The University of Nottingham

School of Computer Science

Data Collection and Analysis of the Linkage between Mental Workload and Spontaneous Facial Expression on Pattern Recognition Task

Zequn Yu

Submitted in part fulfilment of the requirements for the degree of
BSc Computer Science of the University of Nottingham, April 2018

Abstract

The study of spontaneous facial expression and mental workload is giving rise to a wide range of possibilities. In this project, 20 participants were recruited to take part in a pattern recognition task. A face-workload dataset was built. For half of the dataset, various of measures including functional Near Infrared Spectroscopy (fNIRS) were used to evaluate different levels of mental workload. A machine learning model was built to predict mental workload based on fNIRS signals and achieved 73.33% on F1 score. For the other half, the facial data was processed using the state of art analysis tool OpenFace. The result indicated the linkage between mental workload and several facial Action Units.

Acknowledgements

I would like to express thanks to:

- Max Wilson, who gave suggestions to my research study and dissertation writing;
- Horia Maior, who shared his knowledge on fNIRS setup and measurement;
- the staffs at the Mixed Reality Lab, who provided the place for data collection;
- all the participants to the experiment, who agreed on the use of their data;
- Natasa Milic-frayling, who encouraged me to do this research-oriented project;
- my parents, who are always providing unconditional support to me.

The project would not have been finished without you.

Contents

List of Tables

List of Figures

List of Code Snippets

Abbreviations

AFED Automatic Facial Expression Detection

AMSD Automatic Mental States Detection

ANN Artificial Neural Network

EEG Electro Encephalo Graphy

FACS Facial Action Coding System

fNIRS functional Near Infrared Spectroscopy

HCI Human Computer Interfaces

ISA Instantaneous Self Assessment

LDA Linear Discriminant Analysis

SVM Support Vector Machile

SVR Support Vector Regression

1 Introduction

Scientists have been putting efforts to analyse human face for over two centuries [18]. The work to analyse facial expressions had been first taken by psychologists until the second half of the twentieth century, when computer scientists gradually reached the top of technology wave. Face analysis is becoming an active research field and has already resulted in many revolutionary commodity, such as FaceID in Apple iPhone X [35]. In 1978, Ekman and Friesen developed the FACS system [13] to describe human facial movements systematically. The system, consisting of 44 facial Action Units, has been widely used in facial expression analysis since then.

Interestingly, there is another field that was first investigated by psychologists, then also concerned by computer scientists: the evaluation of mental workload. Researchers in HCI use various of subjective and objective methods to measure mental workload, in terms of limited capacity of human mind [41]. Among the methods, physiological measure also attracted attention in recent years. Since discovered in 1992, fNIRS has been playing important role for helping researchers to understand the functionality of human brain, as well as contributing to its application of many commercial products [19]. Many experiments were conducted to measure mental workload on different tasks using fNIRS as physiological data [41] [26] [52]. In this project, the scope of my study embraces the two fields: facial expression analysis and mental workload evaluation.

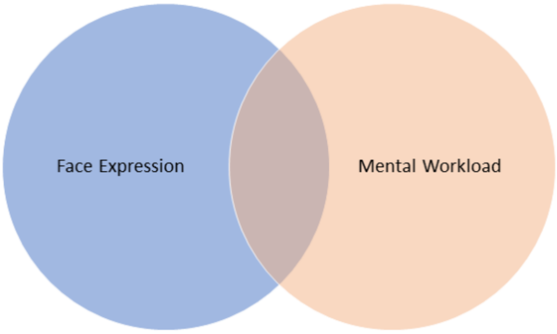


Figure 1: Overview: the scope of my study

1.1 Motivation

The era is witnessing profound changes brought by the development of machine learning. It earned success not only in Computer Vision, but also in Voice Recognition, Natural Language Processing and complex games such as the Game of Go [55]. Among all the fields, the development of comprehensive AFED system is receiving more attention over the past decade. Facial expressions include posed expressions and spontaneous expressions. AFED systems targeted to posed expressions has achieved striking progress. In contrast, because that spontaneous expressions are often subtle and less-intensive, performance on them usually suffers decrease. Even though, Girard et al. verified the viability of performing analysis on spontaneous expressions in complex and unscripted environment [21]. This suggests more potentiality of research focusing on spontaneous expressions.

Due to various difficulties and technical limit, nowadays it is still not possible to build a comprehensive AFED system with high performance in all circumstances. However, Lucey et al. proposed the way to narrow the context of target application and choose face-pain context as their target [40]. Their dataset has been used for training many AFED systems such as OpenFace [4]. Inspired from it, the target context in this project can be regarded as face-workload. In many cases, AFED systems follow the rule that the amount of training samples is positive correlated with the performance. The lack of representative data, both in quantity and high-quality is lasting for years. Involuntary behaviour is becoming the new research focus in the field, yet spontaneous data of affective states is still hard to find [65]. Therefore, the first motivation of this project is that a dataset with spontaneous expressions under different workload levels could contribute to the study of automatic facial expression analysis.

In another study area, the evaluation of human mental states, including affective states and cognitive states, is concerned by many researchers. It is predicted that the study of HCI is moving from computer centred to human centred [8]. An ideal AMSD system would detect subtle changes from users, and interact with users correspondingly. Relevant methods have been attempted by extracting optical [11] or physiological [52] data. These systems could lead to various significant applications. For example, an intelligent car system to predict fatigue could suggest rest alarm for drivers, which may help avoiding accidents and saving drivers life [6] [31]. Measuring engagement is beneficial in teaching environment, because both teachers and learners can use the information to make learning more effective [45].

Mental workload has been suggested relative to human performance [42]. It is intuitive that high workloads cause high stress level, which has been proved by [22]. In this project, the work to evaluate and predict workload could be meaningful in many cases. For instances, psychologist could use the information to monitor and cure patients. and average people could understand and manage their mind better. This becomes the second motivation of the project.

1.2 Objectives

The general aim of this project is to explore the linkage between facial expressions and mental workloads. Concretely, I will **design an experiment, collect data for building a face-workload dataset, use different measures to confirm workload levels, and conduct FACS analysis using the state of art face analysis tool: OpenFace**. The raw dataset should contain: videos of facial data, well-designed questionnaire to raise different levels of workload, subjective measure result, participants performance result and fNIRS brain data. The dataset should be enough to support analysis in this project and inspire future related research. The work would include:

1. to design an experiment for data collection;
2. to present the information sheet, consent form and ethic checklist;
3. to implement a computer based software for the experiment;
4. to recruit participants and conduct the experiment;

5. to apply suitable scenario to synchronise time for fNIRS data, videos and tasks;
6. to pre-process raw fNIRS data and videos;
7. to apply machine learning techniques to fNIRS data for mental workload prediction;
8. to analyse the linkage between facial AUs appearance, intensity and mental workload.

Full version of the dataset should also include frame by frame manually annotated facial AUs as ground truth facial labels. However, this is beyond the scope of this project due to time limit. This will be further discussed in subsection 5.2. The work could be done as an extension of this project.

1.3 Outline

The following sections are structured as below. I first introduce background knowledge and related work. Then I describe details of the experiment design and data collection. The next section is data processing and analysis. Finally, research outcomes and limits will be discussed, and a brief personal reflection conclude the dissertation.

2 Related Work

Topics I study in the project include: facial expression basis, automatic facial expression analysis, existing datasets, mental workload evaluation, workload experimental protocol, and relevant face-workload data analysis. In this section I review literature relating to the above topics to the best of my knowledge.

2.1 Facial Expression Basis

Facial expression is an important channel for human to convey information and express their mental state. In this part I focus on two dimensions of facial expression problem space: level of description, posed or spontaneous.

Level of Description

Facial expressions include moves of eye lids, eye brows, nose, lips and skin texture, which are generated by facial muscles. According to Fasel et al.[18], typical time of muscular changes lasting between 250ms to 5s. The important points of facial expression can be summarised as the location, intensity and dynamics. Intensity refers to the degree of facial expression. Since facial expressions vary from different people, it is hard to agree the absolute degree of intensity. Dynamics of facial expressions is another important factor because some subtle changes of facial expression is hard to observe from static images.

Most of current studies of facial expression are based on prototypic models. A common prototypic model is emotion. The analysis between facial expression and emotion can be tracked back to 1960s, when Paul Ekman started his study as a pioneer in the field. At the time, not many scientists regarded face as a reliable and accuracy source to extract emotional information [14]. In 1970s, Friesen and Ekman proved the universality of facial emotions in both spontaneous and posed expressions without culture differences [15]. They also raised the definition of six basic emotions of facial expression, which are well known as happiness, sadness, fear, disgust, surprise and anger [16]. It is noteworthy that facial expression is not equal to human emotion, which also consists of voice, pose, gestures and other physiological signs. Fasel et al. defined three temporal parameters of facial expressions, which are attack, sustain and relaxation [18]. They also summarised two categories of facial emotions measurement, which are judgment-based and sign-based. In the judgment-based method, the ground truth emotion is agreed by a group of coders. In contrast, it is also possible to measure emotions according to facial AUs appearance, which is called the sign-based method.

However, facial emotions have relatively low occurrence frequency and can not describe all facial activities. Friesen and Ekman devoted to develop a framework to measure face moves systematically. The FACS system describes facial expressions based on 44 facial AUs. 30 of the facial AUs are related to facial muscles and the other 14 AUs are unspecified [24]. Although there are only 44 facial AUs, more than 7000 different AUs combinations have been observed [60]. This represents not only the six basic emotions, but also more possibilities of facial expression, which makes it suitable to describe complex spontaneous facial expressions [65]. Today, FACS system is becoming the most well known and widely used system to measure face moves scientifically.

Posed or Spontaneous

In general, facial expressions can be grouped by posed expressions and spontaneous expressions. Posed expressions are presented by subjects on request, and has no relations with real mental state. For example, a posed smile does not mean happiness of the subject. Despite the shortcoming, posed expressions are usually exaggerated and easy-to-detect, which reduces the difficulty when training AFED systems. On the contrary, spontaneous expressions are more smooth and subtle. The two expressions differ in both influenced muscles and dynamics [64]. From the view of brain activities, posed expressions have been found relative to cortical areas, whereas spontaneous expressions at sub-cortical level.

Differences of the two expressions can be distinguished using FACS system. For example, posed sadness can be easily noticed because few people can perform the actions volitionally. A real sadness often includes the combination of AU 1, 4 and 15, which can be hardly found in posed action without training. This interestingly contributes to the development of lie detection system [33].

For automatic analysis system, performance on the two expressions also varies noticeably. Bartlett et al. compared the performance of a linear SVM to detect several AUs using Ekman-Hager database [5]. Figure 2 illustrates the result. Although we have no knowledge of the details of the unpublished Ekman-Hager database, spontaneous classifier received much poorer performance than posed classifier. This is aligned with my previous experiment, in which our algorithm improved the performance in a dataset of posed expressions dramatically but not in a dataset of spontaneous expressions. According to Whitehill et al., this could be explained by four reasons. First, spontaneous expressions are usually in lower intensity. Second, the dynamics is very subtle. Then, movement such as speech articulation often influences the presence. Finally, head pose is easier to vary during spontaneous expressions [64].

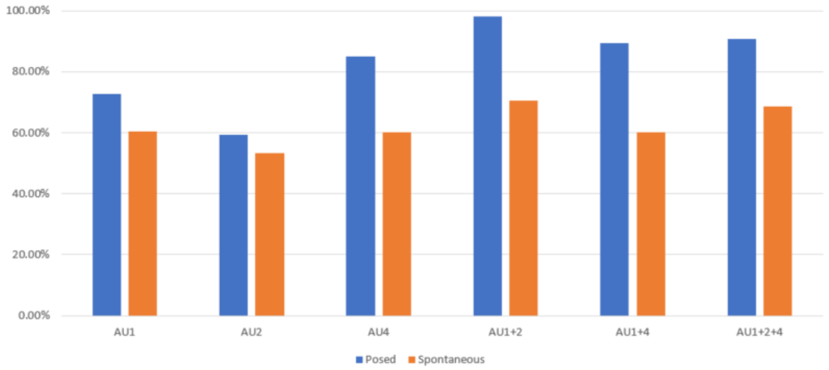


Figure 2: Example of a SVM classifier accuracy performance on several AUs

2.2 Automatic Facial Expression Analysis

As many researchers reported, manual analysis is extremely time-consuming and sometimes unaffordable [56]. For instance, manually labelling a one-minute short video using FACS would cost several hours to a FACS expert [49]. At present, the tremendous amount of raw facial data are impossible to be labelled manually. This highlights the importance of automatic facial expression analysis. With the help of machine learning techniques, this field is making striking progress. AFED system can be used to discriminate fake expressions from real and help with identity security check. A comprehensive review of the field is beyond the scope of my knowledge, and readers are referred to [64] and [18]. Instead, in this part I seek to summarise the general pipeline of AFED system and some available analysis tools. I pay special attention to OpenFace, the state of art face analysis toolkit that I will mainly use in this project. For the review on all current facial analysis tools, please refer to [4].

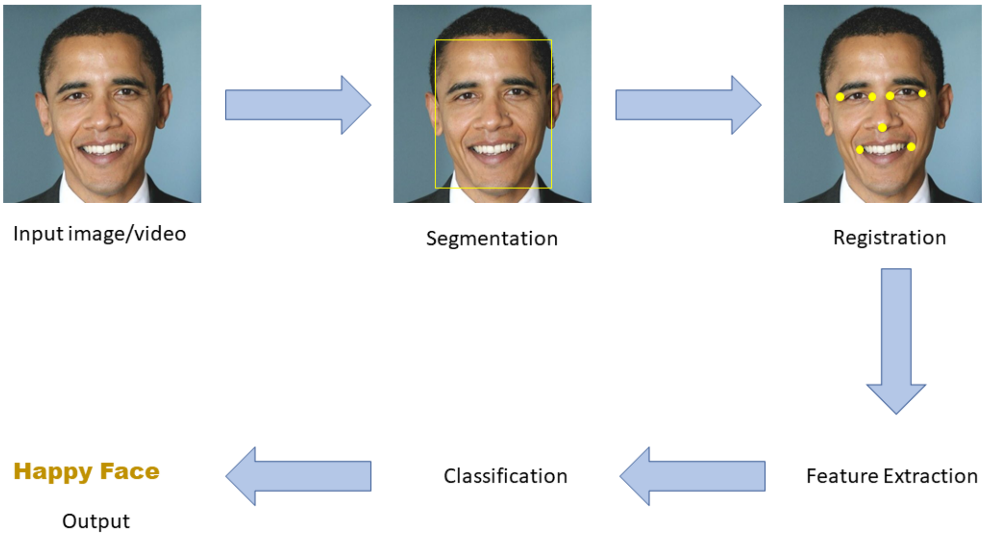


Figure 3: Pipeline of a common automatic facial analysis system

Segmentation is the first step of automatic facial analysis. In this step the system first recognises if an input image contains face. If the input is a video, it can be viewed as a sequence of images. In general, detection methods can be grouped by static and sequential methods. Sequential methods make the use of time differential information. The most famous architecture is proposed by Viola and Jones [63]. Different size of patches from the image are fed into a trained classifier to make decisions. For quick decision, the crucial point is that the classifier can discriminate most of non-face patches using very simple algorithms and only apply complex algorithms to those uncertain [64]. In OpenFace a simple convolutional neural network is used to predict the landmark error [4], which serves similar purpose of quick discrimination.

Facial landmark detection is taken in the face registration step. For the most successful systems being used currently, feature detectors (patch experts) are used to find features by searching image patches [64]. Popular methods for training patch experts include SVR and logistic regression. OpenFace uses a neural network, namely Local Neural Field patch expert, demonstrating better performance than the common SVR method. The patch expert enforces the importance of spatial similarity and sparse constraints [3]. Another component in OpenFace landmark detection model is Point Distribution Model, which helps to capture different variations of landmark shape. Initialised by the face detector in dlib library, the landmark detection model can detect 68 facial landmarks in real time [4].

Feature extraction is the next step. There are a number of different methods to make use of geometric or appearance-based features. Geometric method is based on purely relative positions of eyes, mouth, etc. While it is intuitive to use, evidence shows that appearance-based method is more reliable and lead to better performance [64]. OpenFace combines both the two methods. Appearance-based method is used to extract Histograms of Oriented Gradients following by Principal Component Analysis, which reduces dimensions from 4464 to 1379. Together with another 227 dimensional vectors from the geometric method, there are finally 1606 vectors as the extracted features [2].

The last step is classification. The most popular model for facial expression analysis is SVM. Whitehill et al. pointed out that classification is much less significant than the selection of features [64]. For this project, I focus on AU prediction. In OpenFace, a linear kernel SVM is used for AU presence prediction and a linear kernel SVR for AU intensity prediction [4].

Apart from OpenFace, LAUD [32] is an alternative toolkit for AU prediction. However, LAUD can only deal with static images. Jiang et al. suggested that their LBP method received better performance, but the implementation of LBP has not been available to public. There are a bunch of commercial products such as iMotions[1]. Commercial products are usually delivered in black box, which makes it impossible to explore the details and evaluate the performance of the systems.

2.3 Existing Datasets

Datasets are indispensable for the development of any AFED systems and automatic analysis systems with other purposes. New models need to be trained and algorithms need to be

[1] https://imotions.com/facial-expressions

evaluated using datasets. We particularly focus on datasets containing FACS facial AUs. In this part, I summarise the 7 datasets that were used to train the AU model in OpenFace. Some datasets can also be used to detect other affective states such as pain level. In the mean time, readers are referred to [49], [24], [43] and [65] for more information about existing datasets.

The Cohn-Kanade AU-Coded Facial Expression Database, also known as the CK database, is the most widely used dataset for facial AUs analysis. A total of 504 images, along with both basic emotions and facial AUs labels are available to public [24]. In July 2010, the team made the upgraded version CK+. In this version, more images with validated AU codes were added. Also, a number of spontaneous expressions were available in the dataset [39].

Lucey et al. collected a dataset from patients suffering from shoulder pain. Participants were asked to perform several motions and their spontaneous facial expressions were recorded. They collected 200 videos and annotated 48398 FACS coded frames, along with participants self reported pain level [40]. They published the data as the UNBC-McMaster Shoulder Pain Expression Archive Database, which is an example of narrowing the context of the target application to face-pain analysis. The BP4D database is a spontaneous dataset collected by Zhang et al. The data is built by asking 41 participants to finish a sequence of tasks for triggering emotional expressions. The expressions include the six basic emotions, physical pain, and embarrassment. Overall, the dataset has more than 600 thousands FACS annotated frames, and provides 3D facial data to ameliorate various issues of 2D facial data [66]. The DISFA dataset, in a similar way, collected 27 participants spontaneous expressions by asking them to watch a collection of videos from YouTube. As stimuli, videos for evoking happiness received more than 50% success rate. But for fear and sadness, the success rate was less than 10% [43]. In the SEMAINE dataset, 150 participants were asked to interact with an intelligent HCI system called Sensitive Artificial Listener. Both their face and voice data are recorded [44]. The data includes affective and cognitive states such as thinking and concentration, which have high correlation with the measure of mental workload in this project.

OpenFace has also made use of the datasets with posed expressions such as GEMEP-FERA and Bosphorus. The GEMEP-FERA dataset has been used as the challenge data in the FG 2011 Challenge. The dataset contains 289 static images collected from 10 subjects. They can be categorised into 5 posed emotional expressions, which are anger, fear, joy, relief and sadness [10]. Besides, earlier in 2008, Savran published 4652 posed images with a group of expressions expressed by professional actors [53]. The use of extensive datasets enables OpenFace to achieve the state of art performance in various experimental conditions.

2.4 Mental Workload Evaluation

Over the last 40 years there is no consensus of a widely accepted definition of mental workload. However, many different definitions emphasised the capabilities and limitations of users [7]. In HCI, the evaluation of mental workload in terms of capabilities and limitations is always an important field to study. High mental workload has been proved highly associated with mental fatigue in pilots and drivers driving tasks [6], which can easily lead to severe safety issues.

Researchers have done multiple experiments trying to build a reliable system to evaluate mental workload. The methods can be summed up to task performance measure, subjective measure and physiological measure. Task performance measure can be further divided into primary task and secondary task measure [41].

Performance Measure

In primary task measure, mental workload is estimated based on the difficulty of tasks. It is usually not a good choice to use primary measure as the only method. It is assuming that high difficulty tasks result in high mental workload, which can hardly be true in all time. For example, the Graduate Record Examinations (GRE) is widely used as an admissions requirement to Graduate Schools in the United States. A GRE problem is categorised into one of five difficulty levels in advance. The primary task measure in this case would be predicting high mental workload for participants when doing hard GRE problem. However, this is not necessarily true due to the different background of participants. An English native speaker may find a pre-categorised hard verbal problem easy, which, in result, could not cause high mental workload for them.

A secondary task measure is useful in such cases, when providing additional information of task performance. This could be the measure of response time to answer a GRE problems. However, task performance measure methods are still far from perfect to measure mental workload due to uncontrolled and other unknown factors, and a combination with more measure approaches is often a better choice for the evaluation [7].

Subjective Measure

Since the capabilities and limitations of mental workload vary from different users, in subjective measure users are asked to report the mental workload by themselves. There are two famous evaluation approaches of subjective measure, NASA-TLX and ISA. NASA-TLX is a multi-dimensional rating designed for tasks ranging from simple to sophisticated. The measured scales include Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance [25]. NASA-TLX is so popular that a simple search from Google returns 82900 citations in English and other languages around the world. NASA-TLX provides an overview of workload evaluation, but its complexity makes it not an optimal option for short period evaluation during tasks [41]. On the contrary, the five-point scale rating method ISA is another popular evaluation. The simplicity of ISA makes it the best choice to evaluate short periods. The study in [61] validated subjective measures, and pointed out that subjective measurement approaches should be carefully chose based on the assessment goal.

Physiological Measure

Physiological measure suggests another possibility to evaluate mental workload. Scanners such as EEG and fNIRS are common choices. Today, many commercial brain scanners are available on the market. Emotiv[2] and Neurosky[3] are two cost affordable EEG scanners, and OctaMon[4] is the new fNIRS scanner available in the Mixed Reality Lab at the University of Nottingham. OctaMon produces more scientific data than the other two, which makes it

[2] https://www.emotiv.com/

[3] http://neurosky.com/

[4] http://www.artinis.com/octamon/

suitable for this project. For the complete review of physiological devices readers are referred to [59] and [19].

The study by Frey et al.[36] proved that EEG is beneficial to workload, attention and emotions measures. However, Maior et al.[41] suggested that EEG is very sensible to movement, therefore may not be a good choice to perform tasks that need many moves. Another viable scanner is fNIRS, which depends on blood oxygenation level of electrical voltage [19]. It is reported as a scanner with new brain technology to effectively measure mental workload under different complexities of user interface [27]. The reasoning to measure mental workload using fNIRS is the sensitive changes of blood flow. Oxyhemoglobin (HbO2) and deoxyhemoglobin (HbR) are expected to change in brain tissue under different levels of mental workload [52]. The drawback of fNIRS is that there is a physiological delay for measured blood and actual brain activity, which makes it hard to perform real-time workload evaluation [41] and causes difficulty in synchronising time with other input modalities.

2.5 Workload Experimental Protocol

Most of current studies on mental workload were set up in lab setting environment. In other word, we hardly find any mental workload experiments that are set up in totally natural, spontaneous environment. This means all the studies face a problem: to design proper tasks as stimuli to raise different levels of workload for classification. In this part I focus on tasks that were used in previous experiments.

One type of tasks is working memory task. A famous example is the N-back task. It consists of a sequence of stimuli each appearing in a shot period. Participants need to judge if the current letter matches the one which appears n times earlier in the sequence. Theoretically, workload is higher when people try to memorise the stimulus in more items ago. In practice, researchers found strong relation between the magnitude of n and error rate. 1-back and 3-back were found the best combination to distinguish low and high workload, while 4-back seems too hard for average people to perform [34]. Sassaroli et al. designed a similar task to ask participants to count the colour of sections in a simulative computer based cube. Each side of the cube was displayed in a short period and finally participants were asked to count the total number of different colours in all sides. They defined four levels of difficulty corresponding to 0, 2, 3, 4 different colours [52].

Another type is mathematics task. One of the well known example is called the countdown task. In countdown task, participants are asked to use 6 numbers (each only once) to get as close as possible to the target number within a given time, using the four basic operators: add, subtract, divide, and multiply. For demand evaluation, Pike et al. asked two independent persons to agree the difficulty, and verified it using the Cohen's Kappa test. They recorded the psychological data using fNIRS and found non-verbal tasks tend to cause higher levels of mental workload than verbal tasks [50]. Another commonly used method is digit multiplication task, in which participants are asked to do calculate the product of n digit numbers. Stone and Wei used three levels of difficulty, which are the easy questions (1 digit number multiply 1 digit number), medium (2 digit) and hard (3 digit) [56]. They hardly found distinct differences between medium and hard questions.

Ryu and Myung designed combined tasks to simulate an instrument landing course, in which participants need to find the correct glide slope for landing. They used different speeds of the movement velocities in the simulator as the changes of difficulty levels and found the associations with eye blink interval and heart rate. The tasks also contain arithmetic problems, which are simply two numbers addition. The number of digits is used to separate difficulty levels. They reported its association with alpha suppression derived from EEG data [51].

In addition, some researchers used computer game as the task. Different levels of game represent different levels of difficulty. In Air Traffic Control Game, participants are asked to coordinate departure and landing in an airport. Their result also verified the associations with both ISA and fNIRS measures [41]. Intuitively, controlling more aeroplanes at the same time means more demand on brain. There are parallels between this experiment design and the one used by Marinescu et al., where participants need to shot target balls in another computer game, and the complexity of target balls reflects the difficulty. Marinescu et al. also recorded facial thermography data, and their results showed that nose temperature highly correlates to mental workload measured by subjective approaches in both NASA-TLX and ISA [42].

2.6 Face-Workload Analysis

In this part, we focus on the study between facial expressions and mental states relating to mental workload. Littlewort et al. conducted the experiment to find children's facial expressions during various problem solving tasks including touch-recognition game, spatial puzzle problem and arithmetic problem. They provided evidence that facial AUs can be used to find spontaneous facial expressions in mental states such as frustration. They also used machine learning techniques to find children's age based on AUs appearance [38]. Grafsgaard et al. used their automatic facial AUs detector to find its linkage with engagement, frustration and learning. In their experiment participants are asked to use a computer software to do a Java learning tutorial. They reported indications of some AUs correlating to several mental states. For example, brow lowering (AU 4) has been found associated with the negative mental state frustration [23]. Although they did not perform any mental workload measures, it is intuitive to assume frustration is accompanied by high mental workload. The outcome of the research could provide hints for the study in this project. Stone and Wei designed the task to ask participants to do arithmetic multiplication problems. Videos of their facial data were recorded and mental workload was reflected by EEG data and subjective measures. They regarded the total number and the intensity of measured AUs as extracted features, and reported its relations to mental workload. However, they did not apply any machine learning techniques for analysis and failed to provide analysis on any specific facial AUs [56].

In 2006, Savran et al. collected a dataset with facial images, fNIRS and EEG data for emotion detection by showing participants images. Their work includes the use of multiple devices, therefore the time synchronisation issue was addressed. The drawback of the experiment is that they used an outdated facial expression analysis scheme. Meanwhile, they did not address the issue that fNIRS device covers the forehead of face, resulting in incomplete facial analysis [54]. The two issues will be further discussed in section 3. In spite of performing multiple demand tasks to measure mental workload and stress from facial expressions, Dinges et al. did not use the FACS system for face analysis. Instead, they developed several versions of deformable mask to format the face. Based on this scheme, they also proposed the protocol of an AMSD system to discriminate mental workload automatically and reported the accuracy in 75% - 88% [11].

For all the studies discussed above, an important issue is that they did not attempt to share the collected data and in any sense. Also, most of the studies listed in this part were finished many years ago and seems to be outdated. In recent years, two main progresses contribute to today's study. First, the wide use of the FACS system, and the striking achievement of automatic analysis systems suchlike OpenFace, make facial analysis more systematic and accurate. Second, the use of fNIRS as physiological input provide the new approach to measure mental workload. With the mentioned issues and progresses in mind, in this project I seek to:

- explore mental workload measurement by integrating fNIRS data and other methods;
- take the advantage of the state or art OpenFace for facial analysis;
- build and share a face-workload dataset to support or inspire future related research.

3 Experiment

In this section, I first describe the details of questionnaire design. I also introduce the software that was developed for the computer based experiment, and other devices that were used for data collection. Various of protocol designing problems are discussed, following by the preprocessing and description of the dataset.

3.1 Questionnaire Design

In this project, designing a high-quality task is very important in order to raise mental workload changes and collect meaningful data. To simplify the task, I chose to design the questionnaire to distinguish two levels of question, easy and hard. As summarised in subsection 2.5, I have various options such as working memory, mathematics, or game task. I finally used **pattern recognition** task, in which participants are asked to choose the correct logical pattern based on given options.

I have several reasons for choosing pattern recognition task. First, the N-back working memory task and game task have been taken in similar research projects in [41] and [42]. For mathematics task, there exist some drawbacks. As Sassaroli et al. pointed out, nevertheless fNIRS is less sensitive to participants move than EEG or other brain scanners, recording data will eventually be affected [52]. Besides, for mathematics task, participants will need to look down at scratch paper to calculate the result. This leads to the rotation of the head, therefore making it harder to analyse facial expressions. Second, pattern recognition task is culture fair. This means different language levels will not affect the result of the test. I expect most of participants to finish the test within 30 minutes, and the length is practical for pattern recognition task to be finished. These factors make it the best choice for this project.

The database of selected questions can be found in IQ Research[5]. Originally, there are 40 questions without explicit difficulty level. Intuitively, high level difficulty relates to high level mental workload. Therefore, we choose participants performance on the questionnaire as the *primary measure*. To agree the questions difficulty, I invited three independent evaluators to give their opinion in 5 levels (1:very easy, 2:easy, 3:normal, 4:hard, 5:very hard) for each of the 40 questions. The three evaluators were not invited to take part in the formal experiment. Based on their opinions, I took the arithmetic mean *ave* as the value to annotate the difficulty factor *D* for each question. I used a simple criterion for classification, resulting in 22 easy and 18 hard questions:

$$D = \begin{cases} \text{hard}, & \text{if } ave \geq 3 \\ \text{easy}, & \text{otherwise} \end{cases}$$

All the three evaluators suggested that some questions are too hard to solve within several minutes. This is taken into account as the significance of appropriate difficulty levels. In mental workload experiment, it is important that participants take the task seriously. Otherwise, despite of elaborated task, high level mental workload may not be triggered. An example of this is the arithmetic task taken by Stone and Wei [56]. In their experiment, they received almost the same results from 2 digits and 3 digits multiplication, which were supposed to be different. It is reasonable to suspect that hard questions may be too difficult and participants gave up at some points. Facing these questions, participants may lose their patience and pick a random option in few seconds. For some questions, evaluators can not reach an agreement on choosing standard answer. This impacts later analysis on performance. For those reasons, 10 easy and 10 hard questions were carefully selected for the final version of the questionnaire.

3.2 The PR Software

Computer based test is the most proper way to collect facial expressions. A software is needed for participants to finish the computer based questionnaire. The software was named PR Software and achieved the following functions:

- a User Interface (UI) that allows participants to answer questions
- a UI that allows participants to give ISA feedback
- a timer to record the time for synchronisation
- a proper way to store participants answer for further analysis

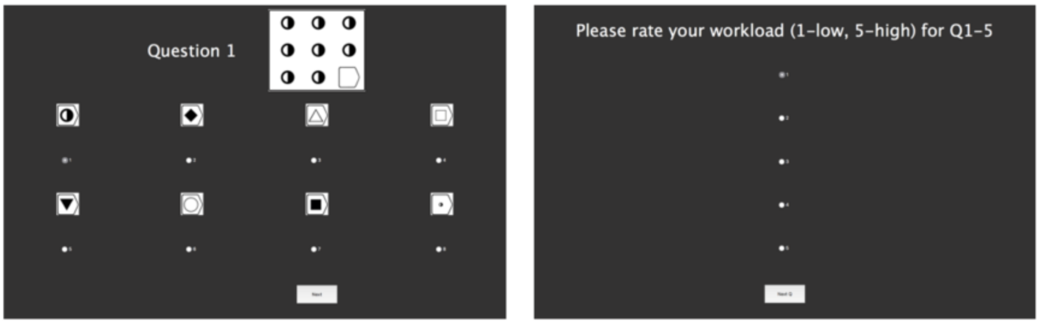
I evaluated some potential choices for programming. Psychology Experiment Building Language (PEBL)[6] is a framework for doing psychological experiments [46]. PEBL provides some easy-to-use features, claiming that 70 behavioural tests can be used and modified. There are some disadvantages for choosing PEBL. The understanding of PEBL C++ source code could be a waste of time. Besides, I tested two available versions of PEBL. The stable version is very old and the UI is not friendly, while the beta version is still being developed and not stable.

I chose Python as the programming language to program the software. The syntax of Python is concise, and I personally had several experiences on it. Python Experiment-Programming

[5] https://iq-research.info/en

[6] http://pebl.sourceforge.net/

Library (PyEPL)[7] is another framework for behavioural experiments written in Python. This framework helps to render 3D environment cross platform [20]. I finally decided not to use existing framework. To program based on my own need saves plenty of time for understanding the code and provides me with more flexibility to achieve extra functions.



Please rate your workload (1-low, 5-high) for Q1-5

Figure 4: Software: screenshot of the question and ISA rate interface

For Python GUI, Tkinter[8] is the standard package packed in Python. This means no other installation is needed to use Tkinter. But for the drawbacks, Tkinter is an old framework and has not been updated for many years. On the contrary, PyQt[9], the Python interpreted framework of Qt toolkit, brings network sockets, SQL databases and various of other features together. The framework is coming to the fifth version and is currently supported by the Qt company. This became my choice of GUI framework.

Implementation Details

In this part, I introduce more coding details of the PR Software. Readers can feel free to skip this part if not interested in these details. It would be very trivial to describe every piece in the dissertation. Instead, for each *.py* file, I first give an overview of its functions, then suggest some coding snippets that I consider essential to understand the code.

main.py

The *main.py* file contains the code to define the base window in PyQt. The base window has a Stack structure to control different pages. In PyQt the window is composed of many widgets.

```python
# add the Stack (widgets list) to the base window
self.widgetList = QStackedWidget()
self.setCentralWidget(self.widgetList)

# to adjust the resolution for different screens, the size was obtained by:
h = app.desktop().screenGeometry().height()
w = app.desktop().screenGeometry().width()
```

Code 1: Software, main.py

[7] http://pyepl.sourceforge.net/

[8] https://wiki.python.org/moin/TkInter

[9] https://www.riverbankcomputing.com/software/pyqt/download5/

welcome.py & finished.py

The *welcome.py* is the welcome/finish page of the software.

```python
# to define a Button and add it to the layout
# after clicked, it calls 'exampleQuestion'
btn = QPushButton("See Example Question")
btn.clicked.connect(self.exampleQuestion)
mainLayout.addWidget(btn, 6, Qt.AlignCenter)

# timer to count a period
# every second (1000 ms) it calls 'updateLabel' to decide whether to stop
self.timer = QTimer(self)
self.timer.timeout.connect(self.updateLabel)
self.timer.start(1000)
```

Code 2: Software, welcome.py

question.py

The *question.py* file contains the page for participants to answer and rate the questions, and relax page between each question set.

```python
# add the image source (question) to the window
q = QLabel(self)
pixmap = QPixmap('questions/q/img_q' + str(self.currentSelectedIndex) + '.png')
q.setPixmap(pixmap)
self.mainLayout.addWidget(q, 0, 3, Qt.AlignCenter)
```

Code 3: Software, question.py

result.py

The *result.py* file is used to store all the information and write it into a *.csv* file. This includes: answer to each question, the time used for answering and rating each question.

```python
# write answers to .csv file
with open('result.csv', 'w', newline='') as f:
  writer = csv.writer(f)
  for i in range(0, 20):
    writer.writerow([i+1, self.answers[i]])
```

Code 4: Software, result.py

3.3 Experiment Protocol

According to the schema of multimodal data collection proposed by Jaimes and Sebe [29], five essential factors shall be underlined.

- The study aims for *spontaneous* expression, so that we do not ask participants to perform any expression deliberately.
- The experiment will be in *lab setting*, not in the wild.
- We tend to measure mental workload, which is an *internal feeling*.
- Participants are fully awarded of the experiment in advance, therefore it is *open recording*.
- The *purpose* is to understand facial expressions under different levels of workload.

Following the aim of this project, the collected data should be used to:

1. **Verify** that the experiment has caused distinct levels of mental workload;
2. **Compare** facial expressions under different levels of mental workload.

For step 1, a proper scheme is needed for mental workload evaluation. Following the suggestions by Cain [7], I used the combination of three different measures. As described above, participants performance is chosen as the *primary measure*. We expect the harder questions to be answered in longer time, so spent time is regarded as the *secondary measure*. For *subjective measure*, ISA task is regarded better than NASA-TLX in order to keep the measure concise and focus on different time periods. Finally, among all the available devices at the stage, the OctaMon fNIRS brain scanner is the best scanner for *physiological measure*. It will be used to record the blood oxygen levels in the pre-frontal cortex.

However, even though the OctaMon is much portable than previous devices, it will still cover the forehead of participants. In this way, face becomes incomplete, which causes significant decline of OpenFace performance because the analysis depends heavily on facial landmark detection. For some participants, many landmarks nearing eyebrows can not be detection. This makes all the subsequent analysis unreliable.

To solve this problem, the dataset was divided into two part. The first part includes fNIRS measure, aiming to achieve step 1. In the second part, participants took the same experiment without wearing the OctaMon, and complete face will be captured for step 2. Besides, time is another variable. For half of the participants, a time limit is required to explore if it can cause higher mental workload compared with unlimited condition. Table 1 summarise the four different study conditions in the project. Subjects for different study conditions was equally collected.

Group Index	fNIRS Record	Time Limit
1	Yes	Unlimited
2	Yes	40s/question
3	No	Unlimited
4	No	40s/question

Table 1: Four different conditions in the study

Facial Data: Moment or Period

The initial plan was to select a static facial image from the video based on the result of fNIRS data. In other word, the image with the peak fNIRS measured value is regarded as the highest workload, and vice versa. In this way, we focus on the *moment* of facial expression. However, there are three defects. The first defect has been demonstrated: wearing fNIRS device, face is incomplete and it will drop the following analysis. Second, there is usually a physiological delay between measured blood and actual brain activity in an indeterminable short period. This delay makes it impossible to capture the accurate moment. The last defect is related to subjective measure. After rating a question using ISA, a 90-120 seconds break time is needed for fNIRS data back to the baseline. For some easy questions, participants only need few seconds to answer. If we measure the peak moment of each question, the break time would take most of the time and make the experiment tedious.

It turned out that for facial data, the analysis of a time *period* is much reasonable than a moment. I finally divided the 20 selected questions into four sets. There are labelled as 2 easy question sets and 2 hard, each consisting of 5 questions. The participants will be asked to rate their general feeling after each question set in ISA. For the time duration of rest time, we chose 90 seconds for the group with fNIRS, and 15 seconds for the other. Before the formal questions, participants were shown an example question to make sure they understand how to finish the questionnaire.

Recording Devices and Software

For the camera information, facial data is recorded using Logitech HD Webcam C525. The webcam is able to provide 720P video under 15fps. To use the webcam, Logitech provide a supported Logitech Webcam Software (LWS)[10] running on Windows system. In practice, another laptop XPS 13 was used to run the LWS.

For fNIRS collection, the OctaMon producer (Artinis Medical Systems, Zetten, The Netherlands) provides an independent HP laptop with the OxySoft software on Windows 10 operating system. In total, 6 transmitters transmit two wavelengths at 839 and 751nm. The HbO2 and HbR data was recorded in 8 different channels and the frequency is set to 10 Hz.

The PR Software is run on a MacBook with OS X El Capitan (version 10.11). In order to separate participants and the researcher, an extended monitor with mouse and keyboard was connected to the MacBook. Then the screen was mirrored in the extended monitor. This allows the researcher to keep watch on participants reaction during the experiment without disturbing them.

Experiment Timeline

The use of different devices leads to another issue: how to synchronise time among different devices. A scheme was developed to cope with it. Figure 5 illustrates the timeline of the whole experiment. The orange time points need to be recorded by an independent timer. For the time period separated by the thin green lines (navy arrows), the PR Software automatically records each node. In this way, with only few recording points, one can easily calculate the time when participants are doing the questions, but not in rating or relax time. Note that for half of the data without fNIRS recording, the first and the last orange points are not needed.

[10] `http://support.logitech.com/en_us/product/hd-webcam-c525/downloads#`

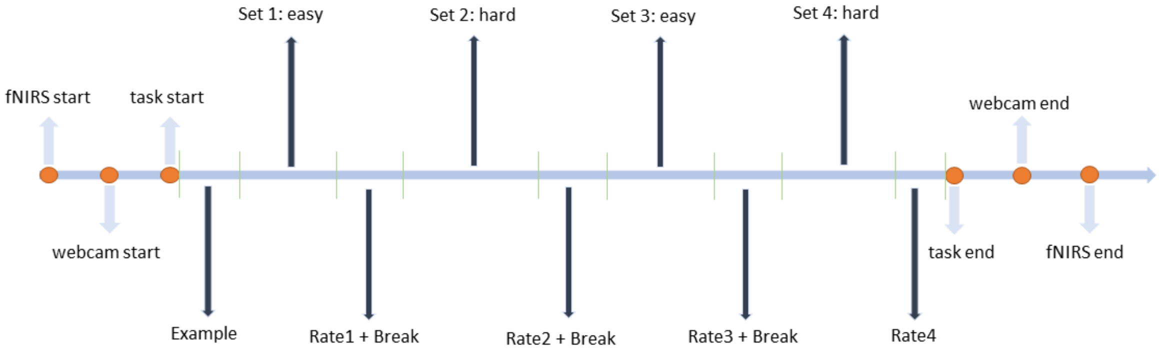


Figure 5: Time synchronization: timeline of the experiment

3.4 The Dataset

Data Collection

In total, 21 subjects were invited to join the experiment at the Mixed Reality Lab, with the average age 21.95 (ranging from 20 to 25). In order to make participants treat the experiment positively and seriously, everyone received a 5 subsidy after they finish the experiment. Participants were first asked to read the information sheet and sign the consent form. One subject decided to withdraw after reading through the information sheet. For each of the four different conditions, 5 subjects were collected. In total, 20 subjects were collected. Practically, the experiment with fNIRS recording lasted around one hour, and the experiment without fNIRS recording lasted around 30 minutes. This includes the time for reading the information sheet and signing the consent form.

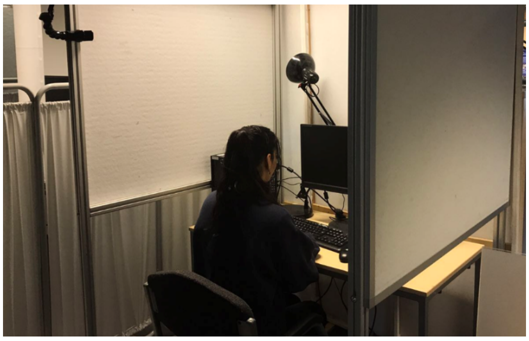


Figure 6: The data collection place at the Mixed Reality Lab

Data Pre-processing

Following the timeline as illustrated in Figure 5, raw video files and fNIRS data were collected. However, the data contains redundant segments and noises. Data pre-processing procedures are required to remove them.

<u>Video Files</u>

The raw videos contains auditory noises, which are meaningless and irrelevant to this project.

The audio/video data processing tool FFmpeg[11] was used to remove them by:

```
ffmpeg -i "input.wmv" -vcodec copy -an "output.wmv"
```

Then, the scheme is to follow the timeline which were introduced in Figure 5. By comparing and calculating the recorded time points, videos corresponding to the 4 question sets were extracted. For video clipping, the python package MoviePy[12] was used. Comparing with other video editing software, MoviePy uses the API from FFmpeg tool and do not perform any format conversion. Therefore, it is significantly faster. Below is the code to trim the video. It follows the timeline in Figure 5.

```python
from moviepy.video.io.ffmpeg_tools import ffmpeg_extract_subclip
startDelay = 147
timeBreak = 90

now = startDelay + timeBreak
till = now + 79.0 # time for answering the question set e1
ffmpeg_extract_subclip("1.wmv", now, till, targetname="1_e1.wmv")

now = till + 12.4 + timeBreak # time to rate and relax
till = now + 203.1 # time for answering the question set h1
ffmpeg_extract_subclip("1.wmv", now, till, targetname="1_h1.wmv")

now = till + 7.2 + timeBreak # time to rate and relax
till = now + 89.4 # time for answering the question set e2
ffmpeg_extract_subclip("1.wmv", now, till, targetname="1_e2.wmv")

now = till + 4.3 + timeBreak # time to rate and relax
till = now + 274.2 # time for answering the question set h2
ffmpeg_extract_subclip("1.wmv", now, till, targetname="1_h2.wmv")

finish = till + 3.6 # time to rate, then finish
```

Code 5: Python code to clip videos

fNIRS Data

The fNIRS brain scanner OxtaMon is produced and delivered together with its software OxySoft. Data is collected and stored in *.oxy3* format. OxySoft provides Excel format for exporting the data. However, exporting data from it is complex and many parameters need to be selected. Besides, the OxySoft has continual bugs with reading and processing the data. Practically, I had the problem of accessing part of the collected data using OxySoft. The Matlab brain data analysis toolbox FieldTrip[13] is introduced to overcome the problem. It can be quickly set up by:

```matlab
addpath PATH/TO/FIELDTRIP
ft_defaults
header = ft_read_header('file.oxy3')
data = ft_read_data('file.oxy3')
```

[11] https://www.ffmpeg.org/

[12] https://zulko.github.io/moviepy/

[13] http://www.fieldtriptoolbox.org/

Code 6: Set up FieldTrip in Matlab

For the two functions, *header* is a Matlab structure containing recording information and *data* returns a N(channels) x M(frames) matrix of the raw data (light intensity). Besides, the structure file *optodetemplates.xml* is needed. For more details of the documentation readers are referred to the FieldTrip official website.

Our next step was to convert the light intensity to the changes of HbO2 and HbR concentration. This could be deduced from the modified Beer-Lambert law by first calculating the value of optical density [58]. The distance between source and detector is 3.5cm [1]. For the extinction coefficients of HbO2 and HbR, we followed the table proposed by Cope [9] and show the value in Table 2. The DPF value was calculated by the formula proposed by Duncan et al. in 1996 [1] [12]:

$$DPF = 4.99 + 0.067 \times Age^{0.814}$$

Practically, I used the SPM-fNIRS toolbox [14], a Matlab analysis of fNIRS signals to generate matrices of HbO2 and HbR concentrations [57]. The function *spm_fnirs_read_artinis* in SPM-fNIRS converts the *.oxy3* file to the SMP-fNIRS required data format.

Wav(nm)	HbO2(mM^{-1}cm^{-1})	HbR(mM^{-1}cm^{-1})
839	1.1018	0.7812
751	0.5554	1.5703

Table 2: The extinction coefficients of HbO2 and HbR in this project

Then, the toolbox provides a UI for users to specify parameters and perform the conversion. After this step, similar way was applied to clip and extract the fNIRS data. Note that a 31-seconds interval during the rest time between the second and third question set was chosen as the *baseline*. Details of the steps can be found in the user manual [62] and the resulting files structure will be further described in appendices.

I was aware of the physiological delay between the measured data and actual brain activities. Besides, the error because of synchronising time between different devices and the run time delay of some code functions were also taken into account. These factors makes it impossible to estimate the delay time. Finally, I chose to keep the data as original, and not to adjust the time delay.

The Dataset

20 subjects took the experiment in 4 different conditions and the index table is shown in Table 3.

At the end of the project, for each subject the dataset has:

- original and clipped facial video without audio input;
- questionnaire answers, answering and rating time;
- fNIRS data with light intensity and haemoglobin concentration values (group 1, 2 only);

[14] https://www.nitrc.org/projects/spm_fnirs/

Group	Participants Index	fNIRS Record	Time Limit
1	1, 2, 3, 7, 20	Yes	Unlimited
2	4, 5, 6, 18, 19	Yes	40s/question
3	10, 11, 12, 14, 15	No	Unlimited
4	8, 9, 13, 16, 17	No	40s/question

Table 3: Participants index on different study conditions

The dataset was then analysed and the details are described in section 4.

4 Data Analysis

In this section I report the analysis procedures based on the dataset. I divided it into two steps, as guided in subsection 3.3. First, the distinction of mental workload under different conditions was verified. Then facial data was processed and analysed by OpenFace. The dataset alone with all the analysis files is clearly structured and the file description is provided at the end of this section.

4.1 Mental Workload

Participants were asked to finish the questionnaire with 4 question sets. It was designed as easy-hard-easy-hard. In this part, the objective is to verify that the actual mental workload is aligned with what is expected. We define *e1* as the first easy question set, and *h1* as the first hard question set, then similarly *e2* and *h2*. The result of traditional measure methods is first discussed. For fNIRS measure, we first performed noise removal and feature extraction. Then, different machine learning techniques were applied for binary classification of high or low mental workload.

Traditional Measurement

I tried to establish the evaluation of mental workload for each question set by building 4 factors: difficulty factor, subjective factor, primary factor and secondary factor. Together, these factors were used to give the general evaluation of the experiment design. For all the factors original values reported in Table 4, I calculated the arithmetic mean of all the 20 subjects.

Difficulty factor derives from the marking of the three independent evaluators. *Subjective factor* is the self reported mental workload from participants. *Primary factor* is the accuracy of the question sets, which reflects participants performance. *Secondary factor* is the average time participants used to finish the question sets. The subjective, primary and secondary factor correspond to the subjective, primary and secondary measure. It is expected that the subjective and secondary measure positively correlate with the difficulty, and the primary measure negatively correlate with it. This means that participants are expected to take longer time, and report higher workload, but perform poorer on more difficult question set. In order

Question Set	Difficulty(1-5)	Subjective(1-5)	Primary(%)	Secondary(s)
e1	1.60	1.25	97	72.40
h1	3.27	3.35	60	173.42
e2	1.73	1.60	89	87.13
h2	3.20	3.50	59	195.93

Table 4: Original values of mental workload on various traditional measures

to compare the four factors, we normalised the difficulty, subjective and secondary factor using:

$$Fd_n, Fsub_n, Fsec_n = \frac{F_n}{F_1}$$

And normalised the primary factor using:

$$Fp_n = \frac{F_1}{F_n}$$

Where $\mathrm{F_n}$ is the factor value of the n^{th} question set. This provides an overview of the three measures on each question set, which is illustrated in Figure 7. We can conclude that the general feedback on the subjective, primary and secondary measure is in line with the expectation.

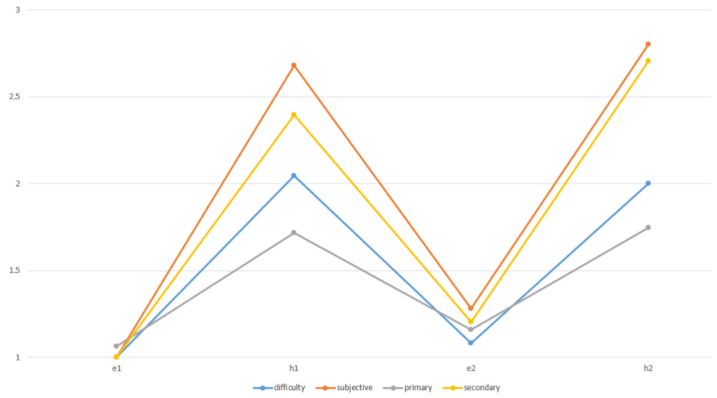


Figure 7: The chart of normalised values of mental workload from Table 4

Time Limit Comparison

In our study, we used different time limit for study conditions. For half of the participant, there is no time limit to finish the questionnaire. While for the other half, each question need to be finished within 40 seconds. Although 40 seconds period is considered enough to finish most of the questions, we designed this condition to see if it leads to higher mental workload.

Figure 8 illustrates the result of participants performance and self reported mental workload. It is clear that participants rate their mental workload higher and suffer a drop in primary task performance in the same questionnaire.

fNIRS Measurement

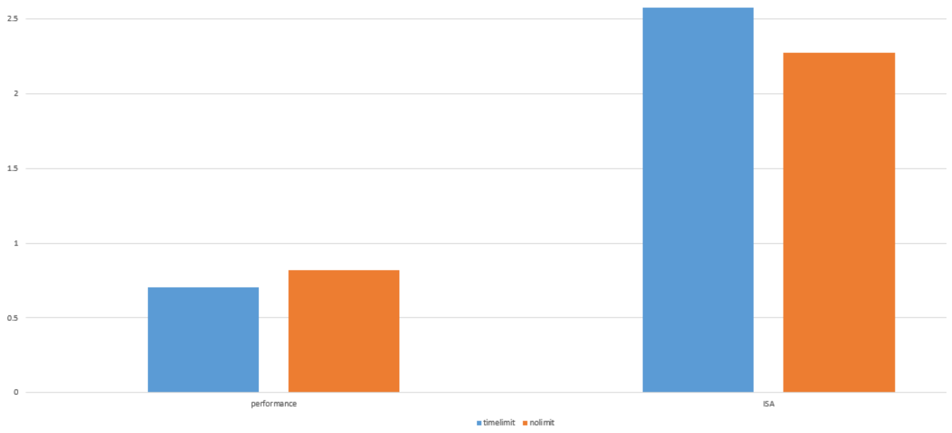


Figure 8: The comparison of the result under different time limit conditions

We performed a sequence of processing steps to measure mental workload using the fNIRS data. Then, for machine learning, we described a binary classification problem and tried to predict mental workload on high or low level.

Noise Removal

The analysis of fNIRS data was based on the 10 subjects of group 1 and 2 as indexed in Table 1. After pre-processing, noises need to be removed from the fNIRS signals. The most common noises in fNIRS is high frequency noises, usually caused by the instrument or the surrounding environment [47]. The low-pass filter is designed to cut off the signals above 0.6 Hz frequency. In FieldTrip, we used the function:

```
output_NxM = ft_preproc_lowpassfilter(input_NxM, sample_frequency, cut_frequency)
```

To further smooth the data, we used the Savitzky Golay filter with the polynomial order 8 and kept 80% frames. The function is provided by Matlab:

```
output_MxN = sgolayfilt(input_MxN, order, frame_length)
```

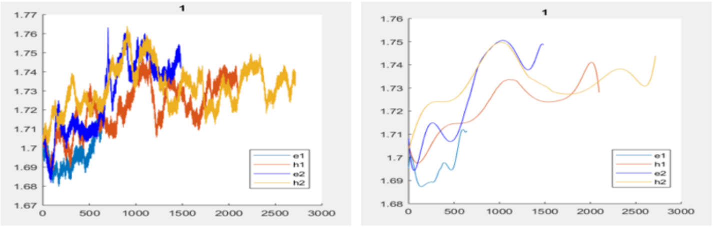


Figure 9: Example: the comparison of fNIRS data before and after noise removal

Feature Extraction

With the increase of mental workload, the value of HbO2 tends to increase, and the value of HbR tends to decrease. This has been confirmed by Herff et al. in the N-back tasks [26]. There are several options for choosing a proper feature, such as HbO2, total hemoglobin (HbO2 + HbR) and exchange hemoglobin (HbO2 - HbR). Nasser and Hong studied several previous experiments and concluded that HbO2 is more robust and suitable for task related brain activities [47]. Koenraadt et al. pointed out the importance of reducing the amplitude differences among participants [37]. Their method is to normalise the differences by dividing the mean value. In this project, we normalised the data using the rest time as *baseline* and define the feature as:

$$Feature = HbO2 - mean(HbO2_{baseline})$$

Then, the appropriate heuristic method should be decided. Common attributes used in fNIRS analysis are peak, mean, variance, slope, skew level and kurtosis [47]. We selected the most frequently used attribute mean value for discrimination and calculated the arithmetic mean for all the signals in each of the 8 recording channels. The code snippet to remove noises and extract features is shown below.

```matlab
function result = extractFeatures(temp)
  % lowpass filter
  highp = 0.6;
  lpe1 = ft_preproc_lowpassfilter(temp.e1_hbo', 10, highp);
  lph1 = ft_preproc_lowpassfilter(temp.h1_hbo', 10, highp);
  lpe2 = ft_preproc_lowpassfilter(temp.e2_hbo', 10, highp);
  lph2 = ft_preproc_lowpassfilter(temp.h2_hbo', 10, highp);
  lprs = ft_preproc_lowpassfilter(temp.rest_hbo', 10, highp);

  % sgolay filter, for the third parameter it must be an odd number
  order = 8;
  ne1 = sgolayfilt(lpe1', order, 2*floor(size(lpe1,2)*0.8/2)+1);
  nh1 = sgolayfilt(lph1', order, 2*floor(size(lph1,2)*0.8/2)+1);
  ne2 = sgolayfilt(lpe2', order, 2*floor(size(lpe2,2)*0.8/2)+1);
  nh2 = sgolayfilt(lph2', order, 2*floor(size(lph2,2)*0.8/2)+1);
  nrs = sgolayfilt(lph2', order, 2*floor(size(lprs,2)*0.8/2)+1);

  % transpose matrices
  sge1 = ne1';
  sgh1 = nh1';
  sge2 = ne2';
  sgh2 = nh2';
  sgrs = nrs';

  % F = HbO2 - mean(HbO2_baseline)
  for i = 1:8
    te1(i,:) = sge1(i,:) - mean(sgrs(i,:));
    th1(i,:) = sgh1(i,:) - mean(sgrs(i,:));
    te2(i,:) = sge2(i,:) - mean(sgrs(i,:));
    th2(i,:) = sgh2(i,:) - mean(sgrs(i,:));
  end

  % extract the mean value of F as the selected feature
  % from two question sets which raise the most distinction
  for k = 1:8
    features_m(k,:) = [mean(te1(k,:)) mean(th2(k,:))];
  end

  % create labels and save the result
  label = [0 1];
  result.data = features_m'*1000;
  result.label = label';
end
```

Code 7: Matlab code to process fNIRS data

Classfication

For each subject, we chose two question sets which lead to the most discrepancy according to the ISA rating. The group with low mental workload was annotated as 0 as the ground truth label, and the one with high mental workload was annotated as 1. In the view of machine learning, this leads to a 8 dimensions binary classification problem. Counting all the 10 subjects, we had:

$$points : [20 \times 8_{double}]$$
$$labels : [20 \times 1_{binary}]$$

Many machine learning techniques can be applied for classification. For fNIRS analysis, LDA, ANN and SVM are the three most commonly used methods. A review of the studies on fNIRS can be found in [47]. In this project, we evaluated the results using ANN and SVM. We used a simple 1-hidden-layer ANN with 10 nodes. For SVM model, we used the one with polynomial kernel.

Because that we only selected 20 samples, 10-fold cross validation is used for evaluation to make use of all the data. To overcome the problem of overfitting, on each of the 10 training sets we used 3-fold cross validation for ANN. For SVM, the same way was used in order to choose the best parameters for training the best model. F1 score is the best known accuracy measure to report classification result. It is defined as:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

We reported the results of F1 score on 10-fold cross validation in both the two methods in Table 5. The best SVM model with polynomial kernel achieved 73.33%. Below is the code snippet to perform 10-k validation using SVM.

```
% random partition of the data
cvp = cvpartition(size(points, 1), 'KFold', 10);
croval_result.rs = cell(10, 1);

% cross validation
for i = 1:cvp.NumTestSets
  % specify unique validation index
  trIdx = find(cvp.training(i));
  teIdx = find(cvp.test(i));

  % find the best hyper parameters for training using 3-fold validation
  best_hp = cross_val_part_svm_bin(points(trIdx, :), labels(trIdx, :), parameters);

  % train a SVM using the best hyper parameters
  model = fitcsvm(points(trIdx, :), labels(trIdx, :), ...
    'KernelFunction', parameters.names{1}, ...
    'BoxConstraint', best_hp.C, ...
    parameters.names{2}, best_hp.V);
  % prediction using the SVM
```

```
    pr = predict(model, points(teIdx, :));

    % input (targets, predicts) for calculating f1 score
    croval_result.rs{i}.f1 = cal_f1_svm_bin(labels(teIdx, :), pr);
```

Code 8: Matlab code to implement SVM and 10-fold validation

Note that we did not omit any bad channels in all steps. Considering that it is not possible to remove all noises and comparing with the similar experiment results in [48], we concluded that fNIRS can be used for mental workload classification on pattern recognition tasks. And we were confident to analyse facial expression using the remaining half of the dataset.

Method	Kernel	$\mathrm{F1}_{\mathrm{ave}}$(%)
ANN	NA	64.44
SVM	rbf	63.33
SVM	polynomial	73.33

Table 5: F1 score of fNIRS classification using different methods

4.2 Facial Expression

The raw videos were collected in *.wmv* format and were clipped corresponding to each question set. We used OpenFace to process the videos and generate the report. OpenFace was installed and configured on a MacBook with OS X El Capitan. The configuration procedure may contain various of issues due to the incompatibility between OpenFace and its dependent librarys. I posted my summary in OpenFace project Issues[15].

In command line, videos can be processed using:

```
./FeatureExtraction -f "videos.wmv"
```

This will produce a *.csv* file containing facial landmark, head pose, and AU analysis. Then, the AU part in the file was extracted to Matlab using the *csvread* function. We used the same scheme to choose 2 question sets out of 4, i.e. to cause the most discrepancy according to the ISA rating. We regarded the one with easy question set as low mental workload, and vice versa. For the 10 subjects, we selected 10 videos of low and 10 of high mental workload.

Ratio Measure

OpenFace analyses the AU intensity and presence frame by frame. For intensity, it yields a value between 0 and 5. For presence, it is a binary classification problem. For AUs on each subject, we defined the *value per frame*, i.e. the sum value of a AU intensity or presence divides by the number of frames within the video.

[15] https://github.com/TadasBaltrusaitis/OpenFace/issues/340

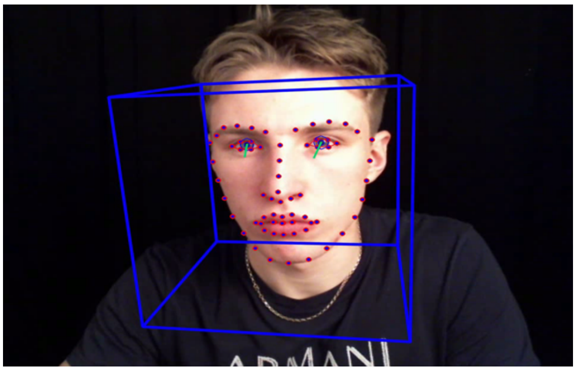


Figure 10: Screenshot: an example of OpenFace analysed video

To compare the AUs between different levels of mental workload, we defined:

$$ratio = \frac{\sum_{n=1}^{S} Re_n}{\sum_{n=1}^{S} Rh_n}$$

Where S is the total number of subjects, which is 10, and Re_n, Rh_n are the *value per frame* of the low, high workload video on the n^th^ subjects. The result for intensity and presence analysis is shown in Table 6 and Table 7. Bigger value indicates that the AU is more likely to appear in low workload state.

AU	**01**	**02**	**04**	**05**	**06**	**07**	**09**	**10**
ratio	0.8489	0.8105	0.8840	1.2390	1.0512	0.9019	0.7784	1.1215
AU	**12**	**14**	**15**	**17**	**20**	**23**	**25**	**26**
ratio	1.2514	1.1745	0.7696	0.9213	0.9157	0.6672	1.1530	0.8192

Table 6: Ratio of low/high mental workload of average AU intensity value per frame

AU	**01**	**02**	**04**	**05**	**06**	**07**	**09**	**10**	
ratio	0.9172	0.9426	0.9061	1.0552	0.0459	1.0336	0.3764	0.6083	
AU	**12**	**14**	**15**	**17**	**20**	**23**	**25**	**26**	**28**
ratio	0.1307	1.4922	0.5987	0.9068	0.5993	0.9503	0.7138	0.5072	0.8789

Table 7: Ratio of low/high mental workload of average AU presence value per frame

Subject Count

We also compared the *value per frame* value of AUs within each subject. Then, for each AU we counted the greater or smaller values separately. For example, 06_{i} l:h = 3:7 means that for 7 subjects out of 10, the intensity *value per frame* of AU5 on high workload is greater than the value on low workload. In this example, we consider the intensity of AU9 on higher mental workload tends to be higher. In Table 8, we only reported the results with a discrepancy more than 4:6 (including 3:7, 2:8, 1:9 and 0:10).

$AU_{i/p}$	05_i	06_i	09_i	15_i	23_i	25_i	15_p	26_p	28_p
count(l:h)	9:1	3:7	3:7	2:8	2:8	7:3	2:8	2:8	3:7

Table 8: Subject count comparison on AUs *value per frame*

Blink Rate

Blink rate (AU 45) is more complex than others. Because a blink usually lasts several frames, we took extra actions to calculate its *value per frame*. OpenFace returns a binary array of AU 45 presence. We first converted the array to a string, such like:

```
string = 0000011111000000011111...
```

Then, we counted the patterns when blink start is detected using:

```
num = count(string,'01')
```

We used this *num* as the number of the occurrence and calculate its *value per frame* for each subject as shown in Table 9. No strong correlations were found with different levels of mental workload.

Condition	**1**	**2**	**3**	**4**	**5**
low(x10^{-3})	13.6674	27.2647	10.4408	25.8929	17.4263
high(x10^{-3})	9.2854	32.8400	16.1290	16.2369	11.8778
Condition	**6**	**7**	**8**	**9**	**10**
low(x10^{-3})	15.8893	11.2245	20.6084	18.2440	14.8233
high(x10^{-3})	17.8326	25.4586	20.2753	24.1158	11.7146

Table 9: Blink *value per frame* on low and high mental workload

Below is the code snippet for face analysis.

```
% read the .csv file processed by OpenFace, and generate the matrices
Ematrix = csvread(easyF, 1, 678);
Hmatrix = csvread(hardF, 1, 678);

% calculate the value per frame for each AU
for i=1:35
  result{j-7}.easy_v(i) = sum(Ematrix(:, i));
  result{j-7}.hard_v(i) = sum(Hmatrix(:, i));
end
result{j-7}.easy_r = result{j-7}.easy_v / size(Ematrix, 1);
result{j-7}.hard_r = result{j-7}.hard_v / size(Hmatrix, 1);

% calculate blink rate
stringE = num2str(Ematrix(:, 35)');
stringH = num2str(Hmatrix(:, 35)');
stringE = stringE(find(~isspace(stringE)));
```

```
stringH = stringH(find(~isspace(stringH)));
blinkR(j-7, 1) = count(stringE,'01') / size(Ematrix, 1);
blinkR(j-7, 2) = count(stringH,'01') / size(Hmatrix, 1);


for k=1:35
  % subject count
  if result{j-7}.easy_r(k) >= result{j-7}.hard_r(k)
    compareM(k,1) = compareM(k,1) + 1;
  else
    compareM(k,2) = compareM(k,2) + 1;
  end

  % calculate ratio
  normalise(j-7, k) = result{j-7}.easy_r(k) / result{j-7}.hard_r(k);
end
```

Code 9: Matlab code to analyse facial data

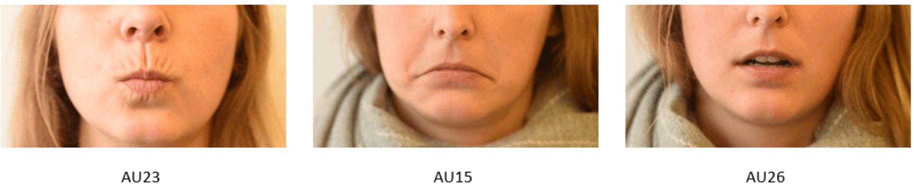


Figure 11: High mental workload: AU23 stronger, and more AU15, AU26 [28]

The result of AUs ratio on low/high workload was calculated and reported in Table 6 and Table 7. If we take 30% as a distinction line, in higher mental workload the intensity of AU23 is stronger and AU9, AU10, AU12, AU15, AU20 and AU26 is more likely to present, while AU14 is less likely. To agree with high mental workload H, we put it together with the subject count in Table 8 and set the criteria to:

$$H = ratio \leq 0.7\ AND\ count \leq \frac{3}{7}$$

We used it to conclude that for higher mental workload, the intensity of AU23 is stronger, and the presence of AU15 an AU26 is more frequent.

Figure 12 is illustrated to sum up the analysis procedures in this project. To verify workload levels triggered by the experiment, participants performance, answering time, and ISA rating were used as primary, secondary and subjective measures. We processed fNIRS data and built train a SVM model using these signals to predict mental workload and further confirm the workload. Then, the *ratio* and *subject count* information was used to find the 3 AUs linkage with higher mental workload. Besides, blink rate was calculated and the evidences show no strong correlations. These findings provide inspiration for future research.

4.3 File Structure

This part is the description of file structure. For each subject, there is an independent folder. In the *n* folder:

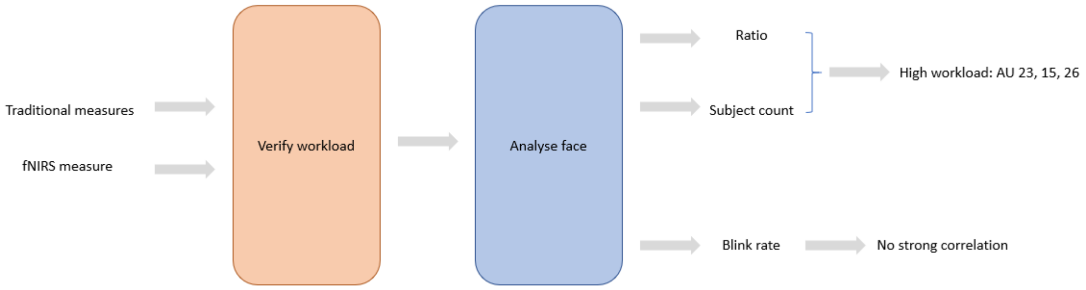


Figure 12: Summary of data analysis in this project

- *n.wmv*: the facial video without audio input;
- *n_e1 n_e2 n_h1 n_h2*: the clipped videos of the first/second (e)asy/(h)ard question set;
- *FPn.oxy3*: the original fNIRS data;
- *FPn_XXX_XXX.csv*: includes participants performance, generated by the PR Software;
- *NIRS.mat*: processed Matlab file after haemoglobin concentration transformation;
- *trim.py*: used to clip the videos;
- *readFdatan.m*: used to clip fNIRS data;
- *\processed*: contains OpenFace processed files.

Besides, the *\files\software* folder includes all the code of the PR Software. In *\files\fNIRS* there are functions to process the fNIRS data and apply machine learning. The features for machine learning are stored in *fnirsinput.mat* and the SVM model is in *svm.mat*.

In addition, inside *\files analysisface.m* is the code for facial data analysis, and *result.xlsx* includes all the tables for mental workload evaluation.

5 Discussion

In this section, I first discuss and interpret the result of mental workload evaluation and facial expression analysis. Then, I evaluate this project in both advantages and disadvantages, and suggest some future work. Finally, in subsection 5.3 I summarise my personal motivation, project management and how I achieved the aim and objectives. A self reflection ends the dissertation.

5.1 Result Interpretation

The result of traditional measures is shown in Table 4 and is normalised in Figure 7. We start the discussion from *difficulty*. The difficulty for question sets was agreed by three independent evaluators and is regarded as the expectation. From Figure 7 it is clear that all the three traditional measures match the trend, i.e. low-high-low-high mental workload. More wild fluctuation in Figure 7 reflects this trend more intensely. From this point of view, self reported ISA rating is the most reliable measure and therefore was used as the ground truth label in training and evaluating fNIRS data.

For fNIRS measure, researchers focus on its comparison with the demand and self reported mental workload level. Maior et al. summarised that their fNIRS measure reflected better on demand than self reported measure [41]. My finding is that for some participants, fNIRS data is more matching with ISA measure instead of expected demand. As shown in Figure 13, the ISA rating for the four question sets is [1, 2, 2, 3] while the demand is around [1, 3, 1, 3]. It is clear that fNIRS measure is more accurate, which further confirmed the advantage of fNIRS to measure mental workload.

For the binary classification, for each subject the question sets with highest and lowest ISA rating were chosen out of 4 question sets. Then, for feature extraction we observed that HbO2 show more correlation with our low-high-low-high trend. This confirms the assertion by Nasser and Hong that HbO2 is more robust than HbR [47]. Mean value of signals is the most frequently used attribute to extract features in fNIRS machine learning, and became the choice in this project. Despite that we applied simple machine learning techniques, our best SVM model reached F1 score 73.33%.

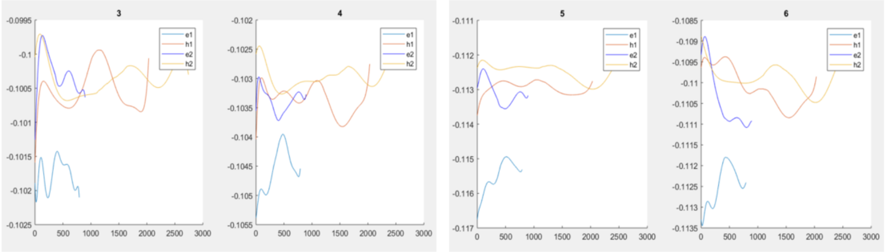


Figure 13: Selected channels of fNIRS HbO2 smoothed data on a subject

We used two methods to evaluate facial expressions, *ratio* and *subject count*. Ratio reflects the average level of intensity or appearance of all collected data (image frames), while subject count emphasise the number of subjects agree with a trend.

Grafsgaard et al. suggested 3 AUs linkage with negative affective states. AU2 and AU4 were associated with frustration and confusion, and AU1 showed a sign of sadness or surprise [23]. In this project, for the ratio analysis in Table 6 and Table 7, all the three AUs shown more intensity and appearance under higher mental workload. Considering that frustration, confusion, sadness and surprise is intuitively linked with high workload, the result confirms their findings to some extent. However, in the subject count analysis their was no significant difference. More samples are needed to verify the findings.

The two methods were used together to conclude the intensity of AU23, the appearance of AU15 and AU26 in relation with higher mental workload. It happens that AU15 and AU23 are also regarded as negative feedback for mental states. This brings a sequence of interesting questions: are these facial AUs directly related to higher mental workload? Or because that higher mental workload raise negative emotions, then the emotions cause the appearance of the AUs? It seems that these questions are beyond the scope of Computer Science. What could be asserted here is that future research on the combination of mental workload and facial expression should pay special attention on these AUs.

5.2 Project Appraisal

Achievements

In general, I had a loose and flexible plan, yet the project reached a satisfying level at the end. The aim was reached and the 8 objectives listed in subsection 1.2 were all achieved.

1. Design an experiment for data collection

This includes the questionnaire design and the experiment protocol. It is thoroughly detailed in subsection 3.1 and subsection 3.3. The Protocol was proved to be a success in subsection 4.1 with limitation presented in subsection 5.2.

2. Present the information sheet, consent form and ethic checklist

The information sheet, consent form and ethic checklist was approved by the ethic committee in the School of Computer Science at the University of Nottingham. The files are attached in appendices.

3. Implement a computer based software for the experiment

The PR Software was developed using Python and PyQt5. It was used to collect data from 20 subjects and was running without issues in practice. The details of the software is described in subsection 3.2.

4. Recruit participants and conduct the experiment

In total, 20 participants took part in the experiment and agreed with the use of their data. The collection was conducted at the Mixed Reality Lab in 5 weeks. More information can be found in subsection 3.4.

5. Apply suitable scenario to synchronise time for fNIRS data, videos and tasks

A timeline was used to synchronise the time between different devices. The scheme was effective in practice. However, the delay issue of fNIRS data described in subsection 3.4 is unsettled.

6. Pre-process raw fNIRS data and videos

Both the fNIRS and facial data were properly processed before analysis. The procedures and coding snippets are shown in subsection 3.4.

7. Apply machine learning techniques to fNIRS data for mental workload prediction

For fNIRS machine learning, SVM and ANN method was used to construct a binary classification of low and high mental workload. The SVM method achieved the highest F1 score. Details can be found in subsection 4.1.

8. Analyse the linkage between facial AUs appearance, intensity and mental workload

Two methods were used to analyse the facial data processed by OpenFace and blink rate was also calculated. The result is presented in subsection 4.2.

Contributions

The contributions of this project can be summed up to three points. First, fNIRS measure can be deemed as an appropriate method to measure mental workload on pattern recognition task. Second, an attempt was made to build the raw version of a face-workload dataset, and the data will be available to future related studies. Third, facial data was processed by OpenFace and some useful clues on facial AUs under different levels of mental workload have been provided.

Limitations

There are nevertheless some limitations in this project. First, I am not be able to provide static images with high or low mental workload. This is discussed in subsection 3.3. The key is to develop a reliable scheme to overcome the delay between fNIRS and actual brain activity. Due to the limit of my knowledge, this could not be achieved within this project.

Second, although fNIRS binary classification received good F1 score, when trying to predict the actual mental workload level (from 1 to 5), the result dropped dramatically. This problem may be ameliorated by recruiting more participants and collecting more data. By the limits of time and finance, in this project only 20 participant were invited. Also, more features on fNIRS signals could be extracted and complex machine learning techniques such as deep learning could be applied.

Third, the shared dataset is not a real dataset due to the lack of manually annotated AU labels as ground truth values. The problem is that without ground truth labels, we can not do any machine learning attempt using facial data. In this project, all the facial data was processed by OpenFace, and it is not possible to evaluate the result without these labels. The accuracy of predicted AUs on OpenFace is not equal, therefore all the findings on facial data shall be only regarded as clues waiting for confirmation by future research.

Finally, as Zhang et al. pointed out, for spontaneous expressions one must keep a good balance between natural expressions and data quality [66]. In this experiment, pattern recognition problems was used as the stimulus to trigger different levels of mental workload and we paid little attention to the stimulus on facial expressions. The result is that for most of the recording time, participants showed little facial expressions and the AU appearance per frame is very low. The drawbacks of imbalanced facial data has been demonstrated by [30]. For the improvement of future face-workload experiment design, one should consider more carefully on the stimulus on both mental workload and emotional expressions, which usually lead to more AU appearance.

Future Work

The first extension to this project, as discussed above, should be adding manually annotated FACS labels. Manual labelling is extremely time-consuming, and the training process could be more than 200 hours, which makes it impossible to be finished within this project. With FACS labels in the future, the abovementioned finding on facial expressions could be verified. We can also use machine learning techniques to predict mental workload based on facial images.

This may or may not be a good idea. El Kaliouby et al. attempted to design a model to measure 6 mental states including *agreeing, concentrating, disagreeing, interested, thinking* *and unsure* based on purely facial expressions [17]. The classification rate of *thinking* received the lowest with only 40.1%. From the view of cognitive and affective states, a mental state would include multiple modalities such as body movement (video), voice (audio) and brain activity (physiological data). In some way, the combination of multiple modalities may become a breakthrough of the of mental states measurement and is worth to try in the future.

5.3 Reflection

Personal Motivation

Last summer, when I was doing a project to label the six basic facial emotions and facial AUs using deep artificial neural network, I started to think about a problem: we try to predict several emotions, dozens of facial AUs, what can we do other than that? For example, can we try to predict the level of contemplation? It is obvious that human mental state system is far more complex than what we are describing now. What directly came to my mind is the significance of improving teaching environment. With an ideal intelligent system, teachers can easily track the reflection of students and change teaching speed accordingly. At the time, I did not know the design of automated teaching systems have been concerned by researchers for years [64]. As time goes on, it turned out that some of the ideas I could achieve within my undergraduate stage, while some I may not. Anyhow, this rough thought became my initial idea of this project.

To implement this idea, I need a dataset which includes facial recordings with mental workload levels. Although there were several previous studies as mentioned in subsection 2.6, their collected data are nevertheless unavailable to other researchers. This leads to another self motivation in this project, which is to build a shareable dataset so that future researchers can make use of it.

Project Management

This is a research-oriented project. To boost my knowledge, I spent most of my time in reading papers. Then, conducting the experiment is time-consuming as well. The Ethics Checklist, Information Sheet and Consent Form are more complex than what I expected, and were updated several times. As discussed through the dissertation, the ideas of experiment design and analysis has always been changing and I was often inspired by new knowledge. Then, I spent more time on deciding to choose pattern recognition questions as the task of the experiment. Apart from that, I believe that the time was in line with expectations. The Gantt chart is provided in Figure 14 at appendices.

Self Reflection

The project is meaningful to myself in both research and programming way. To me, research started from almost zero. At the beginning, I had no ideas on how to quickly grasp points from a long paper, how to judge the quality of papers and how to structure the dissertation properly. For now, progress was made on these issues. I understood the meaning of h5-index and i10-index, and learnt the difference between a top conference paper and a bad paper. I attended the brain meeting every week to discuss fNIRS data analysis, and received many helps from the members in the group.

For programming, I also made progress. Although I had prior experience in Python, I learnt PyQt5 by myself and developed the PR Software. I wrote many Matlab and Python scripts to process and analyse the data. These programming skills could be beneficial to my future study. In addition, I mastered LaTex to produce long papers in beautiful format. This document was edited in LaTex. Apart from the basic requirements, this paper also provides the list of tables, figures and abbreviations. There are hyperlinks to every citation, section, table and figure in the main body. Readers can quickly jump to the part as needed. Also, the code snippets were coloured using *minted*. I believe that mastering LaTex will significantly improve my work efficiency in the future.

I would regard this project as a summary of my undergraduate course. The most important thing I learnt in the past three years is the ability to solve new problems. I have experienced many technical difficulties during this project: the configuration of OpenFace, the installation of *minted* package in LaTex for colourful code snippets, the conversion of fNIRS from light intensity to haemoglobin concentration, etc. It would be trivial and unnecessary to list all the details, and I believe that the provided information is enough for people to rebuild a similar project. This brings the end of this project, but questions are endless, the same as study. I expect this project could be extended and more progress could be achieved in the near future.

References

[1] Artinis medical systems, Einsteinweg 17, 6662 PW Elst, The Netherlands. *Manual OctaMon for OxySoft 3.0.103 and higher, V1610*, unknown.

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.

[3] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 354–361. IEEE, 2013.

[4] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[5] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

[6] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, 2014.

[7] B. Cain. A review of the mental workload literature. Technical report, Defence Research And Development Toronto (Canada), 2007.

[8] J. F. Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238. ACM, 2006.

[9] M. Cope. The application of near infrared spectroscopy to non invasive monitoring of cerebral oxygenation in the newborn infant. *Department of Medical Physics and Bioengineering*, 342:317–323, 1991.

[10] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using phog and lpq features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883. IEEE, 2011.

[11] D. F. Dinges, R. L. Rider, J. Dorrian, E. L. McGlinchey, N. L. Rogers, Z. Cizman, S. K. Goldenstein, C. Vogler, S. Venkataraman, and D. N. Metaxas. Optical computer recognition of facial expressions associated with stress induced by performance demands. *Aviation, space, and environmental medicine*, 76(6):B172–B182, 2005.

[12] A. Duncan, J. H. Meek, M. Clemence, C. E. Elwell, P. Fallon, L. Tyszczuk, M. Cope, and D. T. Delpy. Measurement of cranial optical path length as a function of age using phase resolved near infrared spectroscopy. *Pediatric research*, 39(5):889, 1996.

[13] P. Ekamn and W. Friesen. Facial action coding system (facs): manual, 1978.

[14] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[15] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[16] P. Ekman, W. V. Friesen, and P. Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.

[17] R. El Kaliouby and P. Robinson. Generalization of a vision-based computational model of mind-reading. In *International Conference on Affective Computing and Intelligent Interaction*, pages 582–589. Springer, 2005.

[18] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[19] M. Ferrari and V. Quaresima. A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application. *Neuroimage*, 63(2):921–935, 2012.

[20] A. S. Geller, I. K. Schleifer, P. B. Sederberg, J. Jacobs, and M. J. Kahana. Pyepl: A cross-platform experiment-programming library. *Behavior research methods*, 39(4):950–958, 2007.

[21] J. M. Girard, J. F. Cohn, L. A. Jeni, M. A. Sayette, and F. De la Torre. Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behavior research methods*, 47(4):1136–1147, 2015.

[22] D. N. Glaser, B. C. Tatum, D. M. Nebeker, R. C. Sorenson, and J. R. Aiello. Workload and social support: Effects on performance and stress. *Human Performance*, 12(2):155–176, 1999.

[23] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*, 2013.

[24] R. Gross. Face databases. *Handbook of face recognition*, pages 301–327, 2005.

[25] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50-9, pages 904–908. Sage Publications Sage CA: Los Angeles, CA, 2006.

[26] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz. Mental workload during n-back taskquantified in the prefrontal cortex using fnirs. *Frontiers in human neuroscience*, 7:935, 2014.

[27] L. M. Hirshfield, E. T. Solovey, A. Girouard, J. Kebinger, R. J. Jacob, A. Sassaroli, and S. Fantini. 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*, pages 2185–2194. ACM, 2009.

[28] iMotions. Facial action coding system(facs) - a visual guidebook. `https://imotions.com/blog/facial-action-coding-system/`. Accessed: 2018-04-09.

[29] A. Jaimes and N. Sebe. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1-2):116–134, 2007.

[30] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013.

[31] Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, 36(5):862–875, 2006.

[32] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE, 2011.

[33] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

[34] M. J. Kane, A. R. Conway, T. K. Miura, and G. J. Colflesh. Working memory, attention control, and the n-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3):615, 2007.

[35] M. Khamis, A. Baier, N. Henze, F. Alt, and A. Bulling. Understanding face and eye visibility in front-facing cameras of smartphones used in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 280. ACM, 2018.

[36] J. M. Kivikangas, G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvelä, and N. Ravaja. A review of the use of psychophysiological methods in game research. *journal of gaming & virtual worlds*, 3(3):181–199, 2011.

[37] K. L. Koenraadt, E. G. Roelofsen, J. Duysens, and N. L. Keijsers. Cortical control of normal gait and precision stepping: an fnirs study. *Neuroimage*, 85:415–422, 2014.

[38] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly. Automated measurement of children's facial expressions during problem solving tasks. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 30–35. IEEE, 2011.

[39] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[40] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011.

[41] H. A. Maior, M. L. Wilson, and S. Sharples. Workload alert - using physiological measures of mental workload to provide feedback during tasks. *ACM Transactions on Computer-Human Interaction*, (in Press) 2016.

[42] A. Marinescu, S. Sharples, A. Ritchie, T. S. López, M. McDowell, and H. Morvan. Exploring the relationship between mental workload, variation in performance and physiological parameters. *IFAC-PapersOnLine*, 49(19):591–596, 2016.

[43] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[44] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[45] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2017.

[46] S. T. Mueller and B. J. Piper. The psychology experiment building language (pebl) and pebl test battery. *Journal of neuroscience methods*, 222:250–259, 2014.

[47] N. Naseer and K.-S. Hong. fnirs-based brain-computer interfaces: a review. *Frontiers in human neuroscience*, 9:3, 2015.

[48] H. T. Nguyen, C. Q. Ngo, K. Truong Quang Dang, and V. T. Vo. Temporal hemodynamic classification of two hands tapping using functional nearinfrared spectroscopy. *Frontiers in human neuroscience*, 7:516, 2013.

[49] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[50] M. F. Pike, H. A. Maior, M. Porcheron, S. C. Sharples, and M. L. Wilson. Measuring the effect of think aloud protocols on workload using fnirs. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3807–3816. ACM, 2014.

[51] K. Ryu and R. Myung. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, 35(11):991–1009, 2005.

[52] A. Sassaroli, F. Zheng, L. M. Hirshfield, A. Girouard, E. T. Solovey, R. J. Jacob, and S. Fantini. Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *Journal of Innovative Optical Health Sciences*, 1(02):227–237, 2008.

[53] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008.

[54] A. Savran, K. Ciftci, G. Chanel, J. Mota, L. Hong Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut. Emotion detection in the loop from brain signals and facial images. *eNTERFACE 2006 Workshop*, 2006.

[55] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[56] R. T. Stone and C.-S. Wei. Exploring the linkage between facial expression and mental workload for arithmetic tasks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 55-1, pages 616–619. SAGE Publications Sage CA: Los Angeles, CA, 2011.

[57] S. Tak, M. Uga, G. Flandin, I. Dan, and W. Penny. Sensor space group analysis for fnirs data. *Journal of neuroscience methods*, 264:103–112, 2016.

[58] S. Tak and J. C. Ye. Statistical analysis of fnirs data: a comprehensive review. *Neuroimage*, 85:72–91, 2014.

[59] M. Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002.

[60] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[61] P. S. Tsang and V. L. Velazquez. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381, 1996.

[62] Unknown. Spm-fnirs toolbox. `https://www.nitrc.org/docman/view.php/965/1995/manual_spm_fnirs.pdf`. Accessed: 2018-04-07.

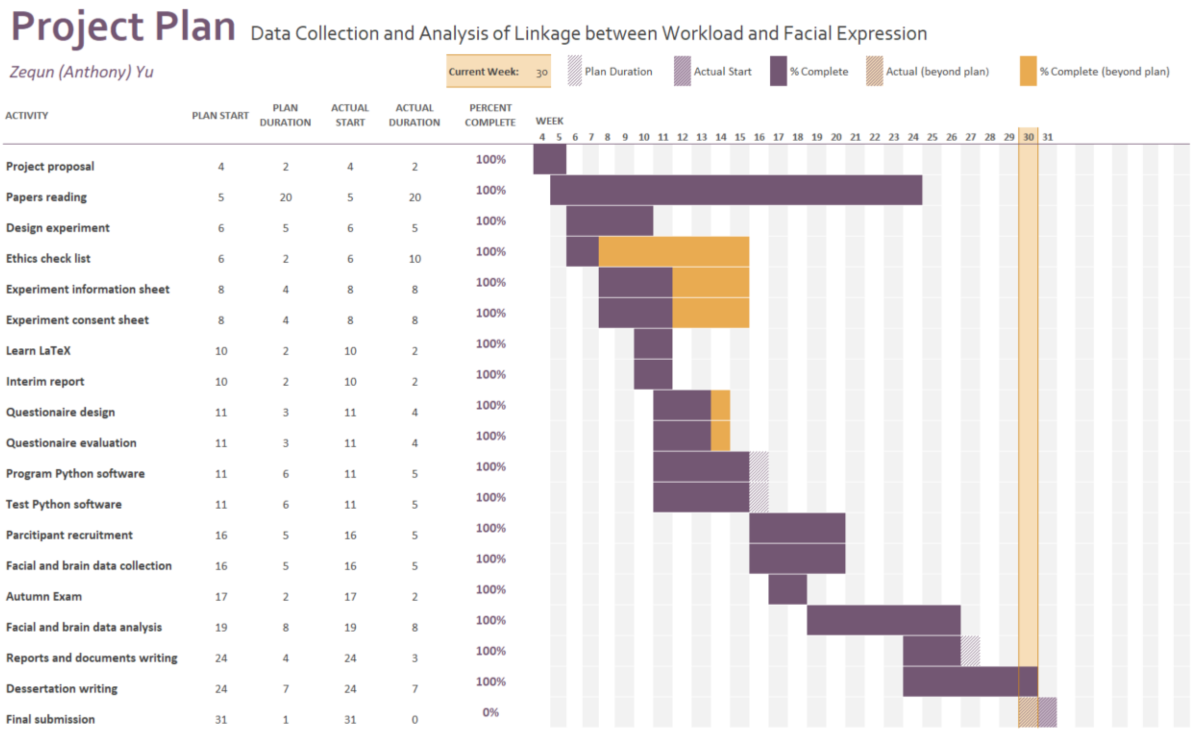
[63] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[64] J. Whitehill, M. S. Bartlett, and J. R. Movellan. Automatic facial expression recognition. *Social Emotions in Nature and Artifact*, 88, 2013.

[65] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[66] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

Appendix A Gantt Chart



Project Plan Data Collection and Analysis of Linkage between Workload and Facial Expression

Zequn (Anthony) Yu

Current Week: 30

▨ Plan Duration ▨ Actual Start ■ % Complete ▨ Actual (beyond plan) ■ % Complete (beyond plan)

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Project proposal	4	2	4	2	100%
Papers reading	5	20	5	20	100%
Design experiment	6	5	6	5	100%
Ethics check list	6	2	6	10	100%
Experiment information sheet	8	4	8	8	100%
Experiment consent sheet	8	4	8	8	100%
Learn LaTeX	10	2	10	2	100%
Interim report	10	2	10	2	100%
Questionaire design	11	3	11	4	100%
Questionaire evaluation	11	3	11	4	100%
Program Python software	11	6	11	5	100%
Test Python software	11	6	11	5	100%
Parcitipant recruitment	16	5	16	5	100%
Facial and brain data collection	16	5	16	5	100%
Autumn Exam	17	2	17	2	100%
Facial and brain data analysis	19	8	19	8	100%
Reports and documents writing	24	4	24	3	100%
Dessertation writing	24	7	24	7	100%
Final submission	31	1	31	0	0%

Figure 14: The Gantt chart of the project