcha challenges and their difficulty to implement (p.728). Moreover, they provided example captcha tests which can be categorized as a good example of an unusable security function. In these examples, users have problems during captcha solution due to reduced readability, which in turn makes the users uncomfortable (p.729). In order to resolve this problem, users mostly reload the captcha image to capture a readable one or cancel the process which they wanted to perform (p.729).

Chew and Tygar (2004) provided three properties that must be satisfied by the presented challenge as: “1) Easy for humans to solve; 2) Hard for scripts to solve; and 3) Easy for tester software to generate and grade” (p. 268). The first two properties imply that there should be a balance between usability and robustness for captchas during the design of a captcha challenge; it must be usable for humans and robust against automated scripts. The development and implementation of challenges providing these three captcha properties is difficult. Chellapilla, Larson, Simard, Czerwinski, & Czerwinski (2005) figured an ideal distribution of HIPs (Human Interactive Proofs) (p.712). A sweet spot range of captcha challenges were postulated in their study. This range is presented in the Figure 5 where its size can be changed over time. In practice, automatic scripts developed by todays computer should not be more successful than 1 in 10,000 (0.01%) and the human success rate should approach at least 90% (p.712).

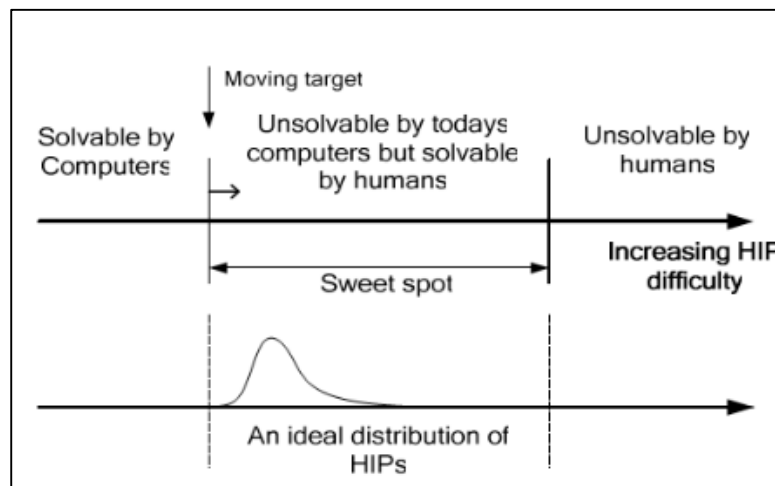


Figure 5: Regions of feasibility as a function of HIP difficulty for humans and computer algorithms
(Retrieved from Chellapilla, Larson, Simard, Czerwinski, & Czerwinski, 2005, p.712)

Nielsen (2003) defined usability of systems/applications with 5 quality components as quoted below (p.1):

Learnability: How easy is it for users to accomplish basic tasks the first time they encounter the design?

Efficiency: Once users have learned the design, how quickly can they perform tasks?

Memorability: When users return to the design after a period of not using it, how easily can they reestablish proficiency?

Errors: How many errors do users make, how severe are these errors, and how easily can they recover from the errors?

Satisfaction: How pleasant is it to use the design?

Our literature review showed that usability studies on captchas mostly address three of these components which are efficiency, errors and satisfaction. Efficiency and error components are categorized as quantitative analysis and can be measured with two criteria for a captcha test; 1) **Accuracy:** how accurately users can pass a captcha challenge, 2) **Response time:** time to solve a captcha whether in correct way or not. On the other hand, satisfaction component of usability is categorized as qualitative analysis and performed with **questionnaires** in surveys mostly in order to understand difficulty perception of users, satisfaction of using a design and lastly mental workload after solving challenges. In the following section literature review about usability studies on text-based captchas will be presented.

2.2.2. Literature review for usability studies on text-based captchas

Baird & Riopka (2005) studied a specific type of text-based captcha called the Scattertype. In the scattertype captchas, within each image, characters are fragmented using horizontal and vertical cuts, and the fragments are scattered by vertical and horizontal displacements. They synthesized this type of captcha that uses variable length pseudo English words. They conducted their experiment in the lab environment with 57 participants asking to solve totally 4,275 captcha challenges with 5 different levels of security and numerous font families. Each captcha had a different difficulty level with different cutting fraction, expansion fraction, horizontal scatter, vertical scatter and character separation levels. For each difficulty level, they measured solving accuracy and for overall tests they produced most confused character matrix.

Chellapilla, Larson, Simard, Czerwinski, & Czerwinski (2005) aimed to better understand human segmentation and recognition capabilities and later to use the results for usable designs of text-based captchas. Their studies were conducted online, asking participants to solve captcha challenges from their own offices. 76 users participated in the first set of experiments. The first set included translation, scaling, rotation and global & local warping effects on captchas. Warping effects are similar to the waving feature that we selected in our experiment. For the second study, 29 users participated, and the stimuli included set of effects such as thin and thick lines that intersected and the ones that did not intersect. The first study showed that for global warping, local warping and local warping plus a combination of other distorting parameters, there was a significant decrease in accuracy. The second study showed that thin lines or non-intersecting thick lines could be used in captcha challenges without affecting usability. For future work, they recommended that each combination of security features could be examined and in addition to accuracy, solving time should be taken in consideration.

Bursztein, Bethard, Fabry, Mitchell, & Jurafsky (2010) conducted a large-scale evaluation of captchas by focusing on the human perspective. They collected and analyzed 21 most popular and widely used captcha schemes (13 images schemes and 8 audio scheme) with the workers from Amazon's Mechanical Turk and with an underground captcha-breaking service to solve more than totally 318,000 captchas. With this study, they were able to compare audio-based and text-based captchas of the same type. They were also able to compare AMT (Amazon's Mechanical Turk) and Underground service workers. For conclusion, they reported that audio-based captchas were harder than text-based captchas in terms of accuracy and solving time. With collecting demographics from AMT, they resulted that non-native speakers of English were slower, though they were generally just as accurate as the others unless the captcha required recognition of an English word. Finally, they reported minor trends indicating that older users were slower but more accurate. For future work, they recommended to investigate more deeply how individual differences influence captcha difficulty (pp.399-413).

Bursztein, Moscicki, Fabry, Bethard, Mitchell, & Jurafsky (2014) aimed to redesign Google's reCaptcha v1 which is a text-based captcha tool. Before their new design, they performed and observed the output of their two studies. One of them was an experiment that measured response time and accuracy, and the other one was a survey. They conducted their first study by presenting generated captchas to users and measuring how different features influenced solving time and accuracy. They examined the effects of each feature in isolation and analyzed the interaction between features by varying multiple features at the same time. They conducted their evaluation on over 27,000 respondents from Amazon's Mechanical Turk (AMT), with millions of captchas. With the power of AMT, they have examined many visual feature effects of captchas such as different character sets, different number of characters, different character overlapping features, variable font size and font family, different foreground and background colors, different number of random dots, different number of lines, line sizes and line type (wavy, straight), up to 20-degree rotated characters, different range of vertically shifted characters. They have also made a survey for user preferences about word and visual feature preferences for qualitative analysis. They achieved nearly 95% human accuracy, indicating 7% improvement compared to the previous designs.

Hsu, & Lee, (2011) conducted an experiment to study the effect of age groups and visual features on the usability of text-based captchas. Twenty-four participants were recruited to take part in the experiment, where twelve of them were in the senior group and twelve in the young group. The types of visual feature examined are Blot Mask, Line Mask, Thread Noise, Global Warp, Geometry Noise. Error rate and response time were measured for each security feature and age groups and mental workload calculated with questionnaires using NASA-TLX scoring. Results of their study showed that different age groups have significantly different response time, error rate, mental workload. For future work they recommended that the same study may be performed to study different age groups and character sets on the usability of text-based captchas.

Tangmanee (2018a) noted that most of security papers focus on technical design and robustness of captcha challenges. On the other hand, usability of captcha challenges has been relatively ignored. Accordingly, letter perception become critical for usable design of text-based captchas. According to some features of lowercase letters, researcher collected and examined correct response rates (accuracy) of them for 6,389 iterations of lowercase letters used in 1,844 captcha challenges. These features are short or not, ascending or not, descending or not. The study showed that the short feature affects the correct response rate of captchas, but extending features of lowercase letters (ascending or descending) does not affect the letter's usability. The study offers 12 characters with short features that should be used in captchas, namely " where they are a,c,e,m,n,o,s,u,v,w,x,z.". Similarly, some popular web site captchas like Yahoo, Baidu, Google's reCaptcha use a small set of letters based on an unknown rule. His future work recommendation is to examine what a set of letters are both usable for humans and robust from bot attack together in terms of features examined in this study.

Tangmanee (2018b) investigated usability of captchas according to their letter case with a set of selected features. He used the same data set that as before. At this time the features he examined were uppercase or lowercase, uppercase with straight line or not, uppercase with diagonal or not, lowercase with straight line or not, lowercase with slender or not. The variables were correct response rates (accuracy) of users for each 5 features using 8,928 iterations for which both uppercase and lowercase letters used in 1,844 captcha challenges. This study showed that the uppercase letters used in captchas are less usable than lowercase letters. Another result is the uppercase letters with the straight-line feature is more usable than uppercase letters without the straight-line feature. The Uppercase Diagonal feature does not significantly change the result. Similarly, for lowercase, the feature straight-line is more usable than without it, but the slender feature of lowercases does not significantly change the result. His future work recommendation is to examine the set of letters that are both usable for humans and robust against automated attacks.

Brodić, Amelio, & Janković, (2018) analyzed the effect of captcha characteristics on users' response times. In this study, an experiment was reported, which was based on two populations of users for text and image-based captchas, separated by demographic features as their age, gender, education level and internet experience. Their study had a distinct difference than the previous studies; they conducted their experiment with a larger population (102 internet users) with higher number of demographic factors, and their study presented a more general analysis including a population of students, employees, clerks, teachers and engineers of age between 18 and 52 years and also it explored nine captcha types, including different types of text and image-based captchas. Two text-type captchas were text only and numeric only, seven types of image-based captchas to click animals in the wild, home numbers, face of an old woman, animated face, worried face, surprised face, and the picture of the captchas among a picture set. They examined response time results statistically to prove or disprove a set of hypotheses. Their analyses showed that educational level significantly affects response time but gender does not. Lastly, the results showed that younger users solve the challenges more quickly compared to elderly.

Belk, Fidas, Germanakos, & Samaras (2015) reported two experiments that investigated the effect of users' cognitive styles and cognitive processing abilities (speed of processing, controlled attention, working memory capacity) on preference and response time. Before the experiment, they categorized participants according to their cognitive styles using their web-based psychometric test instruments developed for this study. Later, they conducted first part of their experiment with 131 participants in order to understand their preference between text-based and image-based captchas and measured their accuracy and response time. By their second experiment with 125 participants, they explored at this time, the effect of users' cognitive processing abilities on accuracy and response time for different levels of difficulty. For both text-based and image-based captcha challenges, these levels determined the added noise and distortion ratios on the characters or images.

Nanglae, & Bhattarakosol (2015) reported a questionnaire that included samples from three countries and found that the nationality of users also has impact in using text-based captcha. They collected their data from 188 random participants from ages between 10-60 years old, 70 users are Bhutanese, 48 users are from Indian, and 70 users are from Thai. The questions are grouped into four categories which are text-based captcha appreciation, text-based captcha Familiarity, text-based captcha recommendation, text-based captcha usability. Similarly, Norbu & Bhattarakosol (2012) had done previously the same study with participants from Bhutan and Indian only.

Gafni, & Nagar (2016) focused on the users' attitudes and experiences related to use of the different types of captchas. A questionnaire and an experiment were conducted with 212 participants. The questionnaire with 61 questions was made in order to understand user experiences and an experiment was conducted in order to measure accuracy and response time for five different type of captcha challenges. These captcha types are text-based, arithmetic operation based, image based, game based and "No CAPTCHA". Researchers recommended for future research, to examine the effects of different devices in solving captcha tests, such as through mobile phones or computers in terms of user experience, accuracy, and response time.

Beheshti, & Liatsis (2015) examined the captcha usability issues such as length of characters, size of characters, language used in context and distortion levels for security measurement with 100 participants. In order to conduct their study, a website was developed which used the reCaptcha with a specific type of text-based captcha and it allowed for user feedback through an online survey form with 13 questions. They collected also accuracy rates and response time for different backgrounds, ethnic and age participated in the experiment. Together with quantitative and qualitative data, they have compared the output results. Novelty of the present study is to employ a brain imaging technique to study the role of captcha features on mental workload, besides behavioral results. The following section presents technical background for fNIRS and studies using this technique.

2.3. fNIRS (Functional Near Infrared Spectroscopy)

2.3.1. Technical background for fNIRS

Near Infrared Spectroscopy (NIRS), first described by Jobsis in 1977, is a neuroimaging technology used for measuring absolute levels or changes in concentration levels in oxygenated (HbO) and deoxygenated hemoglobin (HbR) in the brain. Total hemoglobin (HbT) is calculated with the summation of HbO and HbR. Basically, three types of NIRS technique are commonly used based on the specific type of illumination: “1) CW-Continuous Wave 2) FD-Frequency Domain 3) TD-Time Domain” (Ferrari & Quaresima, 2012, p.924). The most commonly used technique and commercially most available one is CW-based NIRS due to its technological simplicity, low-cost and portability (p.924). However, this type of technique measures only changes (not absolute values) of oxygenation levels in hemoglobin. Conversely, FD-based and TD-based NIRS techniques can measure the absolute oxygenated and deoxygenated hemoglobin levels in the blood (p.924). The instrumentation for fNIRS used in the present study is not FD-based or TD-based, so more details about the principles of CW-based NIRS technique is presented in this section.

Spectroscopy is the study of the absorption and emission of light signals. Principles of Near Infrared Spectroscopy (NIRS) technology are based on the facts presented below:

- The human tissues are relatively transparent to near-infrared (NIR) light in the spectral window 650-1000 nm (Ferrari & Quaresima, 2012, p.923).
- Near-infrared (NIR) light is either absorbed or scattered in the human tissues (Ferrari & Quaresima, 2012, p.923).
 - It is able to penetrate human tissues. (Scattering)
 - It is attenuated mainly due to hemoglobin in the human tissue. (Absorption)
- Absorption of oxygenated and deoxygenated hemoglobin differs depending on the Near-infrared (NIR) light's wavelength (Jobsis, 1977)

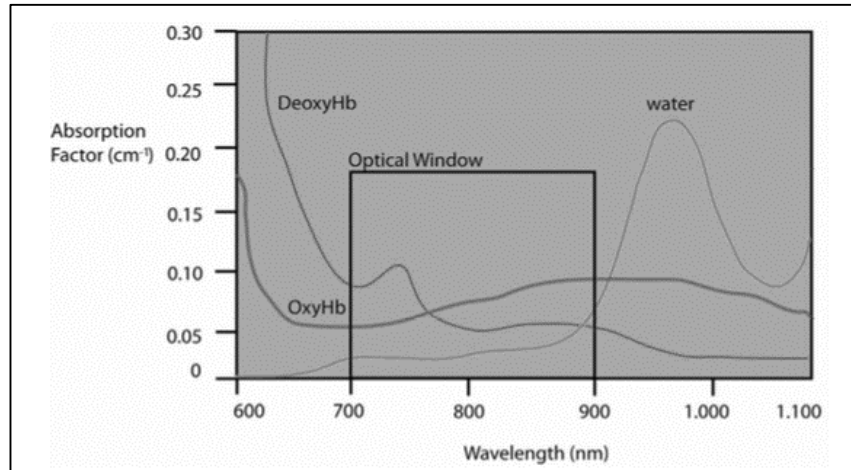


Figure 6: Absorption spectrum in NIR window (Retrieved from Leon-Carrion & Leon-Dominguez, 2012)

With taking advantages of these facts, NIRS technology is able to measure changes in HbO and HbR levels by transmitting NIR light onto the scalp with two or more wavelengths. These levels are calculated by using a Modified of the Lambert-Beer's Law (MBLL) based on the reflected (scattered) light. Details of this MBLL calculation are explained in this section. While HbO absorption factor is higher for $\lambda > 800$ nm, HbR absorption is higher for $\lambda < 800$ nm (Pinti, Tachtsidis, Hamilton, Hirsch, Aichelburg, Gilbert, & Burgess, 2018, p.3). But the wavelength ranges that between 700-900 nm (see Figure 6) is the optimal and non-invasive light spectrum for cognitive studies, named the optical window (Leon-Carrion & Leon-Dominguez, 2012, pp.49-50). So, there should be at least two components for measurement of reflected light: 1) light source for transmitting NIR light and 2) light detector for quantifying scattered NIR light where spatial distribution is like banana-shape that is shown in the Figure 7 (Ferrari, Mottola, & Quaresima, 2004, p.465). The distance between the source and the detector effects penetration depth of the light where the longer distance, the deeper it penetrates (Pinti, Tachtsidis, Hamilton, Hirsch, Aichelburg, Gilbert, & Burgess, 2018, p.4). Typical source and detector distance values are 4 to 5 cm for adult studies and 2 to 3 cm for child and baby studies (Quaresima & Ferrari, 2019, p.51).

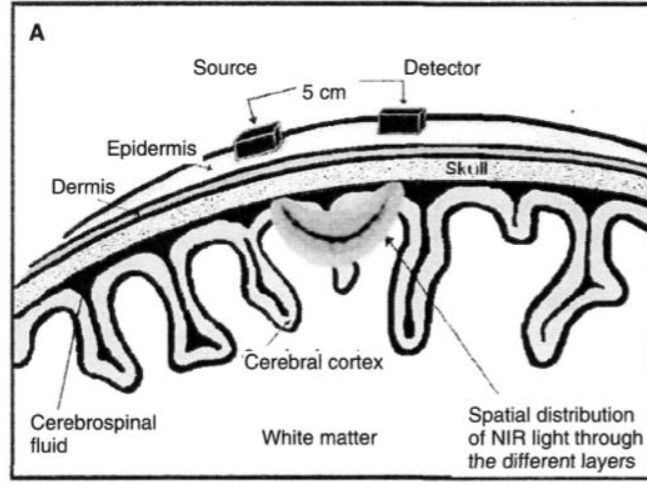


Figure 7: Spatial distribution of NIR Light (Retrieved from Ferrari, Mottola, & Quaresima, 2004)

NIRS light source and detector pair is called an "optode", and the region of tissue interrogated by the NIR light is called a "channel". It is located in the middle of the source and the detector with a depth of around the half of the source-detector pair distance. Commonly, optodes are distributed uniquely on the head with fixed distances. As it was mentioned at the beginning of this section, briefly there are three types of technique (CW-based, FD-based, TD-based) that are commonly used for NIRS. Let's assume, I_0 is the transmitted light signal and I is the reflected signal that will be attenuated differently related to changes in hemoglobin concentration. In a different technique, the signal shapes and quantification method differs as illustrated in the Figure 8 (Scholkmann, Kleiser, Metz, Zimmermann, Pavia, & Wolf, 2014, p.7).

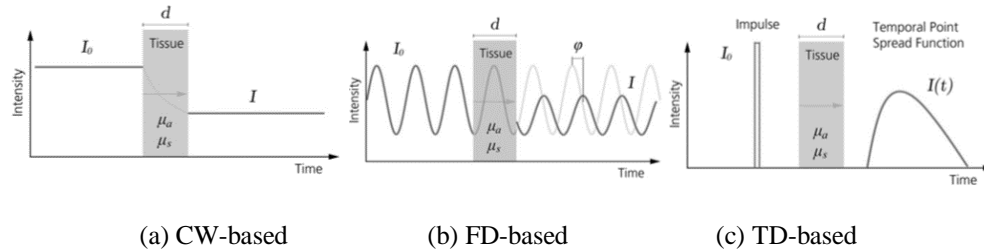


Figure 8: Illustration of three types for NIRS technique (Retrieved from Scholkmann, Kleiser, Metz, Zimmermann, Pavia, & Wolf, 2014)

For the CW-based NIRS technique, based on the transmitted and reflected light signal, HbO and HbR levels are calculated by using a Modified of the Lambert-Beer's Law (MBLL) as shown below (Pinti, Tachtsidis, Hamilton, Hirsch, Aichelburg, Gilbert, & Burgess, 2018, p.4):

$$\text{Absolute Attenuation } (A) = -\log_{10} \left(\frac{I_0}{I} \right) \quad (1.1)$$

Initially, the first absolute attenuation measurement is calculated and later measurements are subtracted by this value in order to remove the effect of scattering, melanin, and water concentrations which makes it as differential spectroscopy (p.4).

$$\text{Differential Attenuation } (\Delta A) = \varepsilon(\lambda) \cdot \Delta c \cdot d \cdot DPF(\lambda) \quad (1.2)$$

Where ε : is the molar absorption coefficient at a certain wavelength λ

Δc : is the changes in chromophore concentrations (either HbO or HbR)

d : is the distance of the source and detector

DPF : is the differential pathlength factor

In the present study, the MBLL function was used, which was defined in the fNIRS software (Biopac, United States), to calculate oxygenation levels from collected fNIRS spectroscopy signals. The details of how this function was applied will be mentioned in the methodology chapter. Functional neuroimaging uses the NIRS technology to study the relationship between brain activity in certain brain areas and specific mental functions.

The term fNIRS refers to functional neuroimaging (Functional Near Infrared Spectroscopy). fNIRS has been employed in domains of research that are also addressed by functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG) and positron emission tomography (PET). In 2018, Pinti and his colleagues summarized advantages and disadvantages for fNIRS over fMRI, EEG, MEG technologies.

The previous studies have shown that certain functional or cognitive tasks like mental arithmetic, music imagery, emotion induction, etc. activate certain areas of the brain in the prefrontal cortex (PFC) to support oxygen demands, where HbO increases and HbR decreases (Pinti, Tachtsidis, Hamilton, Hirsch, Aichelburg, Gilbert, & Burgess, 2018, p.5). This is called the "hemodynamic response" and it can be measured by fNIRS or similar technologies (p.5). The main brain areas that are activated during cognitive tasks in the prefrontal cortex is illustrated in the Figure 9 (Carlén, 2017, p.3).

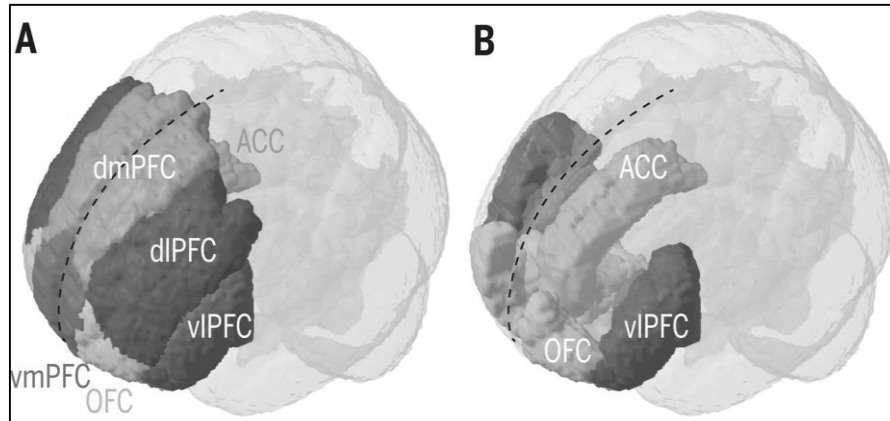


Figure 9: Functional Division of the human prefrontal cortex (Retrieved from Carlén, 2017)

Studies using fNIRS technology recently cover various domains including social cognition, psychology and education, economics, linguistics and medicine both in healthy or neurological and psychiatric disorders besides functional neuroimaging (Quaresima & Ferrari, 2019, pp.59-60). Our study is example of functional neuroimaging where we measure hemodynamic correlates of mental activity during the task completion related with cyber systems authentication, in order to understand mental workload.

2.3.2. Literature review for usability studies using fNIRS

In addition to behavioral data such as response time, accuracy measurements, and subjective measurements of mental workload using qualitative questionnaires, usability testing may be conducted using neuroimaging tools and techniques for quantitative analysis (Hill & Bohil, 2016, p.1). Functional near-infrared spectroscopy (fNIRS) is a neuroimaging tools that can be used to measure neural activity during cognitive tasks (p.1). Due to its non-invasive and portability properties, usability testers can conduct their experiments and collect data in operational environments using fNIRS tools with their flexible caps (p.2).

Hirshfield, Solovey, Girouard, Kebinger, Jacob, Sassaroli, & Fantini, (2009) presented a novel process for experiments that could be used to measure the mental workload during participants performing some cognitive tasks. A web page was designed, which included a set of cognitive tasks that demanded the use of verbal or spatial working memory. They conducted their experiment with 10 participants. Their measurements were based on oxygenized hemoglobin changes (HbO) where data were collected by a frequency-domain based fNIRS tool. Their results indicated that different levels of workload cognitive tasks could be detected by fNIRS measurements, which in turn could help improving the User Interface (UI) design processes.

Lukanov, Maior, & Wilson, (2016) investigated usability issues on the design of three different form layouts for a car insurance form filling process using continuous-wave based fNIRS tools. 15 right-handed persons participated in the experiment. The layouts

used in the experiment are shown in Figure 10. Before the experiment, a video was shown to participants which contained information that they needed to fill in the forms. In addition to the collected fNIRS data, for subjective measurement, participants filled in multidimensional subjective workload scale NASA-TLX battery that included Mental demand, Physical demand, Temporal demand, Performance, Effort, and Frustration. Their measurements were based on oxygenized hemoglobin changes (HbO). The results showed that the divided form layout (see Figure 10) needs more mental workload, in contradiction of expectations concluded from the questionnaires.

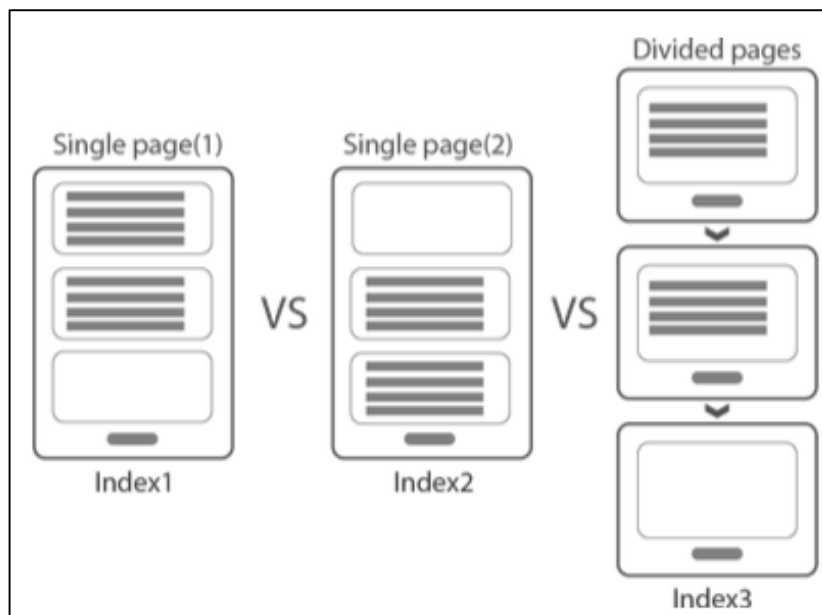


Figure 10: Used different web layouts during fNIRS Analysis (Retrieved from Lukanov, Maior, & Wilson, 2016)

Bhatt, Agrali, McCarthy, Suri, & Ayaz (2019) investigated usability of a new web-based tool using fNIRS, eye-tracking tools, survey measures, and performance measures. They conducted their experiment with 37 participants, which consisted of three groups (first time users, current users and employees). There were two levels of task difficulty for all task types (37 different tasks) grouped as *easy* and *difficult*. As expected for the difficult-level tasks, it was measured higher oxygenated hemoglobin changes.

Neupane, Saxena, & Hirshfield (2017) studied users' neural signals during processing of legitimate vs illegitimate and familiar vs unfamiliar websites using fNIRS. This study an example for an fNIRS study used in the cybersecurity context. They selected 30 websites from Alexa Top 50 list and used the for studying phishing in the experiments. They conducted their experiment with 20 participants. They used statistical Friedman's test and Wilcoxon Singed-Rank Test (WSRT) in order to determine existence of significant changes related with oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) on PFC. Their analysis showed that participants experienced higher cognitive workload during assessment of validity of the visited web sites.

Our review of the literature shows that researchers who want to conduct experiments in human computer interaction laboratory environment should take care of some artifacts for proper fNIRS studies. Solovey, Girouard, Chauncey, Hirshfield, Sassaroli, Zheng, Fantini, & Jacob (2009) identified these artifacts and gave guidelines for researchers how they should be aware of several considerations. Head movement, facial movement, ambient light, ambient noise, respiration and heartbeat, muscle movement and lastly slow hemodynamic response were identified as potential sources of noise and artifacts. They also examined keyboard input, mouse input, head movement and facial movement in their experiments. In their study, they didn't examine ambient light, noise, respiration and heartbeat, slow hemodynamic response due to many correction methods had been previously proposed and got successful result earlier but they gave guidelines for them. They noted that forehead and major head movements should be avoided and minor head movements, heartbeat and respiration may be corrected using filtering for experimental studies. These filtering methods may be adaptive finite impulse response (FIR) filtering, Weiner filtering, adaptive filtering. The fNIRSoft software that is used in our study has the FIR filtering capability for cleansing the collected data, which was also employed in the present study. From the previous studies and the results of their own experiment, they gave some guidelines for the other artifacts: for ambient light wear isolating cap should be used, minimizing ambient external noise, collecting signal during a clicking only task for mouse click artifacts, designing test interfaces with in a time span 6-8 seconds in order to avoid slow hemodynamic responses are useful outputs of their study. All these considerations are summarized in their paper. In our study, we designed our tests according to their guideline.

Anderson, Kirwan, Jenkins, Eargle, Howard, & Vance (2015) defined neurosecurity as “applying neuroscience to behavioral information security to better understand and improve users’ security behaviors. One ultimate goal of neurosecurity is to design more effective user interfaces (UIs) that can help users make informed decisions” (p.2884). They published research related to neurosecurity using various neuroimaging technologies such as fMRI, EEG, Eye Tracking, and fNIRS. Mostly their studies were related to information security risk perception or behaviors of users against warnings such as legitimate web sites. Our study can be categorized as neurosecurity experiment, as we investigate human brain responses using fNIRS tools during solving captcha challenge which is important cybersecurity countermeasure in authentication.

CHAPTER 3

METHODOLOGY

The research design of our study, the materials used for data collection, hypotheses to find the answers of our research questions, the procedure which includes conducted steps before and during the experiment, and lastly participant's information are presented in this chapter.

3.1. Research Design

The focus of our study is to determine if fNIRS device can be used to analyze hemodynamic response during a text-based captcha solving task in order to understand usability of them. If any significant response can be detected during captcha solving task, another focus of our study is to find which regions of the prefrontal cortex (PFC) are significantly affected by the selected security features used in captchas.

Numerous visual features are used in text-based captcha designs, such as the addition of visual clutter as noise, collapsing characters and so on. Following-up our literature review we have selected three of them to be analyzed during our experiment. They were the main features of the final design of Google's reCaptcha v1. These features were determined by an experiment conducted with massive number of participants from Amazon's Mechanical Turk (see Chapter 2 for the literature review). We have designed these three visual features with different levels of difficulty. First levels of all the features are the levels with no features applied. The second levels of all the features are the level of difficulty, when these levels of features are applied. At those levels, the expectation was a minor effect on the usability of captchas. The third levels of the features are the most difficult levels, where usability was expected to be affected negatively.

The first selected feature is the *line* feature with 3 different levels. For the line-feature effect, we have added a line in the middle region of a captcha passing through all the characters. The second level of the line-feature is a thin line and the third level of line-feature is a thick one. The second feature is the *waving* feature with 3 levels. For the waving-feature effect, we have relocated the orientation of captchas on a sinusoidal line at the second level of difficulty. On the third level of the waving-feature, in addition to the second level, more waving effect are applied on each captcha character. The last feature is the *rotating* feature with 2 levels. At the second level of difficulty, we have rotated each character independently with the angle randomly selected from an array including -20, 15, 15, 20 degrees. This feature has no more levels, because rotated characters with more degrees start overlapping by decreasing the accuracy dramatically. All three visual features are summarized in Table 2 with their difficulty levels.

Table 2: Selected features to be analyzed with difficulty levels

Features	# of Difficulty	Security features defined for each level		
Line	3 Levels	No Line	Line Thin	Line Thick
Waving	3 Levels	No Wave	Wave Light	Wave High
Rotation	2 Levels	No Rotation	Rotation (-20,-15,15,20)	-

In order to analyze the effect of each feature and the difficulty level, all the combination of these different features are analyzed. A total of eighteen types of text-based captcha are generated where features of each type are given in Table 3. The captcha with no security feature was chosen as baseline captcha.

Table 3: Final security features of eighteen types of text-based captcha used in our experiment

Type#	Line Features	Waving Features	Rotation Features
1	No Line	No Wave	No Rotation
2	Line Thin		
3	Line Thick		
4			Max Rotation 20
5		Wave Light	
6		Wave High	
7	Line Thin		Max Rotation 20
8	Line Thick		Max Rotation 20
9	Line Thin	Wave Light	
10	Line Thin	Wave High	
11	Line Thick	Wave Light	
12	Line Thick	Wave High	
13		Wave Light	Max Rotation 20
14		Wave High	Max Rotation 20
15	Line Thin	Wave Light	Max Rotation 20
16	Line Thin	Wave High	Max Rotation 20
17	Line Thick	Wave Light	Max Rotation 20
18	Line Thick	Wave High	Max Rotation 20

* Max Rotation 20 = Rotation (-20, -15, 15, 20)

Following up the literature review; font family, font size, color of characters and background color are selected from the set of visual features that affect the usability

minimally in order to focus only the visual features analyzed in our study. We determined to use a black foreground, a gray background and selected Arial regular true-type font family with 22 font size to generate each character of the captchas. With the same objective, it is selected only 6-digit numeric character set that was randomly generated for each participant.

In order to generate captchas in our stimuli, an open source tool named cool-php-captcha¹ was used with some little code adjustments to match exact visual features expected in our design. The details of these adjustments are given in Appendix B. One sample of generated captcha images for each single feature is given in Figure 11 and three examples for all eighteen types of captcha generated by our application are also listed on Appendix A.



Figure 11: Example images generated for each single captcha feature

The literature reviews also showed that solving a text-based captcha with usable accuracy takes 3-7 seconds. In order to activate hemodynamic response in the brain, we need to give a task that lasts at least 8 seconds or more. In order to deliver and guarantee this objective, the same type of captcha was repeated five times in a block design shown in Figure 12. Between each of captcha in the block, two seconds little rest is designed to be given.

¹ <https://github.com/josecl/cool-php-captcha>

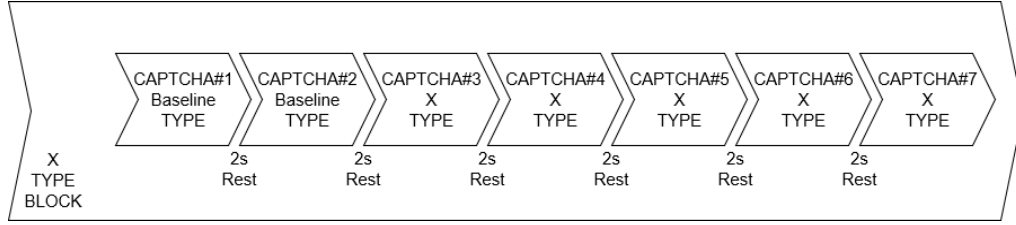


Figure 12: Sequence of captcha types shown in one main block design

For each captcha type with the features as specified in Table 3, a total of 18 blocks were designed and displayed to each participant in a random order. Between each block, 10-second rests are placed. At the beginning of the blocks, a test-run block was displayed, which aimed to make the participants comfortable with the user interface, physical conditions and materials used in the laboratory. Within the test-run block, types of captchas were randomly selected as shown in the Figure 13.

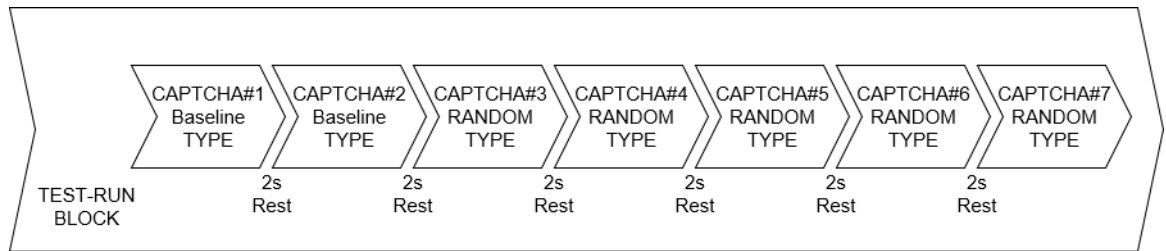


Figure 13: Sequence of captcha types shown in test-run block design

After the test-run and before the experiment session, a 20-second rest was introduced. So a total of 19 blocks are displayed during the experiment. The overall block design is depicted in Figure 14.

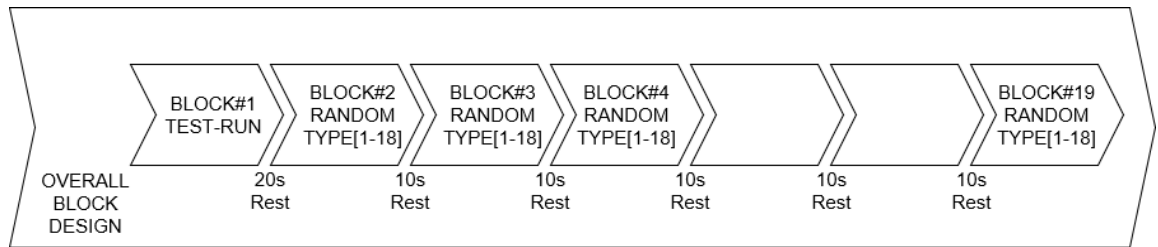


Figure 14: Overall block design used in the experiment

In order to display the captchas a web application was developed using PHP programming language. The application had its own user interface and collected all the data needed for response time and accuracy measurements, as well as demographic data of the participants. In addition to correct and answered values, the displayed captcha images have been also recorded in base64 form in order to use in troubleshooting if needed. The application communicated with the fNIRS device by sending a marker for

synchronization. These markers define a start of a block, end of baseline captchas and lastly end of a block which will be helpful for data analysis for collected fNIRS measurements.

The application included a simple graphical interface, mainly in grayscale consisting of a captcha image, a text input field and a submit button. The design was kept simple in order to keep the participants not distracted during the experiment. The captcha characters consisted of only numbers, therefore the input validation of the form field was activated so that only 6 digit numbers could be entered. The warning messages were arranged accordingly. All designed user interfaces are given on Figure 15.

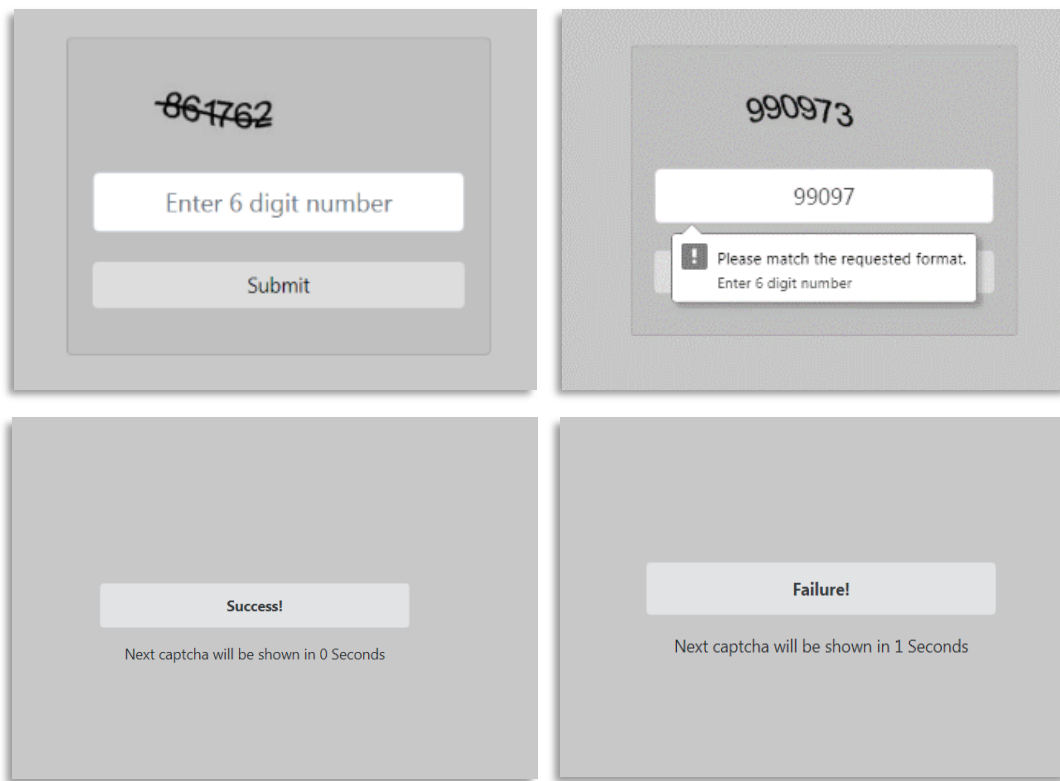


Figure 15: Screenshots for user interface of PHP application

3.2. Hypotheses

According to research questions defined in our study, we have formulated two hypotheses. In order to test them, for each visual feature and their interactions, we will check the existence of any significant mean change in HbO, HbR, HbT or Oxy values during the course of solving a text-based captcha. If we can answer “YES” to our first question, we planned to examine the mean changes in order to understand effect of difficulty levels on them. HbO will be our focus among fNIRS datas which is first indicative of

mental workload. Our expectation is HbO values will increase when the difficulty level of visual feature increases. If we got “NO” answer for the first question, we will conclude selected three features cannot produce differentiable hemodynamic response or at least it cannot be detectable using fNIRS device. If we got significant mean changes on some optodes, we will also report the affected regions of human PFC.

3.3. The Procedure

First of all, an ethical approval has been received from METU Human Subjects Ethics Committee in order to conduct the experiment with human participants using fNIRS device (Appendix C). After the development of the application and the laboratory environment setup, a pilot experiment was conducted with one participant to evaluate the developed software. The application was developed further upon the observations made in the pilot study.

Before the experiment, firstly, a signed informed consent was collected from each participant. Secondly, a Turkish version of the Edinburgh Handedness Inventory questionnaire form (Oldfield, 1971) was filled by participants, in order to determine if they were right-handed or not. In addition, all the participants filled in a demographic data form. With the purpose of following best practices for privacy of collected information, no private data that can refer individuals was included in the form. Lastly, all instructions related to avoid motion artifacts that should be followed during an fNIRS experiment have been given to participants before the experiment.

During the experiment, each participant tried to solve a total of 133 captchas in approximately 20-30 minutes. Accuracy and response time data were collected by our software and fNIRS data collected by Cobi Studio. The details of all data collection process are described in the next section.

3.4. Data Collection and Materials

All experiments were conducted in the laboratory settings of the Informatics Institute. The experiment setup is illustrated in Figure 16. In order to measure fNIRS light, fNIRS1000 model device was used. Collected data were recorded by the Cobi Studio Software. Participants interacted with a numpad that was used to enter numeric captcha values on a display. An fNIRS headband was used with 16 optodes and 48 channels.

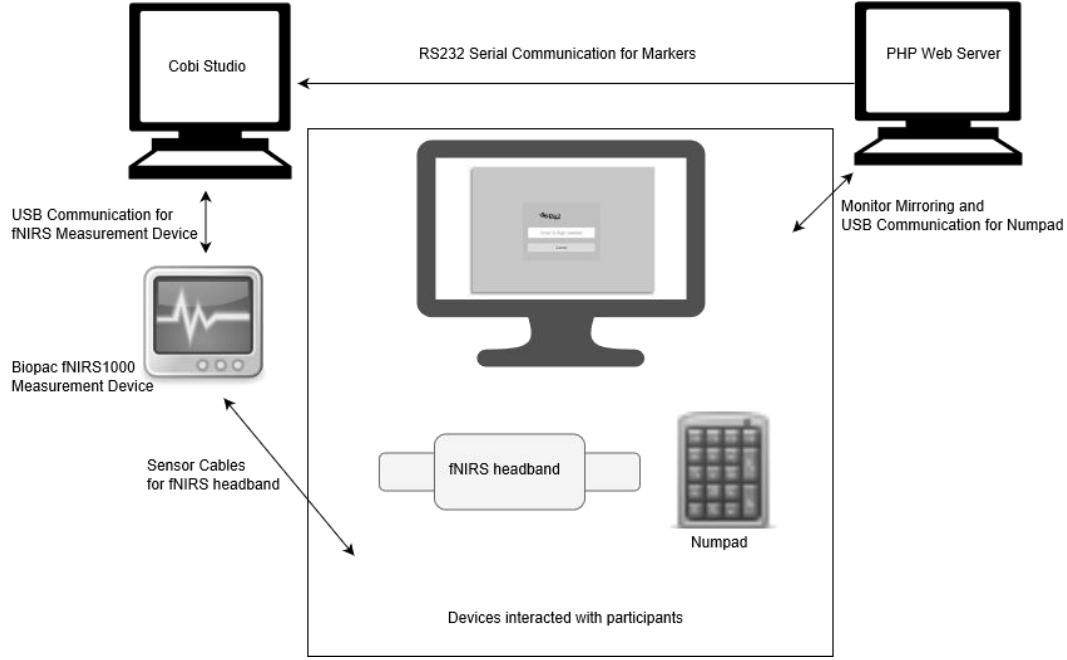


Figure 16: Experiment Setup

The developed application was deployed on a different computer than the one where Cobi Studio software was run. Our software also sent markers that defined the start of a block, the end of baseline captchas and lastly the end of a block through RS232 Serial Communication Ports to Cobi Studio software. Moreover, at the end of a block, supplementary marker was sent in order to follow the type of captcha shown in the block. The markers and their positions used in the experiments are illustrated on the Figure 17.

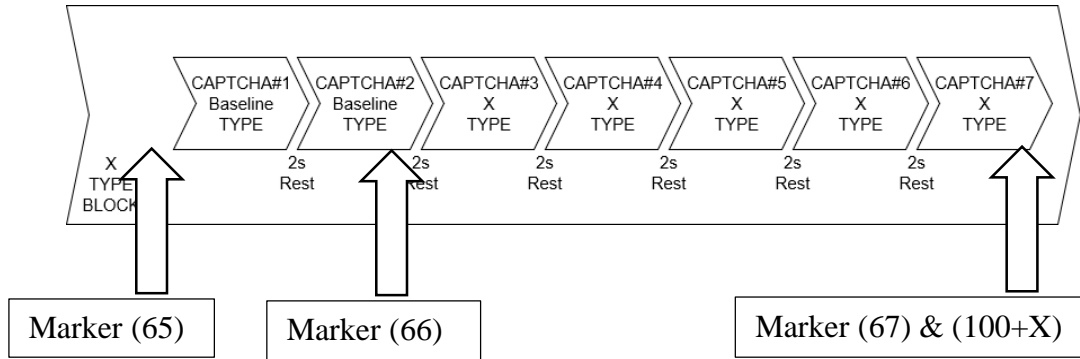


Figure 17: Marker codes and actual positions in a block design

All data need to measure accuracy and response time for each captcha types are collected by our software. The preparation of collected fNIRS data for analysis have been done by fnirSoft Professional Software. The details of this preparation process is given with in results of study on the next chapter. The picture of experiment environment in the laboratory is given on the Figure 18 with the permission of one participant.



Figure 18: The picture of experiment environment in the laboratory with the permission of one participant

3.5. The Participants

The experiment was conducted with 25 participants, who were students mostly from the Middle East Technical University, Turkey. All participants were right-handed. The mean age of the participants was 25.0 ($SD = 2.37$). Over half the sample (13) was female; 9 participants had eye-glasses and one of them had contact lens during the experiment. There was no color-blind participant as long as reported in the demographic data form. Eight participants were undergraduate student, 10 had a Bachelor's degree and the rest had a Master's degree. The results of the study are presented in the following chapter.

CHAPTER 4

RESULTS

The results of accuracy and response time collected from PHP web application are presented in Section 4.1 and the results of data recorded by Cobi Studio are presented in Section 4.2. All collected data are analyzed by IBM SPSS v25, and preparation of collected fNIRS data was performed by fNIRSoft software whose process also will be described in Section 4.2.

4.1. Behavioral Data (Accuracy and Response Time) Results

Twenty-five participants solved totally 3,325 numeric text-based captchas during experiment including test run blocks. Test run and baseline captcha tests were always the two initial stimuli in each block. Accordingly, they were excluded from the analysis reported in this section. Also, due to a technical problem of one captcha type shown for one participant, the related data were excluded. Finally, we analyzed remaining 2,160 captcha results for twenty-four participants.

The mean accuracy of overall tests is 78.7 in percentage ($SD = 28.15$) and the mean response time of overall tests is 5.5 in seconds ($SD = 29.40$). The descriptive statistics for results of accuracy and response time for each type are presented respectively in Table 4 and 5. The response time of each solved captcha was measured by a JavaScript function, namely `performance.now`, which count time in high precision after a page loads until the submit button clicked. The graphical representations of the results are also illustrated in the Figure 19 and 20. Minimum values for response time shows that none of participants did not give up completing the task during the experiment. Also in order to make this inference, every answers of every participant have been analyzed.

Table 4: Descriptive Statistics for collected accuracy in percentage

Conditions	N	Minimum	Maximum	Mean	Std. Deviation
TYPE1	24	80	100	99.16	4.08
TYPE2	24	80	100	98.33	5.64
TYPE3	24	80	100	96.66	7.61
TYPE4	24	60	100	91.66	13.07
TYPE5	24	80	100	99.16	4.08
TYPE6	24	60	100	84.16	14.42
TYPE7	24	60	100	93.33	11.29
TYPE8	24	40	100	80.83	19.09
TYPE9	24	60	100	96.66	9.63
TYPE10	24	0	100	70.00	22.06
TYPE11	24	60	100	95.83	10.17
TYPE12	24	20	100	58.33	22.778
TYPE13	24	40	100	88.33	18.57
TYPE14	24	20	100	60.83	29.76
TYPE15	24	60	100	83.33	15.22
TYPE16	24	0	80	28.33	26.32
TYPE17	24	60	100	71.66	14.34
TYPE18	24	0	60	20.00	18.65
Valid N (listwise)	24				

Table 5: Descriptive Statistics for collected response time in 1000*seconds

Conditions	N	Minimum	Maximum	Mean	Std. Deviation
TYPE1	120	2132	10272	4163.11	1294.46
TYPE2	120	1993	8631	4453.62	1359.14
TYPE3	120	2242	12800	4446.03	1541.83
TYPE4	120	2307	16646	4536.71	1836.94
TYPE5	120	2321	12305	4360.14	1511.64
TYPE6	120	2424	11143	4945.89	1597.79
TYPE7	120	2564	12479	5069.89	1754.78
TYPE8	120	2619	19483	5516.67	2717.53
TYPE9	120	2321	8286	4307.98	1213.42
TYPE10	120	3309	20813	6055.42	2743.07
TYPE11	120	2291	15733	4607.24	1795.74
TYPE12	120	3039	21280	6696.56	3027.23
TYPE13	120	2325	10173	4573.71	1437.84
TYPE14	120	2799	28500	6320.12	3355.22
TYPE15	120	2442	16047	5328.64	2276.68
TYPE16	120	2870	31182	9445.60	5356.57
TYPE17	120	2639	19683	5659.98	2420.20
TYPE18	120	3433	24566	8815.26	4322.88
Valid N	120				

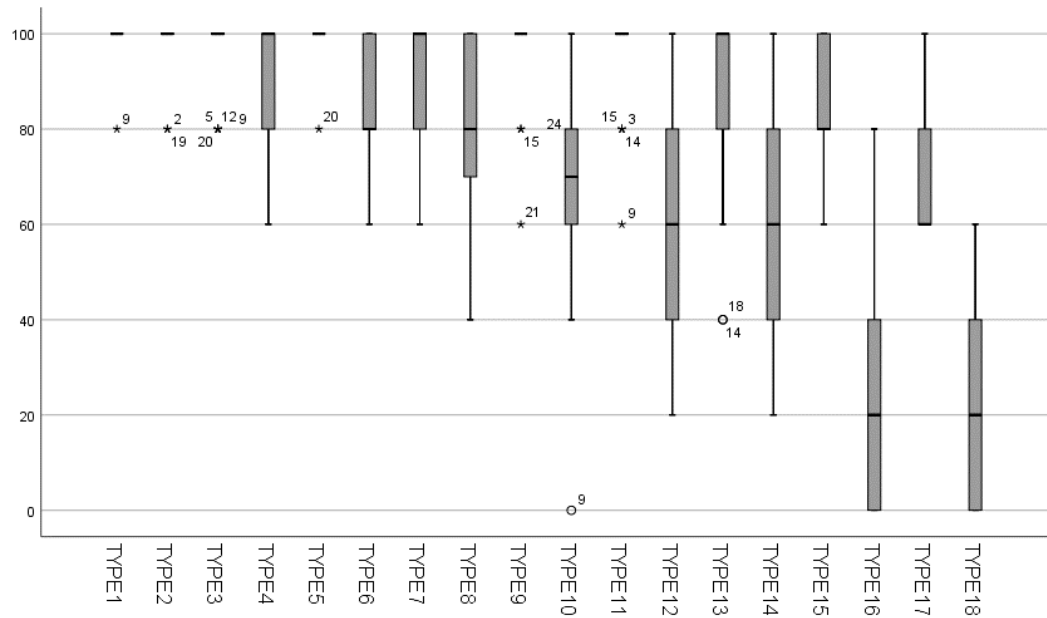


Figure 19: Box-plot for collected accuracy in percentage

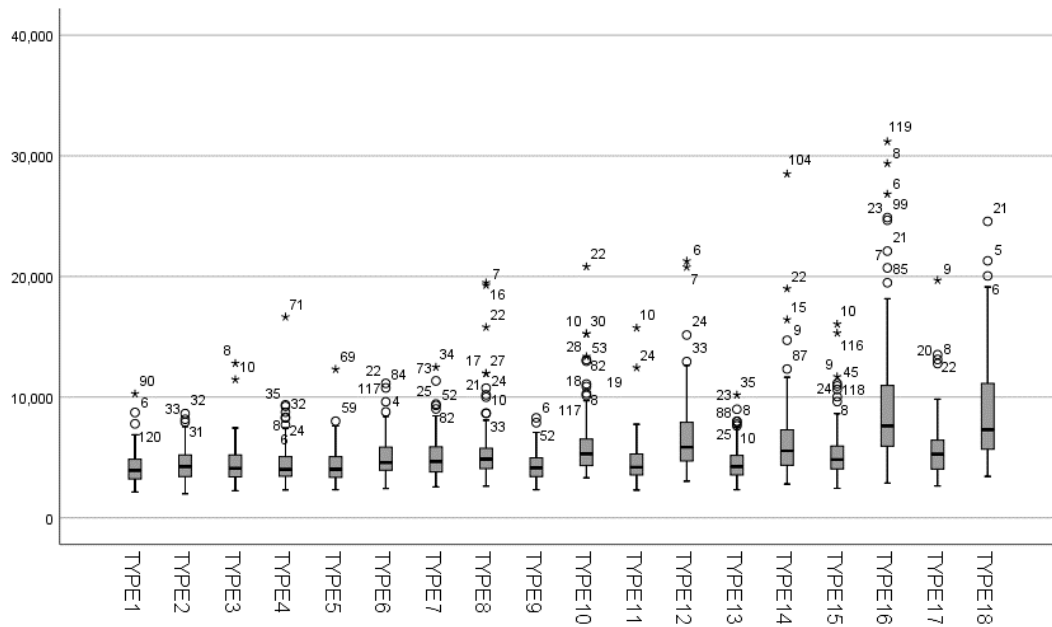


Figure 20: Box-plot for collected response time in 1000*seconds

A three-way within subjects (or repeated measures) ANOVA was conducted to compare the effect of difficulty levels of line, wave and rotation visual features on accuracy and response time in eighteen different conditions. Results showed that; There was a statistically significant main effect of line feature on accuracy, $F(2, 46) = 39.663$, $p < 0.05$. As sphericity assumption has been violated, Greenhouse-Geisser statistics is described for

this data, there was a statistically significant main effect of wave feature on accuracy, $F(1.41, 32.450) = 225.605$, $p < 0.05$. There was a statistically significant main effect of rotation feature on accuracy, $F(1, 23) = 177.429$, $p = 0.05$. There was a statistically significant main effect of line feature on response time, $F(2, 46) = 29.947$, $p < 0.05$. As sphericity assumption has been violated, Greenhouse-Geisser statistics is described for this data, there was a statistically significant main effect of wave feature on response time, $F(1.13, 26.13) = 87.795$, $p < 0.05$. There was a statistically significant main effect of rotation feature on response time, $F(1, 23) = 77.537$, $p < 0.05$.

The plots for estimated marginal means for accuracy performance data with the factor of line and waving visual features are presented on the Figure 21. The plots show that accuracy decreases when the difficulty levels of visual features increase especially with in the wave high conditions.

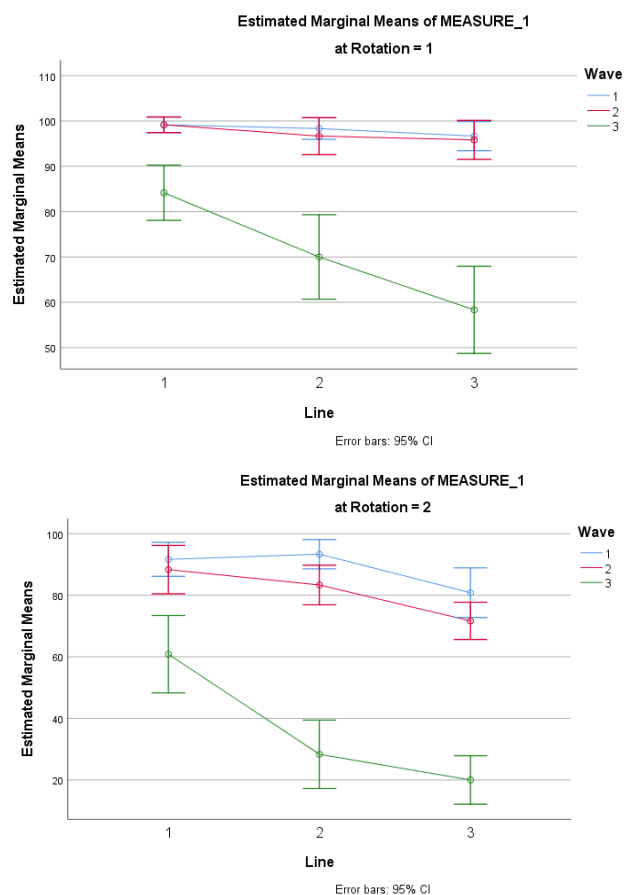


Figure 21: Estimated marginal means for accuracy performance data

The plots for estimated marginal means for response time performance data with the factor of line and waving visual features are presented on the Figure 22. The plots show that response time performance data increases when the difficulty levels of visual features increases. Although minimum values of response time data show participants did not give

up, the plots of average response times of users on difficult conditions give signals of engagement on participants. At the condition wave high and rotation feature applied, although during difficulty level of line feature increasing from level 1 to level 2, response time increases, during difficulty level of line feature increasing from level 2 to level 3, response time decreases.

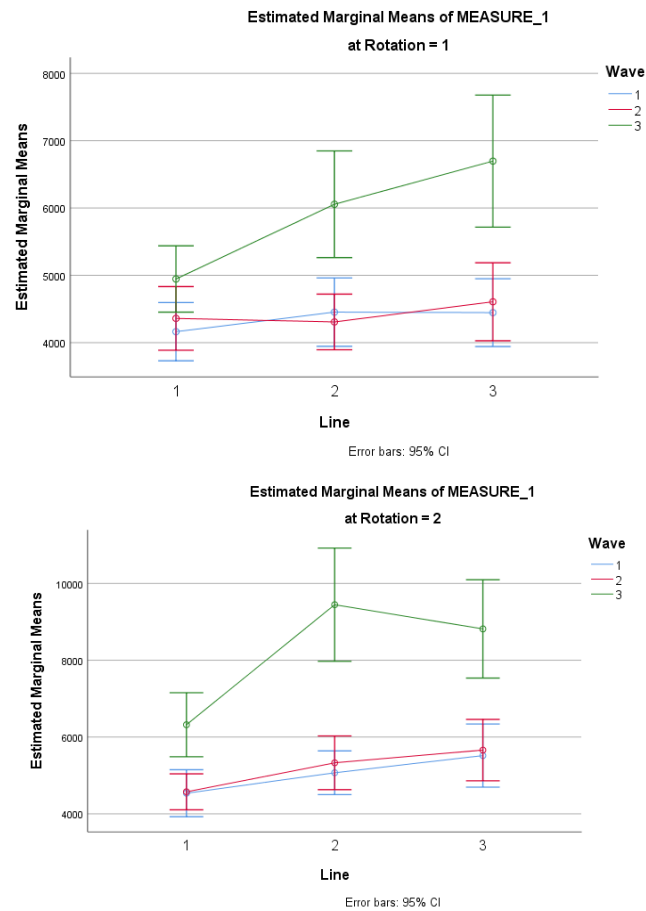


Figure 22: Estimated marginal means for response time performance data

4.2. fNIRS Data Results

Before analyzing collected fNIRS data, preparation process has been done on raw light intensity data which is summarized on the Figure 23.

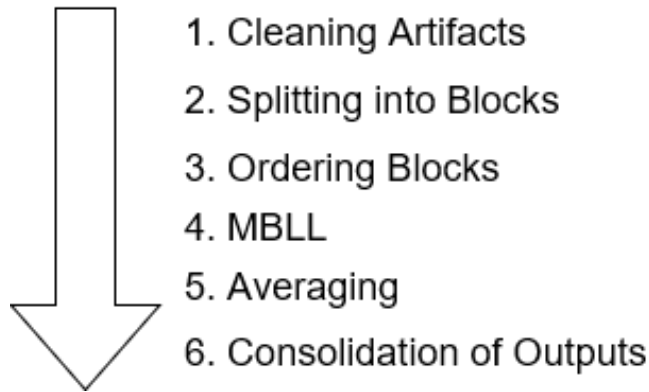


Figure 23: The preparation process of collected fNIRS data

The first step includes enhancements to deal with artifacts caused by noises or minor movements of the participants during the experiment. In this step firstly, layout views by optodes were analyzed and rejection operation were done. In most cases, those artifacts are caused by bad contact of the headband and the optodes with participants' head and/or contact of the sensors on headband with hair or eyebrows of participants. Optodes having raw signal values higher than 4000mV caused by bad contact and raw signal values lower than 400mV caused by hair or eyebrows have been rejected after each subjects' data examined. Totally 42 optodes out of 400 optodes for all participants to be rejected is detected. Later to remove noises caused by heartbeat and respiration, a Finite Impulse Response Digital Filter (FIR filter) with an order of 20 has been applied. Accordingly, we eliminated 10% of data points based on a mutual agreement of the researchers (the author and the supervisors).

As the second step, refined collected fNIRS light data have been splitted into blocks with the help of the markers sent by the application as a part of our experiment design. These markers are recorded by Cobi Studio during experiment. During the split operation, technical problems have been detected related with markers on 13 main blocks out of 450 main blocks. This problem most probably caused by USB RS232 dongle instead of using embedded one which holds some markers on buffer and later sends it sometimes. Related markers have been updated to form original blocks correctly with the help of information collected by our web application considering response time and rest time. Totally 950 variables have been generated after splitting applied.

On the third step, captcha types shown to for each participant in a main block have been ordered by checking both the markers recorded by COBI Studio and logs recorded by our PHP application. Both data were consistent. By using this order list, the mbll (Modified Beer Lambert Law) function embedded in fNIRSsoft software was applied for every block

in order to generate each participant's HbO, HbR, HbT and Oxy values for each type of captcha. The two initial captchas in each block were taken as the baseline parameter of this function and so totally 3,600 variables were created considering all four types of calculated values. Finally, after taking the average of values in each variable, the results were consolidated to construct one table that is suitable for statistical analysis.

After outlier analysis, data from three subjects were eliminated based on mutual agreement of the researchers. Later, a repeated measures ANOVA with factor TASK analysis was conducted by using different types of visual features defined in our experiment design. Different groups were created according to optodes (o1-o16) and dependent variables (HbO, HbR, HbT, Oxy) before starting the analysis. The results are presented in Table 6 and their Mauchly's test of Sphericity are presented in Table 7.

Table 6: Detected significant changes in mean values (Tests of Within-Subjects Effects)

				Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
hbo	o1	Line * Wave * Rotation	Sphericity Assumed	1.657	4	0.414	3.892	0.006	0.170
	o11	Line * Wave * Rotation	Sphericity Assumed	1.410	4	0.352	3.427	0.012	0.146
	o12	Line * Wave * Rotation	Sphericity Assumed	1.044	4	0.261	2.587	0.044	0.126
	o13	Line * Wave * Rotation	Sphericity Assumed	1.039	4	0.260	2.554	0.046	0.118
	o16	Wave * Rotation	Sphericity Assumed	0.742	2	0.371	3.686	0.034	0.156
		Line * Wave * Rotation	Sphericity Assumed	1.092	4	0.273	3.176	0.018	0.137
	o2	Line * Wave * Rotation	Sphericity Assumed	1.203	4	0.301	2.736	0.034	0.120
	o3	Line * Wave * Rotation	Sphericity Assumed	1.115	4	0.279	2.588	0.043	0.115
	o4	Line * Wave * Rotation	Sphericity Assumed	1.603	4	0.401	3.730	0.008	0.172
	o6	Wave * Rotation	Sphericity Assumed	0.206	2	0.103	3.712	0.034	0.171
hbr	o1	Line * Wave * Rotation	Sphericity Assumed	1.665	4	0.416	4.024	0.005	0.175
	o10	Line * Rotation	Sphericity Assumed	2.098	2	1.049	4.180	0.023	0.180
	o12	Line * Wave * Rotation	Sphericity Assumed	1.156	4	0.289	2.542	0.047	0.124
	o14	Line * Wave * Rotation	Sphericity Assumed	1.172	4	0.293	2.664	0.040	0.143
	o16	Line * Wave * Rotation	Sphericity Assumed	1.091	4	0.273	3.185	0.018	0.137
	o2	Line * Wave * Rotation	Sphericity Assumed	1.150	4	0.287	2.624	0.041	0.116
	o4	Line * Wave * Rotation	Sphericity Assumed	1.285	4	0.321	2.663	0.039	0.129
	o8	Line * Rotation	Sphericity Assumed	2.027	2	1.013	3.586	0.038	0.166
oxy	o1	Line * Wave * Rotation	Sphericity Assumed	1.827	4	0.457	3.425	0.013	0.153
	o10	Line * Wave * Rotation	Sphericity Assumed	1.517	4	0.379	2.850	0.029	0.130
	o11	Line * Wave * Rotation	Sphericity Assumed	2.120	4	0.530	3.554	0.010	0.151
	o13	Line * Wave * Rotation	Sphericity Assumed	1.631	4	0.408	2.533	0.047	0.118
	o3	Line * Wave * Rotation	Sphericity Assumed	2.029	4	0.507	3.022	0.022	0.131
	o4	Line * Wave * Rotation	Sphericity Assumed	2.004	4	0.501	4.108	0.005	0.186
	o9	Line * Wave * Rotation	Sphericity Assumed	1.884	4	0.471	3.464	0.012	0.148

Table 7: Mauchly's test of Sphericity of optodes for each detected significant changes in mean values

Method			Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Epsilon ^b Huynh-Feldt	Lower-bound
hbo	o1	Line * Wave * Rotation	0.393	16.263	9	0.063	0.748	0.903	0.250
	o11	Line * Wave * Rotation	0.549	11.030	9	0.276	0.817	0.996	0.250
	o12	Line * Wave * Rotation	0.357	16.922	9	0.051	0.747	0.913	0.250
	o13	Line * Wave * Rotation	0.551	10.385	9	0.322	0.795	0.973	0.250
	o16	Wave * Rotation	0.997	0.058	2	0.971	0.997	1.000	0.500
		Line * Wave * Rotation	0.533	11.599	9	0.239	0.813	0.990	0.250
	o2	Line * Wave * Rotation	0.652	7.880	9	0.548	0.835	1.000	0.250
	o3	Line * Wave * Rotation	0.575	10.203	9	0.336	0.764	0.918	0.250
	o4	Line * Wave * Rotation	0.454	12.957	9	0.167	0.797	0.989	0.250
hbr	o6	Wave * Rotation	0.758	4.711	2	0.095	0.805	0.872	0.500
hbt	o1	Line * Wave * Rotation	0.516	11.520	9	0.244	0.794	0.972	0.250
	o10	Line * Rotation	0.778	4.515	2	0.105	0.818	0.885	0.500
	o12	Line * Wave * Rotation	0.509	11.078	9	0.273	0.772	0.951	0.250
	o14	Line * Wave * Rotation	0.570	8.093	9	0.528	0.805	1.000	0.250
	o16	Line * Wave * Rotation	0.636	8.348	9	0.501	0.801	0.972	0.250
	o2	Line * Wave * Rotation	0.616	8.915	9	0.447	0.828	1.000	0.250
	o4	Line * Wave * Rotation	0.693	6.029	9	0.738	0.865	1.000	0.250
	o8	Line * Rotation	0.896	1.871	2	0.392	0.906	1.000	0.500
oxy	o1	Line * Wave * Rotation	0.403	15.813	9	0.072	0.725	0.869	0.250
	o10	Line * Wave * Rotation	0.517	11.499	9	0.245	0.825	1.000	0.250
	o11	Line * Wave * Rotation	0.638	8.278	9	0.508	0.853	1.000	0.250
	o13	Line * Wave * Rotation	0.676	6.823	9	0.657	0.862	1.000	0.250
	o3	Line * Wave * Rotation	0.481	13.478	9	0.144	0.740	0.882	0.250
	o4	Line * Wave * Rotation	0.392	15.356	9	0.083	0.784	0.968	0.250
	o9	Line * Wave * Rotation	0.421	15.943	9	0.069	0.753	0.902	0.250

Repeated measures ANOVA results show that; There was a statistically significant three-way interaction between line, wave and rotation on optode1 HbO values, $F(4,76) = 3.892$, $p < 0.05$. There was a statistically significant three-way interaction between line, wave and rotation on optode11 HbO values, $F(4,80) = 3.427$, $p < 0.05$. There was a statistically significant three-way interaction between line, wave and rotation on optode12 HbO values, $F(4,72) = 2.587$, $p < 0.05$. There was a statistically significant three-way interaction between line, wave and rotation on optode13 HbO values, $F(4,76) = 2.554$, $p < 0.05$. There was a statistically significant three-way interaction between wave and rotation on optode16 HbO values, $F(2,40) = 3.686$, $p < 0.05$. There was a statistically significant three-way interaction between line, wave and rotation on optode16 HbO values, $F(2,80) = 3.176$, $p < 0.05$. There was a statistically significant three-way interaction between line, wave and rotation on optode2 HbO values, $F(4,80) = 2.736$, $p < 0.05$. There was a statistically significant three-way interaction between line, wave and rotation on optode3 HbO values, $F(4,80) = 2.588$, $p < 0.05$. There was a statistically significant three-way interaction between line, wave and rotation on optode4 HbO values, $F(4,72) = 3.730$, $p < 0.05$.

The plots for estimated marginal means for HbO measurements with the factor of line and waving visual features are presented on the Figure 24.

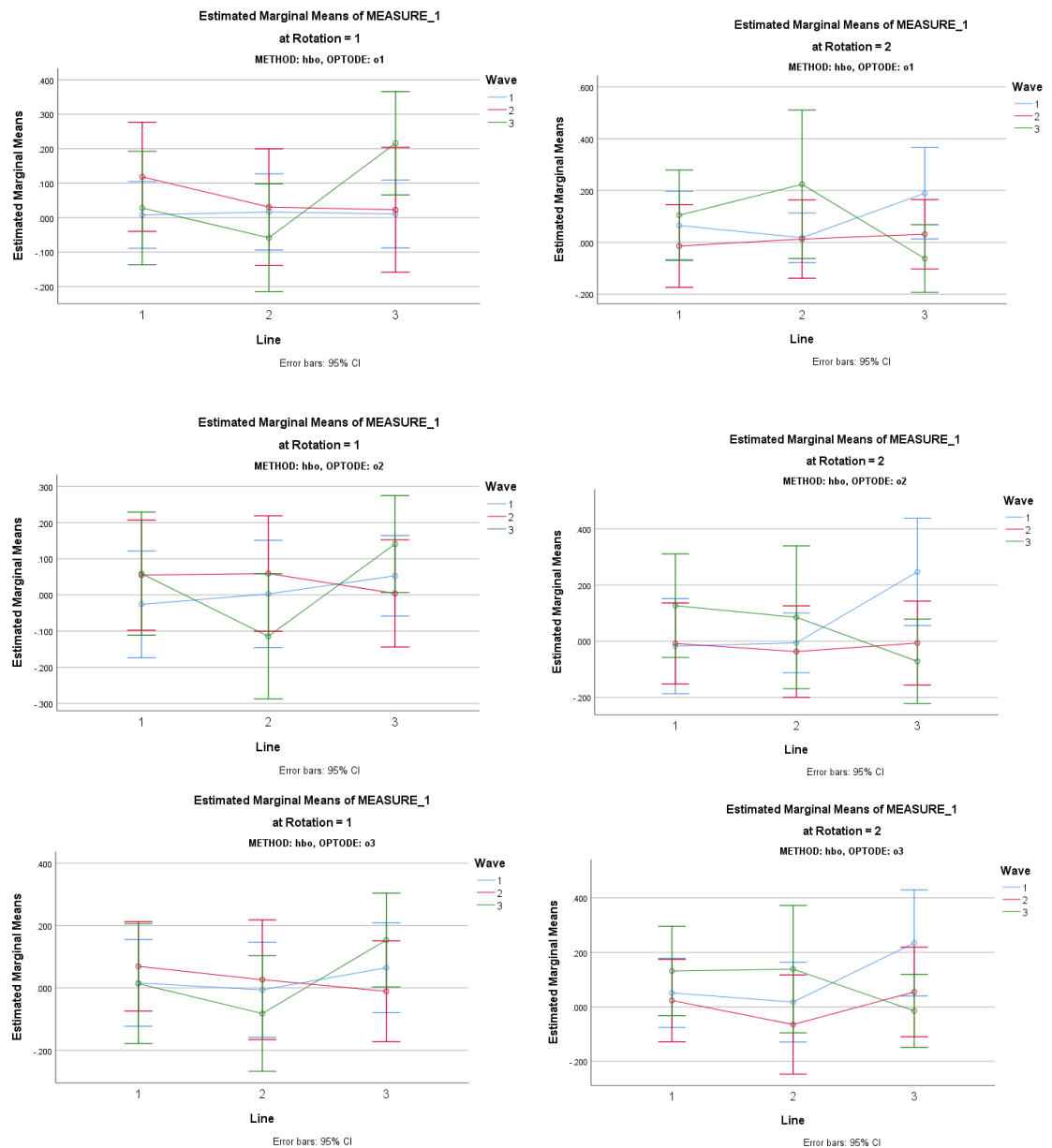


Figure 24: Estimated marginal means for HbO measurements

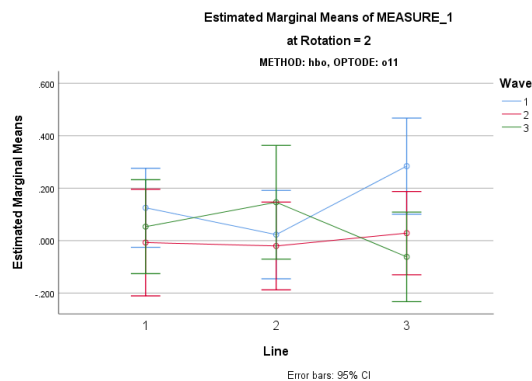
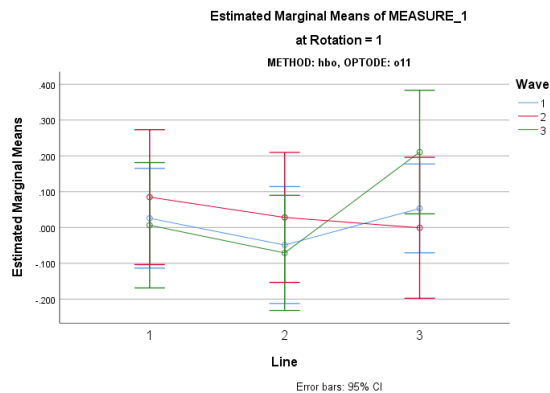
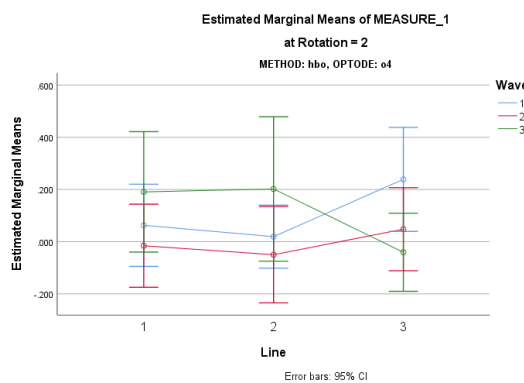
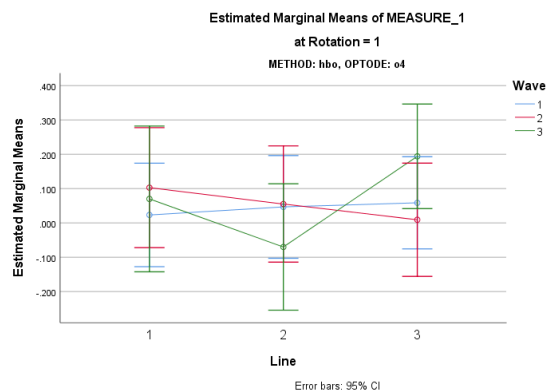
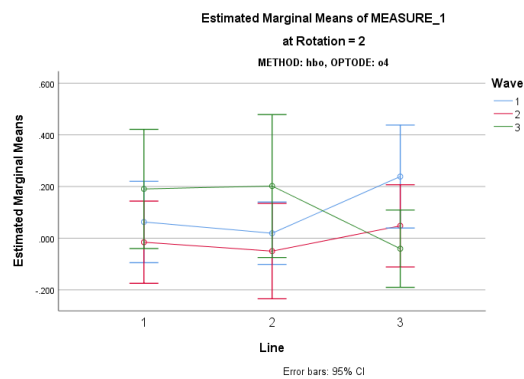
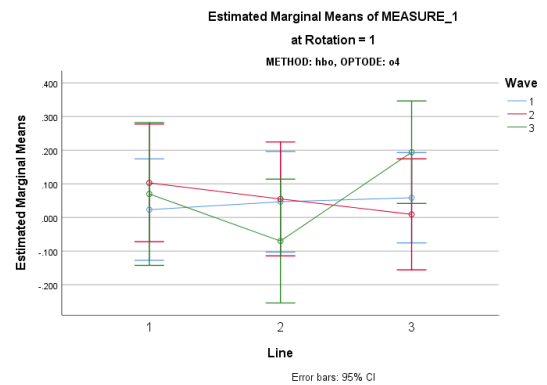


Figure 24(Cont'd): Estimated marginal means for HbO measurements

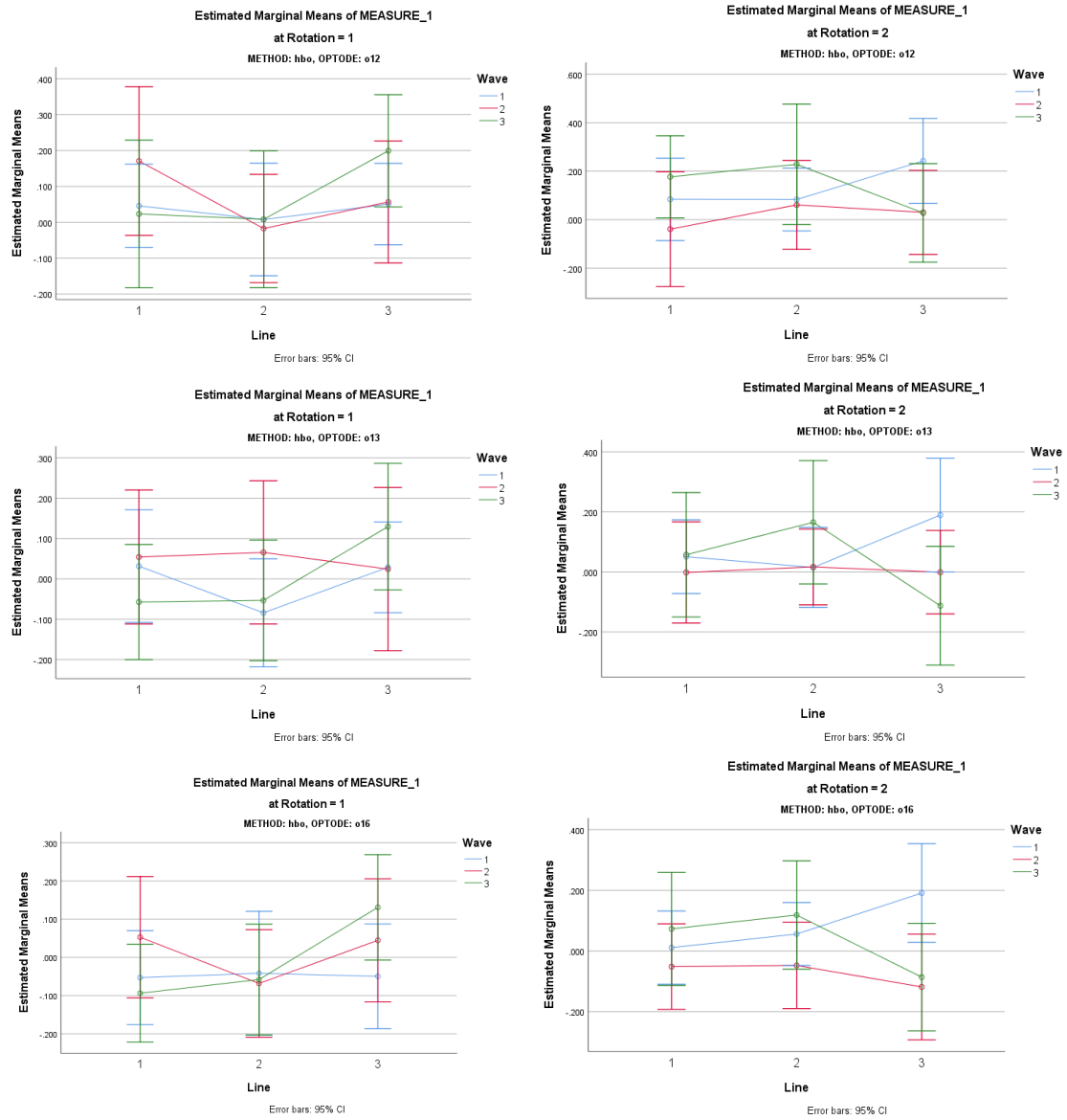


Figure 24(Cont'd): Estimated marginal means for HbO measurements

4.3. Summary of Results

This chapter presented the methodology data analyses. The analyzed data included both accuracy and response time collected by the PHP application, and fNIRS measurements collected by Cobi Studio.

A three-way within subjects (or repeated measures) ANOVA was conducted to compare the effect of difficulty levels of line, wave and rotation visual features on accuracy and response time in eighteen different conditions. The results showed that there was statistically significant main effect of three visual features on accuracy and response time behavioral data. Although minimum values of response time data show participants did not give up, the plots of average response times of users on difficult conditions give signals of engagement on participants.

After preprocessing the fNIRS measurements, a repeated measures ANOVA has been conducted by using factors of different levels of visual features. There was no observed main effect that returned a significant difference for any of the HbO, HbR, HbT or Oxy measurements. However, interactions of the three visual features (line * wave * rotation) returned significant differences on some of the optodes given in Table 6. Affected optodes are also marked with dark color on the optode layout in the Figure 25 for only HbO measurements. The plots for estimated marginal means shows the direction of changes for HbO measurements, as shown in the Figure 24. The findings are divergent for each feature on different optodes. Interesting pattern was found on during interaction of three feature. Increasing the thickness of line with the high waving feature increased oxygen consumption when the rotation feature was not applied, but decreased consumption when characters were rotated. Accordingly, these results show that the selected features do not exhibit a linear, additive effects.

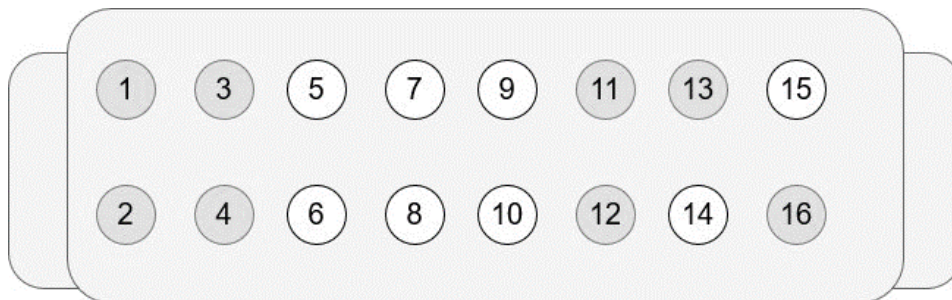


Figure 25: Optode layout on the headband and affected optodes

CHAPTER 5

CONCLUSION

The conclusion, contribution to research literature and limitations and future works are presented in this chapter.

Captcha challenges have an important role as a countermeasure of automated web application attacks by distinguishing automated attacks accomplished by scripts and humans. Studies on captchas address three aspects, namely design, robustness and usability of captchas. In the present study we focused on the usability of text-based captchas. A review of the literature showed that most of the usability studies on captchas have used behavioral measures, in particular accuracy and response time as measurement criteria for quantitative analyses, as well as questionnaire forms for assessing mental workload (e.g., NASA-TLX scoring).

Recently optical neuroimaging techniques have found limited use in usability and mental workload relevant to cyber security. As of our knowledge, no research has investigated the effects of captcha solving using neuroimaging techniques. We propose that using brain imaging has high potential to improve our understanding of mental workload in. Given that Functional Near Infrared Spectroscopy (fNIRS) is a neuroimaging technique that is also widely used in neuroscience due to its low-cost, portability, and non-invasiveness, in the present study, we employed fNIRS and reported the analysis of an experiment that was carried out with 25 participants while they were solving a group of text-based captcha with different visual features.

The specific purpose of the present study was to understand whether a text-based captcha solving task produces a differentiable hemodynamic response in human brain. If it were produced, secondly our next aim was to examine the effects of selected three visual features (Line, Wave, Rotation) on recognition in order to understand which regions of the PFC were affected by them. The features were selected based on the previous research reported by Google reCAPTCHA v1 final design. During the experiment, eighteen types

of captcha were presented to the participants, as combinations of the selected features with different levels of difficulty.

The results showed that the combination of the types with different difficulty levels has affected the behavioral measurements (accuracy and response time). The waving high feature interaction with the rotation and line features has low usability according. As the difficulty increased, usability decreased in terms of accuracy and response time. Although minimum values of response time data show participants did not give up, response times of users on difficult conditions give signals of engagement on participants

On the other hand, fNIRS results revealed different patterns compared to the behavioral results. There was no observed main effect that resulted in a significant difference for any of the HbO, HbR, HbT or Oxy measurements. However, interactions of three features (line * wave * rotation) led to significant differences on the most of the optodes. The direction of the changes with the feature difficulty for HbO measurements revealed divergent patterns on the optodes. A specific pattern that was compatible with the behavioral findings was that increasing the thickness of the line with the high waving feature led to an increase in oxygen consumption when the rotation feature was not applied, but a decrease in consumption when the characters were rotated. The results show that the selected visual features (line, wave, rotation) do not exhibit a linear, additive behavior. It is likely that a relative disengagement from the task may have been experienced, which was reflected in their performance reflected. Another possibility for non-linearity may be the combination of the visual features led to emergent Gestalt effects that influenced the perception of the overall captcha figure.

For summary, using fNIRS device may not be helpful during the analysis of individual visual features on mental workload at least for selected features, but it may be helpful during the analysis of combined (aggregated) features. For example, final candidate designs which has more than two visual features can be compared in terms of mental workload in quantative way.

In the field of cyber security, it is known that the importance of authentication for web security. In order to offer strong authentication mechanisms, captcha challenges also play an important role in the mitigation factor against both the brute force and other unwanted automated attacks. Information disclosures may occur as a result of trials with brute force attacks. Brute force or dictionary attacks can be target on user password fields or on other critical fields. If it is an automatic attack on different than authentication mechanism, undesirable use of an application's resources can be achieved by collecting data or attempts to exhaust the system.

Although the captcha challenges are a common prevention mechanism against both bruteforce and other automatic attacks, it is evaluated by many researchers that whether the difficulty of challenges is too easy or too hard has an impact on usability as well as robustness of security mechanisms. In the challenging captchas, users have problems during completing the task due to reduced readability, which in turn makes the users

uncomfortable. In order to resolve this problem, users mostly reload the captcha image to capture a readable one or cancel the process which they wanted to perform.

Cyber security and usability should always be in balance in corporate systems. Stronger challenges used in the sensible systems make it easy to decode the code automatically, and this may result in security incidents. On the other side, if it is too hard it can lead to user dissatisfaction. This research study provides a novel methodology that can be used to calibrate captcha configurations in the design step of secure software development processes performed by business units that demands usability and cyber security units that demands robustness. Considering that wireless fNIRS devices are becoming widespread, smaller and cheaper, it will be possible for the end user to have these measuring tools and to be part of the system within human in the loop paradigm. Users with different capabilities can use an application simultaneously. Further studies may reveal usage of similar methodologies on the design of security mechanism with self-calibration properties according to the user profile or their current mental state.

The present study has limitations that should be addressed by future work. The fNIRS method, despite its broad use is subject to motion artifacts. Although the participants were expected not to move their head during the experiment, approximately 20 minutes, it was a challenging issue for the participants some movements were indispensable. These motion artifacts were eliminated by filtering and manually by checking anomalies in data patterns. These limitations are common in any physiological data collection technique and further developments in sensor and data processing technologies will improve data cleaning. Another limitation of the study was the selection of only three visual features and their combinations due to experiment practice. The role of additional futures would improve the coverage of the study with the cost of longer experiment duration and complexity. For future studies, it is planned to broaden the set of text-based captchas in terms of visual features. Another potential domain of research is a comparison between text-based captchas and image-based captchas in terms of mental workload. Also instead of expecting from participants to enter shown characters on display by using numpad or keyboard, another experiment design may be conducted such as expecting from participants to repeat out loud in order to analyze and minimize motion artifacts.

REFERENCES

- Ahn, L. von, Blum, M., Hopper, N.J., & Langford, J. (2000). The Official CAPTCHA Site Retrieved June 23, 2019, from <http://www.captcha.net>.
- Anderson, B. B., Kirwan, C. B., Jenkins, J. L., Eargle, D., Howard, S., & Vance, A. (2015, April). How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2883-2892). ACM.
- Baird, H. S., & Riopka, T. P. (2005, January). Scattertype: a reading captcha resistant to segmentation attack. In *Document Recognition and Retrieval XII* (Vol. 5676, pp. 197-207). International Society for Optics and Photonics.
- Beheshti, S. M. R. S., & Liatsis, P. (2015, December). CAPTCHA Usability and Performance, How to Measure the Usability Level of Human Interactive Applications Quantitatively and Qualitatively?. In *2015 International Conference on Developments of E-Systems Engineering (DeSE)* (pp. 131-136). IEEE.
- Belk, M., Fidas, C., Germanakos, P., & Samaras, G. (2015). Do human cognitive differences in information processing affect preference and performance of CAPTCHA?. *International Journal of Human-Computer Studies*, 84, 1-18.
- Bhatt, S., Agrali, A., McCarthy, K., Suri, R., & Ayaz, H. (2019). Web Usability Testing With Concurrent fNIRS and Eye Tracking. In *Neuroergonomics* (pp. 181-186). Academic Press.

- Brodić, D., Amelio, A., & Janković, R. (2018). Exploring the influence of CAPTCHA types to the users response time by statistical analysis. *Multimedia Tools and Applications*, 77(10), 12293-12329.
- Bursztein, E., Aigrain, J., Moscicki, A., & Mitchell, J. C. (2014). The end is nigh: Generic solving of text-based captchas. In *8th {USENIX} Workshop on Offensive Technologies ({WOOT} 14)*.
- Bursztein, E., Bethard, S., Fabry, C., Mitchell, J. C., & Jurafsky, D. (2010, May). How good are humans at solving CAPTCHAs? A large scale evaluation. In *2010 IEEE symposium on security and privacy* (pp. 399-413). IEEE.
- Bursztein, E., Martin, M., & Mitchell, J. (2011, October). Text-based CAPTCHA strengths and weaknesses. In *Proceedings of the 18th ACM conference on Computer and communications security* (pp. 125-138). ACM.
- Bursztein, E., Moscicki, A., Fabry, C., Bethard, S., Mitchell, J. C., & Jurafsky, D. (2014, April). Easy does it: more usable CAPTCHAs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2637-2646). ACM.
- Carlén, M. (2017). What constitutes the prefrontal cortex?. *Science*, 358(6362), 478-482.
- Chellapilla, K., Larson, K., Simard, P., Czerwinski, M., & Czerwinski, M. (2005, April). Designing human friendly human interaction proofs (HIPs). In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 711-720). ACM.
- Chen, J., Luo, X., Guo, Y., Zhang, Y., & Gong, D. (2017). A Survey on Breaking Technique of Text-Based CAPTCHA. *Security and Communication Networks*, 2017.
- Chew, M., & Tygar, J. D. (2004, September). Image recognition captchas. In *International Conference on Information Security* (pp. 268-279). Springer, Berlin, Heidelberg.
- Elson, J., Douceur, J. J., Howell, J., & Saul, J. (2007). Asirra: a CAPTCHA that exploits interest-aligned manual image categorization.
- Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage*, 63(2), 921-935.
- Ferrari, M., Mottola, L., & Quaresima, V. (2004). Principles, techniques, and limitations of near infrared spectroscopy. *Canadian journal of applied physiology*, 29(4), 463-487.
- Gafni, R., & Nagar, I. (2016). CAPTCHA–Security affecting user experience. *Issues in Informing Science and Information Technology*, 13, 063-077.

- Gordieiev, O., Kharchenko, V. S., & Vereshchak, K. (2017, May). Usable Security Versus Secure Usability: an Assessment of Attributes Interaction. In *ICTERI* (pp. 727-740).
- Hill, A. P., & Bohil, C. J. (2016). Applications of optical neuroimaging in usability research. *Ergonomics in Design*, 24(2), 4-9.
- Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J., Sassaroli, A., & Fantini, S. (2009, April). Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2185-2194). ACM.
- Hsu, C. H., & Lee, Y. L. (2011). Usability study of text-based CAPTCHAs.
- Hsu, C. H., & Lee, Y. L. (2011, July). Effects of age groups and distortion types on text-based CAPTCHA tasks. In *International Conference on Human-Computer Interaction* (pp. 453-455). Springer, Berlin, Heidelberg.
- International Organization for Standardization/International Electrotechnical Commission. (2011). ISO/IEC 25010-Systems and software engineering—systems and software Quality Requirements and Evaluation (SQuaRE)—system and software quality models. Authors, Switzerland.
- Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323), 1264-1267.
- León-Carrión, J., & León-Domínguez, U. (2012). Functional near-infrared spectroscopy (fNIRS): principles and neuroscientific applications. v *Neuroimaging-Methods. IntechOpen*.
- Lukanov, K., Maior, H. A., & Wilson, M. L. (2016, May). Using fNIRS in usability testing: understanding the effect of web form layout on mental workload. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4011-4016). ACM.
- Moradi, M., & Keyvanpour, M. (2015). CAPTCHA and its Alternatives: A Review. *Security and Communication Networks*, 8(12), 2135-2156.
- Morein, W. G., Stavrou, A., Cook, D. L., Keromytis, A. D., Misra, V., & Rubenstein, D. (2003, October). Using graphic turing tests to counter automated DDoS attacks against web servers. In *Proceedings of the 10th ACM conference on Computer and communications security* (pp. 8-19). ACM

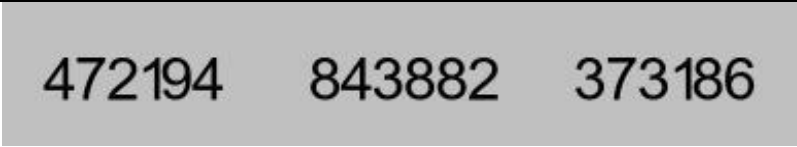




- Nanglae, N., & Bhattarakosol, P. (2015, June). Attitudes towards Text-based CAPTCHA from developing countries. In *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (pp. 1-4). IEEE.
- Neupane, A., Saxena, N., & Hirshfield, L. (2017, April). Neural underpinnings of website legitimacy and familiarity detection: An fnirs study. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1571-1580). International World Wide Web Conferences Steering Committee.
- Nielsen, J. (2003). Usability 101: Introduction to usability.
- Norbu, K., & Bhattarakosol, P. (2012, December). Factors Towards the effectiveness of CAPTCHA. In *2012 7th International Conference on Computing and Convergence Technology (ICCCT)* (pp. 566-570). IEEE.
- OWASP, T. (2017). Top 10-2017 The Ten Most Critical Web Application Security Risks. Retrieved June 23, 2019, from: https://www.owasp.org/images/7/72/OWASP_Top_10-2017_%28en%29.pdf.pdf
- Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2018). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*.
- Quaresima, V., & Ferrari, M. (2019). Functional near-infrared spectroscopy (fNIRS) for assessing cerebral cortex function during human behavior in natural/social situations: a concise review. *Organizational Research Methods*, 22(1), 46-68.
- Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Pavia, J. M., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *Neuroimage*, 85, 6-27.
- Shirali-Shahreza, S., Ganjali, Y., & Balakrishnan, R. (2011). Verifying human users in speech-based interactions. In *Twelfth Annual Conference of the International Speech Communication Association*.
- Singh, V. P., & Pal, P. (2014). Survey of different types of CAPTCHA. *International Journal of Computer Science and Information Technologies*, 5(2), 2242-2245.
- Tariq Banday, M., & A Shah, N. (2009). Image flip CAPTCHA. *The ISC International Journal of Information Security*, 1(2), 105-123.


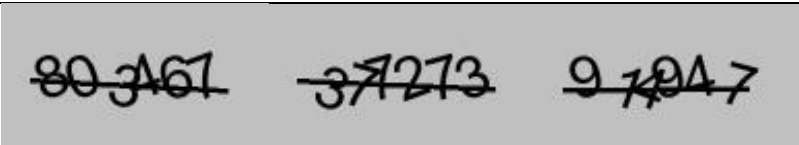
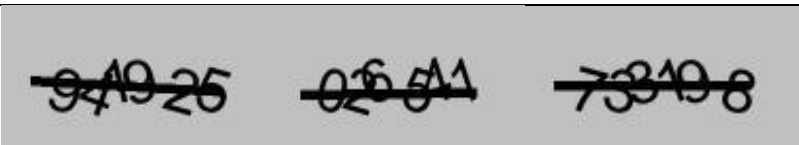
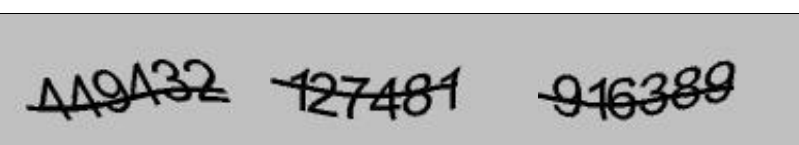
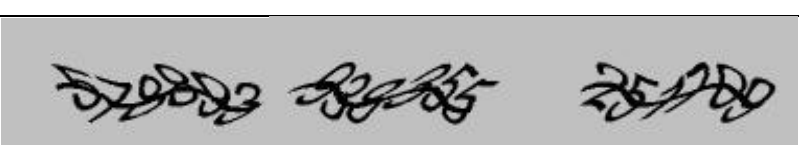


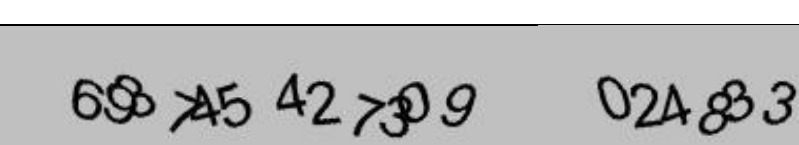
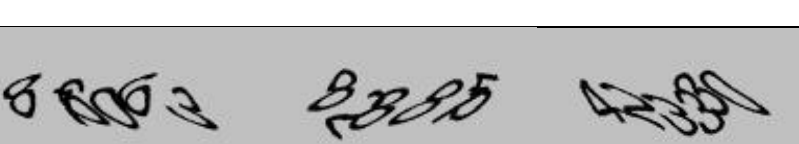
APPENDICES

APPENDIX A

SAMPLES OF ALL GENERATED CAPTCHA TYPES

Three examples for each types of captcha generated during our experiment is listed below.

Type: 1 Baseline	
Type: 2 Line Thin	
Type: 3 Line Thick	
Type: 4 Rotation20	
Type: 5 Wave Light	

Type: 6 Wave High			
Type: 7 Line Thin+Rotation20			
Type: 8 Line Thick+Rotation20			
Type: 9 Line Thin+Wave Light			
Type: 10 Line Thin+Wave High			
Type: 11 Line Thick+Wave Light			
Type: 12 Line Thick+Wave High			
Type: 13 Wave Light+Rotation20			
Type: 14 Wave High+Rotation20			

Type: 15 Line Thin+Wave Light+Rotation20	
Type: 16 Line Thin+Wave High+Rotation20	
Type: 17 Line Thick+Wave Light+Rotation20	
Type: 18 Line Thick+Wave High+Rotation20	

APPENDIX B

MODIFICATIONS ON COOL-PHP-CAPTCHA

In order to generate a captcha image, cool-php-captcha code was used with a little adjustment. The code changes related with some visual and security features to generate exact captchas designed in our experiment are given below. Trivial changes in configuration values used in the code like background, foreground color, font family, font size are not included in this appendix.

For Line Feature:

```
protected function WriteLine() {  
-    $x1 = $this->width*$this->scale*.15;  
+    $x1 = $this->width*$this->scale*.07;  
  
-    $y1 = rand($this->height*$this->scale*.40, $this->height*$this->scale*.65);  
-    $y2 = rand($this->height*$this->scale*.40, $this->height*$this->scale*.65);  
+    $y1 = rand($this->height*$this->scale*.50, $this->height*$this->scale*.60);  
+    $y2 = rand($this->height*$this->scale*.50, $this->height*$this->scale*.60);  
}
```

For Rotation Feature:

```
protected function WriteText($text, $fontcfg = array())  
  
+    $array = [-20,-15,15,20];  
    for ($i=0; $i<$length; $i++) {  
-        $degree = rand($this->maxRotation*-1, $this->maxRotation)*$this->difficulty;  
+        if ($this->maxRotation == 0){  
+            $degree = 0;  
+        } else {  
+            $degree = $array[rand(0,3)];  
+        }  
    }
```

For Waving Feature:


```
protected function WaveImage() {  
    // X-axis wave generation  
-    $xp = $this->scale*$this->Xperiod*rand(1,3) * $wdf;  
-    $k = rand(1, 100);  
+    $xp = $this->scale*$this->Xperiod*(($this->difficulty==2)?1:2) * $wdf;  
+    $k = rand(0, 100);  
    // Y-axis wave generation  
-    $yp = $this->scale*($this->Yperiod)*rand(1,2) * $wdf;  
+    $yp = $this->scale*($this->Yperiod)*(($this->difficulty==2)?1:2) * $wdf;
```


APPENDIX C

ETHICAL APPROVAL FOR THE EXPERIMENT

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

DUMLUPIHAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 69
ueam@metu.edu.tr
www.ueam.metu.edu.tr



ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

Sayı: 28620816 *1640*

11 ARALIK 2018

Konu: Değerlendirme Sonucu

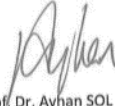
Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

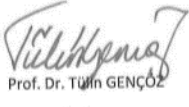
Sayın Doç.Dr. Cengiz ACARTÜRK

Danışmanlığını yaptığımız Emre MÜLAZIMOĞLU'nun "CAPTCHA Kullanılabilirliğinin FNIRS ve Eye Tracking Araçları ile Ölçülmesi" başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay 2018-EGT-172 protokol numarası ile araştırma yapması onaylanmıştır.


Saygılarımla bilgilerinize sunarım.



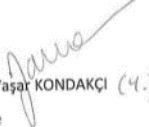
Prof. Dr. Ayhan SOL
Üye



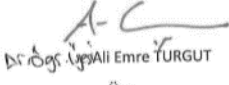
Prof. Dr. Tülin GENÇOZ
Başkan




Prof. Dr. Ayhan Gürbüz DEMİR
Üye




Prof. Dr. Yaşar KONDAKÇI (4.)
Üye



Doç. Dr. Ali Emre TURGUT
Üye



Doç. Dr. Emre SELÇUK
Üye



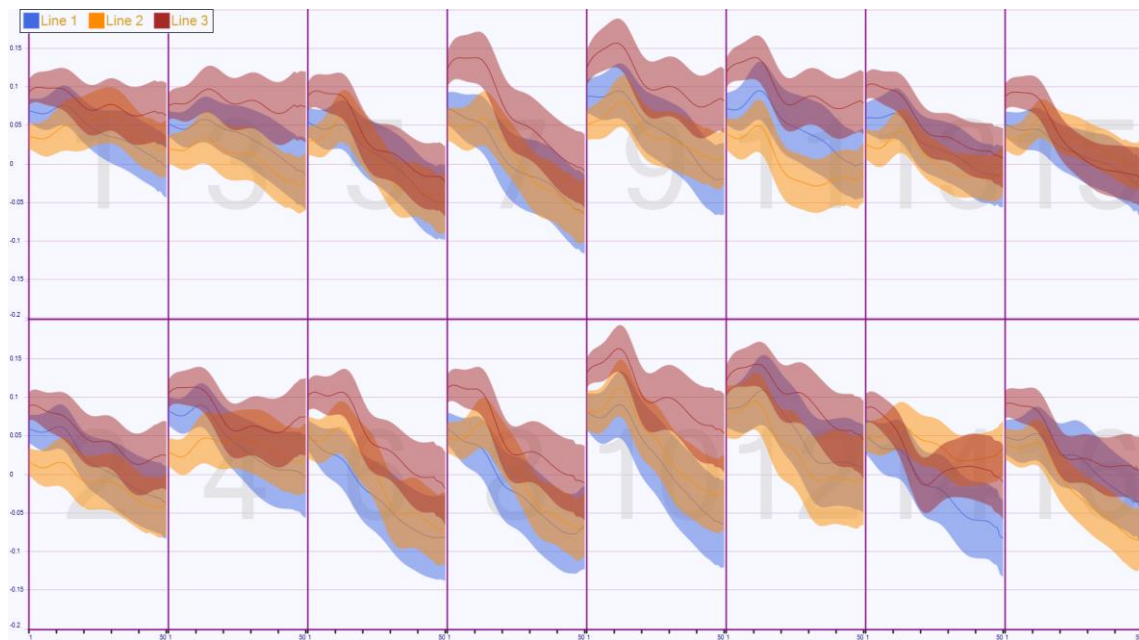
Doç. Dr. Üyesi Pınar KAYGAN
Üye

APPENDIX D

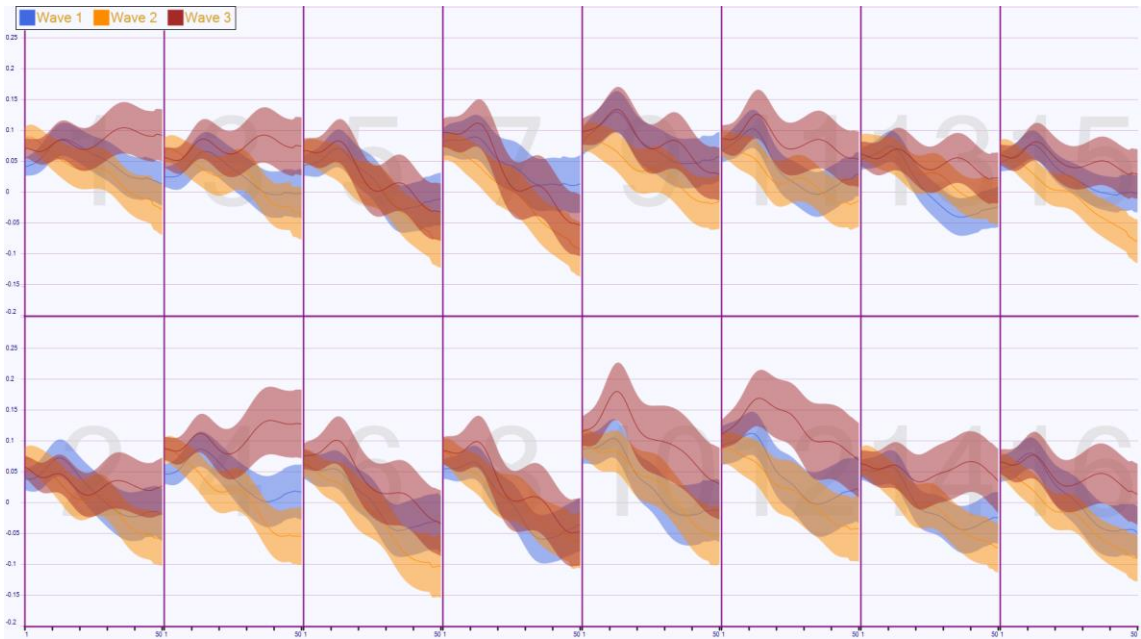
MEAN CHANGES OF HBO ON TIMELINE FOR EACH TYPE

HbO values for each optode in time-series are presented below. In each graph, the values for different difficulty levels are shown with different color described as in the legend on the top of figures.

For the main effect of visual feature **Line**:



For the main effect of visual feature **Wave**:



For the main effect of visual feature **Rotation**:

