The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA
School of Computer Science

4331522

Supervisor: Dr Susan Ali

Module Code: COMP3003

The effectiveness of using augmented reality technology for children during Covid-19

Submitted May 2021, in partial fulfilment of

the conditions for the award of the degree **BSc Hons Computer Science**

4331522

School of Computer Science

University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature ________________________

Date 21/05/2021

Acknowledgements

Throughout the writing of this dissertation, I have received a great deal of support and assistance. I would first like to thank my supervisor, Dr Susan Ali, whose expertise was invaluable in completing my dissertation. Without her suggestions, ideas, and valuable comments, this dissertation would not be possible. I would also like to thank my friend, Sen Tho, for providing me with his notes on the ‘Machine Learning’ module. It was highly beneficial as it allowed me to understand the core concept without having been enrolled on the module. Finally, I would also like to thank my parents and our country's government for providing me with financial support to successfully complete my studies.

Abstract

Due to Covid-19, a sudden shift towards e-learning has been made. This shift has largely been negative since e-learning fails to provide a meaningful incentive for children to study at home. Hence, there is a need to identify a suitable e-learning method that can solve this problem. In this dissertation, we propose the use of Augmented Reality (AR) to help overcome this obstacle. We will be measuring the actual benefit of AR by comparing AR to other forms of e-learning methods, these methods include: learning via video calls and watching pre-recorded educational videos. To successfully make the comparison, we will be focusing on two aspects: “the engagement factor” and “memory retention”. “The engagement factor” involves measuring the level of engagement a certain learning method offers. “Memory retention” involves measuring the volume of information children can retain after learning using a certain learning method. This research paper would propose the use of two variations of automated engagement tracking software to measure “the engagement factor”. One of the variations would use facial tracking software, whereas the other variation would measure the participants' eye-blink rate.

TABLE OF CONTENTS

List of Acronyms

LIC	Low-income countries
MIC	Middle-income countries
HIC	Hiddle-income countries
AR	Argument Reality
COVID-19	Coronavirus disease 2019
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
NPIs	Nonpharmaceutical interventions
FACS	Facial Action Coding System
AU	Action Unit
RBF	Radial Basis Function
SVM	Support Vector Machine
CERT	Computer Expression Recognition Toolbox
FPS	Frames per second
KNN	K-Nearest Neighbors
Bagging	Bagged Decision Trees
CNN	Convolutional Neural Network
HOG	Histogram of oriented gradient
HOG-SVM	Histogram of oriented gradient with a combination of Linear SVM
GPU	Graphics processing unit
CPU	Central processing uni
EAR	Eye aspect ratio
EAR_{avg}	The average eye aspect ratio
C	Cost / penalty coefficient
Electrodermal activity	EDA
Standard deviation	SD
Root mean square error	RMSE

1 Introduction

1.1 Definitions

Words that may need further clarifying are underlined throughout the thesis and are defined below:

Two-way synchronous learning allows both the learner and the teacher to communicate via a medium in real-time; it often involves online chats and videoconferencing via Skype, Microsoft Teams

One-way asynchronous learning consists of content being delivered one way (towards the learner); it often involves watching pre-recorded educational videos, listening to broadcast radio shows, etc. One-way asynchronous learning is more flexible than two-way synchronous learning as the learners can learn at their desired time.

Head pose is defined as the orientation of a person's head relative to the view of a camera. It is calculated by measuring the relative location and rotation of the person's head with respect to the camera.

Blink rate is the number of blinks a person performs per minute

Facial Action Coding System (FACS) is a comprehensive, anatomically based system for describing all visually discernible facial movements. It breaks down facial expressions into individual components of muscle movement, called Action Units (AU) [16].

OpenFace is open-source software that can recognize faces and facial features using deep neural networks in real-time. It can detect facial landmarks, head pose, AU and track eye gaze. Out of 44 AUs, it can recognize a subset of AUs, and these include AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, AU45. The responsibility of each AU mentioned above is shown in table 2.

AU number	Responsibility
01	Inner brow raiser
02	Outer brow raiser
04	Brow lowerer
05	Upper Lid Raiser
06	Cheek Raiser
07	Lid Tightener
09	Nose Wrinkler
10	Upper Lip Raiser
12	Lip Corner Puller
14	Dimpler
15	Lip Corner Depressor
17	Chin Raiser
20	Lip Stretcher
23	Lip Tightener
25	Lips Part
26	Jaw Drop
45	Blink

Table 2- shows the list of AU and their responsibility.

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression

Feature selection is the process of reducing the number of variables input in machine learning when developing a predictive model. It often involves discarding variables input that has little to no significant relevance. The number of variables input are often reduced for many reasons; this includes: simplifying the models to make them easier to interpret by researchers/users, lessening the computational cost and hence shorter training times, to avoid the 'curse of dimensionality', and finally enhanced generalization by reducing overfitting.

Computer Expression Recognition Toolbox (CERT) is an system that can detect facial expression in real-time. It claims to be able to detect 40 continuous dimensions and 30 AUs from the FACS. While CERT was a free system for use, it is only available to the institution for a given price.

P-value or calculated probability is the probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true [17].

Randomized search is a technique where random combinations of the hyperparameters are used to find the 'best' solution for the built model and is faster than grid search.

Grid search is a technique where every combination of the hyperparameters are used to find the best solution for the built model and is slower than randomized search but is more accurate.

K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. It operates by randomly splitting the data into k folds (groups) of approximately equal size, with one of the folds being treated as a testing set and the remaining k-1 fold as the training set. The process repeats itself K times such that the fold used for testing is unique each time.

OpenCV is a library which is used for real-time computer vision.

Dlib is an open-source library that contains machine learning algorithms and other tools for solving real-world problems. As of 2021, it contains major algorithms such as numerical, machine learning, graphical model inference, image processing, threading, networking, data compression and integrity, etc [18].

HOG-SVM HOG is a feature descriptor that is used to extract features from an image. It extracts features by overlaying a pixel grid onto the image and then calculates the gradient and the orientation of each pixel. It can assess if a face is present and its location by using the extracted feature as an input into an SVM algorithm for classification.

C is often labelled as the cost/penalty coefficient and is a hyperparameter in SVM. It is responsible for trading off the correct classification of the training dataset against the functional margin. A good balance of C is required since a large 'C' correctly classifies all training points but leads to a smaller margin and overfitting, thus yields low bias and high variance. Whereas a small 'C' leads to a larger margin but incorrectly classifying training points yielding a high bias and lower variance.

Gamma is also a hyperparameter in SVM, and it is defined as the distance of influence of a single training data point.

Image target is a unique image used by the Vuforia's camera to trigger an AR object on the screen. While Vuforia does not explicitly state how their tracking algorithms works [19] it is apparent that they use a variation of a natural feature tracking algorithm; therefore, features such as sharp details with high contrast corners are easily tracked [20]. Hence, an online Augmented Reality Marker Generator [21] was used to generate images with all these features in this thesis.

Pearson correlation coefficient is used to measure the linear correlation between two sets of data. Table 15 shows the Pearson correlation coefficient value and the corresponding direction and strength of correlation.

Pearson correlation coefficient	Direction and strength of correlation.
-1	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakly negative
0	No association
0.2	Weakly positive
0.5	Moderately positive
0.8	Strongly positive
1	Perfectly positive

Table 15- shows the Pearson correlation coefficient value and the corresponding direction and strength of correlation.

Root mean squared error (RMSE) measures the standard deviation of the residuals, where the residuals are measure of the distance between the regression line and data points. The formula for RMSE is shown on equation 3, where O_i are the observations, S_i are predicted values of a variable, and n is the number of observations available for analysis.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2} \quad (3)$$

1.2 Motivation

Due to the severity of the ongoing COVID-19 pandemic caused by SARS-CoV-2 [1] , over 2 million people have lost their lives as of the current date - January 19, 2021 [2]. To slow down the spread of the deadly virus, many countries have resorted to NPIs [3] - these interventions include social distancing, isolation, and quarantine; ultimately leading to 1.6 billion children being unable to attend school at its peak [3]. As a result, different forms of e-learning techniques were adopted during this period. For instance, in LIC, lessons were broadcasted through live television and radio shows, whereas in MIC and HIC, both one-way asynchronous and two-way synchronous online classes were used.

Contrary to popular belief, classroom learning isn't more effective than e-learning, as research has shown learners understand more information using computer-based instruction rather than the traditional classroom methods [4]. Hence learners retain more information using e-learning relative to classroom learning. However, one must note that while e-learning is more effective at helping students retain information, it is plagued with a high dropout rate, with 20 to 80 % of people failing to complete an e-learning course as the learners often fail to find the content either engaging or exciting [5]. This is especially the case with children as they are less attentive than adults and are easily distracted when the content is unengaging. Therefore, there is a need to find an ideal form of e-learning method that has the added benefit of being effective (in terms of information retention) but is also engaging, especially for children (age: 7 to 11) who are the most affected.

AR provides a promising solution to this problem as research has shown that participants who use AR show an increased level of engagement [6], along with other additional benefits, especially in the educational field, as summarised in table 1. Therefore, one of the aims of this research paper is to investigate this promising solution, hence explore the effectiveness of using AR in the academic environment on children relative to other forms of e-learning methods that are currently being used during the pandemic; where effectiveness would involve measuring “the engagement factor” and “memory retention”. This research will only focus on the ‘simple’ AR teaching method which involves the use of an image target to show rudimental AR objects/images on a mobile device. Only rudimental AR objects/images would be considered as they do not require in-depth 3D modelling knowledge, hence if one were to use AR in a real-life scenario during the pandemic as a teaching method, it can be quickly deployed in the education field without the need for school teachers to understand the core concept of 3D modelling. Also, one must note that this research will not consider the use of high-tech equipment (such as AR headset) due to the limited funding for this research and also due to the Covid-19 restrictions.

Advantage of using AR	Research paper title
Increases students interest in learning	Applications of augmented reality-based natural interactive learning in magnetic field instruction [7].
Increases students motivation in learning and makes learning entertaining	An Interactive Mobile Augmented Reality Magical Playbook: Learning Number with the Thirsty Crow [8].
Develops spatial ability	Affordances of Augmented Reality in Science Learning: Suggestions for Future Research [9].
Increases collaboration	Augmented reality: An overview and five directions for AR in education [10].
Increases engagement	Using the Augmented Reality 3D Technique for a Convex Imaging Experiment in a Physics Course [11].

Table 1- Studies show the advantage of using AR in the educational field.

In order to compare “the engagement factor” of AR relative to other forms of e-learning methods in the academic area, one must measure how engaged a child becomes whilst learning using such methods in real-time. Before we can devise different ways of measuring engagement, one must first answer the simple question: What is engagement? In layman terms, engagement is referred to the degree of attention, curiosity, interest, optimism, and passion that students show when they are learning or being taught, which extends to the level of motivation they have to learn and progress in their education. When people refer to “engagement” in the education sector, they are actually referring to the three different forms of engagement. These three forms of engagement are affective engagement, behavioural engagement, and cognitive engagement [12]. Affective engagement is the measure of emotional reactions the user displays while completing the task [13]. For instance, the greater the positive emotion relative to a user’s negative emotion, the greater the user’s affective engagement. Behavioural engagement measures the student’s physical involvement in the learning, by assessing their participation in the given academic tasks [13]. For instance, a student who is completing a given task and is generally answering questions and staying “on-task” is said to have a greater behavioural engagement. Cognitive engagement measures a student’s willingness to take on the learning task at hand [14]. For instance, a student showing a high level of cognitive engagement will invest a significant amount of effort and time in completing a task, even if it is perceived as complex. All three types of engagement can be measured using either internal or external forms of observation. An internal form of measurement includes self-report, and an external form involves the use of a third-party observer – this can be either an automated machine or another human [12]. Internal forms of observation are deemed to be inferior to external forms since they are extremely subjective and are often misrepresented as participants tend to inflate their level of engagement in a certain task. Therefore, only external forms of observation would be considered in this dissertation.

In terms of whether to use an automated machine or another human for external forms of observation, we will consider the use of an automated machine. Soley since this concept of automated machine to monitor student's engagement can be of great use for teachers during Covid-19 and hence can be directly used outside of this research confinements. This is since teachers cannot monitor each student whilst teaching online due to remote learning. They cannot identify students who are struggling with their task, who are not 'paying attention', or who are not even present- especially when the cameras during online video call are turned off. Therefore, using a form of engagement measuring technique can help overcome these obstacles and get added benefits such as instant feedback on their students' engagement levels. This can also help teachers identify students who are not attentive and can also help teachers reflect on their own teaching strategies, essentially leading to improved teaching and learning environments. Therefore another aim of this research is to implement and compare different forms of engagement measuring techniques. Finally, since there seems to be a correlation between engagement and academic performance [15], this dissertation will investigate the actual relationship between learning and engagement and the possibility of predicting one's grade by their level of engagement.

1.3 Limitations

While this research paper seeks to find the effectiveness of using AR in the educational field for children, this paper will only focus on one subject – French. This is to make the experiment brief and appealing for the participants such that they can be easily recruited. For instance, an experiment focusing on only one subject requires ~30 minutes, with each additional subject requiring an additional 10-15 minutes. Suppose this research was to focus on all 12 subjects that are taught to children during primary school. In that case, it will require ~150 – 210 minutes, which is impractical and extremely time-consuming for the participants, ultimately making it very difficult to recruit multiple participants for such an experiment; hence only one subject was chosen. French was selected because it is a subject that requires students to retrieve information via memory recall which can be examined relatively easily – with a written assessment/quiz (see section 4.3).

Another limitation of this project is that it only focuses on three forms of e-learning. These include the use of AR, pre-recorded and online video calls/videoconference (most common being skype/zoom classes). Pre-recorded and online video calls/videoconference were chosen for this study because they both capture the most popular form of e-learning methods currently employed during the pandemic. For instance, pre-recorded is a form of one-way asynchronous learning and is the most popular method used in LIC, whereas online video calls/videoconference is a form of two-way synchronous learning and is the most popular method being employed in both HIC, MIC [22]. The main reason this thesis only focuses on only three forms of e-learning methods is due to the limited number of participants. For instance, in the research, the nine participants are split into different groups, with each group having three members and a different form of e-learning method (see section 4.1). If we also considered other forms of e-learning, it would require a more significant number of participants, which we were, unfortunately, unable to recruit.

Other major limitations include the use of a limited dataset to train the face tracking classification model due to inadequate CPU's processing power (see section 3.1.1.). This limitation of the CPU's and GPU processing power also meant that an inferior Dlib algorithm was used to detect and locate faces (i.e HOG-SVM algorithm was used instead of CNN – see section 3.2.1), which negatively impacted the accuracy of the eye-blink rate program.

1.4 Aims and Objectives

This research consists of two main aims and one minor aim: It aims to investigate the effectiveness of using AR in the educational field on children relative to other forms of e-learning. Furthermore, this project also aims to implement and compare different forms of engagement measuring techniques; this includes the use of facial tracking software and a program that measures participant's eye-blink rate. Finally, this project will investigate the actual relationship between learning and engagement and the possibility of predicting one's grade by their level of engagement.

The aims can be split into the following objectives:

1. Test and train a machine learning model that can classify students' engagement by using an output from facial tracking software.
2. Design and develop an eye-blink rate program that can measure students' engagement by measuring their eye-blink rate.
3. Teach the participants using their assigned e-learning methods; this includes AR, pre-record videos and video calls.
4. Investigate the effectiveness of each e-learning method (where effectiveness refers to "memory retention" and "the engagement factor")
5. Investigate the performance of both facial tracking software and eye-blink rate program against a ground truth
6. Find the relationship between learning and engagement

1.5 Hypotheses

The hypothesis being tested are as follows:

1. Children retain more information when learning using AR relative to learning via video calls or pre-recorded videos.
2. Children are more engaged when learning using AR relative to learning via video calls or pre-recorded videos.
3. There is a *strong* correlation between learning and engagement.
4. The eye blink-rate measuring program will be more effective at measuring student's engagement.

2 Related Works

For this project to contribute to the field of computer science, it must provide new ideas / be a novel study, or it must improve on the current work that has already been done in the same field. Therefore, it is imperative to review studies that are related to this project. These reviews can be split into three sections; they are as follows:

- The use of e-learning for children (Section 2.1)
- The use of facial tracking software for measuring engagement (Section 2.2)
- The use of eye blink tracking software for measuring engagement (Section 2.3)

2.1 The use of e-learning for children

While there exists an extensive amount of research on the use of e-learning, upon further examination, it has been revealed that the data which have been collected are often biased. For instance, a paper [23] completed a meta-analysis on 111 abstracts; out of the 111 abstracts, sixty-one studies conclude the effectiveness of e-learning. The paper showed ~90% of the sixty-one studies classified e-learning as ‘effective’, but upon further investigation, it became clear that most researchers who made such claims appeared to have a stake in the success of e-Learning. Therefore, this raises a question on the credibility of research that has been concluded in this field. Thus, the paper [23] suggest that a further unbiased investigation of e-learning in the education field is needed. This is indeed a focus of this dissertation as it will be able to contribute to the field of e-learning and hence improve on previous biased work.

Very few studies are available that directly compare AR learning to other forms of e-learning. Out of the very few, one of the most relevant studies made a direct comparison between augmented reality, virtual learning environment and learning via the use of smartphone applications [24]. The study concluded that learning using smartphone applications is far superior to learning using AR or virtual learning environments as it is more convenient for both teachers and learners. However, this study only made a qualitative analysis which is subjective to the researcher. No actual quantitative data was captured or studied regarding which learning method is more engaging or yield better information retention, which is one of the main targets of this dissertation. Moreover, whilst mobile learning is indeed a form of e-learning, it is not the most prominent form of e-learning that is used during the pandemic. Therefore, this dissertation will bring the novelty of comparing AR learning to learning via video calls / by watching pre-recorded videos which are the most popular forms of e-learning methods used during the pandemic in both HIC, MIC as mentioned in the section 1 and will also aim to collect qualitative analysis.

One must note that a wealth of studies exist on purely the use of AR in the education field, as research has shown that AR promotes one of the main interactions needed in education - learner-content interaction [25] [26]. This is since learner-content interaction has shown to enhance memory, imagination and understanding of tasks [27] [28]. The other added benefits can be seen in Table 1. Moreover, research conducted on the subject of visual art courses also showed an increase in confidence, satisfaction and overall student’s motivation [25]. This study was fascinating since it involved students (age 13–16, M = 13.7), which is vaguely similar to the age range of students that will be considered in this dissertation. Students were taught about famous art (Mona Lisa, The Starry Night, etc.) using AR during their experiment. The AR application consisted of images of the art and their description. Pre and post-experiment, a survey was completed by the students to assess their learning motivation, this is a use of an internal form of observation. This dissertation would use similar methods; however, in terms of novelty, as previously mentioned, it would *compare* the effectiveness of AR to other e-learning methods and will use an external form of observation (more specifically automated machine), since an internal form of observation is extremely subjective as previously mentioned in section 1.2.

2.2 The use of facial tracking software for measuring engagement

Many studies have been conducted to measure user’s engagement by observing user’s facial features and their expression and movements. Some of the facial features include head pose, eye gaze, FACS. A study performed in 2017 used OpenFace to extract student’s FACS, head poses and eye gaze in real-time of 10 participants (age: 21-26) [29]. The study then categorizes the participant’s engagement as either “engaged” or “unengaged” using automated and manual techniques. The automated techniques consisted of using two different machine learning algorithms, such as SVM (Linear and RBF kernel) and Logistic Regression, with varying features (feature selection). The manual techniques consisted of splitting the recorded videos into 10 seconds slices. 2280 different samples were extracted, and each slice was manually labelled as either “engaged” or “unengaged”. Using the manually labelled annotation as ground truth, the accuracy, F1 score, precision, recall of each machine learning algorithm, and its variant were observed. The result shows the best accuracy, F1 score, precision, recall were obtained when 8/27 features were selected; the eight features include the standard deviation of rotation of the head around x,y axes, the standard deviation of the gaze vector in x,y, z direction, AU02, AU05, AU17. In this dissertation, a similar feature selection approach will be applied; however, in terms of novelty, instead of only considering two machine-learning algorithms to classify the user's engagement state, a wide variety of algorithms will be considered. The most optimum algorithms will then be used, hence providing more accurate data. Also, in their study, the participant's ages varied from 21-26; in this dissertation, the age group will range from 7-11. This will not only help answer our hypotheses but also help further the understanding of facial tracking software for different age groups.

A study conducted by Whitehill et al. in 2014 [30] also measured student's engagement using facial tracking software. However, in this study, CERT was used to extract 19 different AU in real-time. Similar to the previous method report [29], this study also used human observation to obtain the ground truth. The ground truth was also obtained by splitting the recorded video into 10 seconds slices. However, in this research, multiple teachers annotated the videos into four different categories of engagement, ranging from one to four, with one being 'not engaged' and four being 'very engaged'. The study stated that the majority of the teachers agreed with each other, and their observation was reliable when discriminating low versus high degrees of engagement, but when a fine level of discrimination was required (four distinct levels), the reliability decreases significantly. Therefore, in this dissertation, to make the data more reliable, engagement will only be classified into two groups: engaged and unengaged.

2.3 The use of eye blink tracking software for measuring engagement

Studies have shown animals perform fewer blinks when focused on an object to minimize any potential loss of visual information, especially when the information is being perceived visually[42][43][44]. This has lead to an increase in numbers of research surrounding the use of blink rate to predict task difficulty [31], mental workload [32], cognitive performance [33]. However, the research surrounding the use of eye-blink rate to measure engagement is still very limited. Out of the limited studies, one of the most prominent research includes a study conducted in 2017 [34], in which the paper looks at the relationship between visual attention (engagement) and eye blink frequency. In this paper, electrodermal activity (EDA) was used to measure eye blink frequency as they claimed that there is a correlation between EDA and eye blink frequency. An EDA sensor was used to measure electrodermal activity. EDA sensor operators by measuring the electrical signal between two electrodes in contact with the skin. The study showed that the blink rate was lower when participants were engaged (when visual attention was needed). The blink rate was higher when the participants were unengaged (when visual attention was not needed). This dissertation aims to further their studies in many ways. For instance, using the notion of there being a relationship between engagement and eye blink frequency, we will use this relationship to predict the participant's engagement level by measuring their eye-blink frequency/rate. Also, as mentioned in section 1.1, a part of this dissertation is to find a practical method of measuring students engagement level during the pandemic to help enhance teaching and learning for both teachers and students. Hence, using an EDA sensor to measure engagement in a real-world situation is not practical and is inconvenient due to many reasons. One of the main reasons is that EDA sensors consist of rather sensitive and fragile wires that need to be attached to the students at all times and will require students to remain stationary for an extended period as even a minor movement can cause the EDA sensor to produce false data. Also, since EDA sensors are research equipment that is not readily available in people's homes, it will be impractical to ask the general public to purchase EDA sensors to measure their engagement in a real-world scenario - during Covid-19. Much suitable alternative would be the use of a webcam/camera device in a smartphone, tablets, etc. Therefore, in this research, we provide the novelty of measuring an engagement / eye-blink rate using a simple webcam or a camera in a smartphone rather than EDA sensors. In terms of methods, this dissertation will create an eye-blink rate program to measure data for this research and also provide an open-source program that can be used to further this study. The eye-blink rate would be calculated by detecting facial landmarks [35] [36] [37] and by extracting measurements of the participant's eye – see section 3.2.

Another study [38]performed a similar experiment. However, in this case, the participants (21 adults) watched a movie scene, with some of the participants being told to count the amount of time they encounter a 'land scene'. Whilst, the other participants were told to count the amount of time they encountered 'water scene'. The results showed that the group counting the numbers of 'land scene' showed decreased blink rate during the 'land scene' and increased blink rate during the water scene. Similar, the group counting the numbers of 'water scene' showed a decreased blink rate during the 'water scene' and increased blink rate during the land scene, thus showing the effect of engagement on eye blink rate. While this dissertation would vary significantly from their study [38] in terms of different age groups(participants being children age range from 7-11 rather than adults), the approach in which the eye blink rate would be measured (using smartphone/web camera), the main difference would be the context in which the participant's eye-blink rate is measured (for instance, this dissertation will be focused on measuring eye-blink rate during the student's learning process rather than merely measuring the engagement of participants whilst they are viewing a movie scene).

3 Implementation and Implementation Analysis

3.1 Implementation of Facial Tracking

Facial tracking software will not be created in this dissertation since there are wide ranges of free (for research purposes) and highly accurate software with various functionality, as shown in Table 3.

Software	Extract head pose	Extract AUs	Extract eye gaze	Extract features in real-time
KeenTools FaceTracker [38]	-	✓		✓
dlib [18]	-	-	-	✓
Chehra [39]	✓	-	-	-
OpenPose [40]	✓	-	-	✓
Openface	✓	✓	✓	✓
WebGazer.js	-	-	✓	✓

Table 3- shows the list of facial tracking software and its capabilities.

By looking at table 3, it is apparent that OpenFace is the most optimum software for this research as it is the only software that can extract head pose, AU, eye gaze in real-time.

OpenFace processes the raw real-time footage and outputs a .cvs file containing:

- frame number
- the timestamp of each frame
- values for extracted AU, which is measured in intensity ranging from 0 to 5
- values for the extracted head pose, which is measured in millimetres for the head location and radians for the head rotation
- values for eye gaze, which is measured in radians

This output has to be processed and classified by a machine learning model to obtain data regarding the students' level of engagement.

3.1.1 Dataset

In order to process the raw data obtained from OpenFace, a binary classification model must be trained and tested using a dataset. In this dissertation, DAISEE [41] dataset was chosen. DAISEE dataset contains 9068 video snippets captured from 112 users, each of which is 10 seconds long. The emotional state of each video was divided into four groups: boredom, confusion, engagement and depression. In DAISEE, each emotional state was given a rating out of 4, where zero represented a person showcasing a very low case of that given emotion, whereas 4 represents the person showcasing a very high level of that given emotion. Since not all videos were labelled, out of the 9068 video clips, only 8925 were usable. Since this paper only focuses on the degree of student engagement level, the values for boredom, confusion and depression emotional state were discarded. Table 4 shows the distribution of the videos and their corresponding level of engagement.

Engagement level	Numbers of videos
1	61
2	455
3	4422
4	3987

Table 4- shows the distribution of the videos and their corresponding level of engagement.

As previously mentioned in section 2.2, [30] states that finding the ground truth is very difficult and highly unreliable when four distinct levels of engagement are considered. Thus, the four levels of engagement merged to form two groups of engagement: engaged and engaged unengaged. The engaged group consisted of engagement levels one and two, whereas the unengaged group consisted of engagement levels three and four.

The videos were then inputted into the OpenFace software, which then extracted the head pose, eye gaze, and AUs, resulting in a new form of a dataset that can be used to train and test machine learning models. Due to the CPU's processing limitation, one must note that not all of the footage from table 4 were used. Only 52 randomly selected video clips were used for this research, with 50% of the video from the engaged group and the other 50% from the unengaged group, which is a limitation of this research. Since the OpenFace process 30 FPS, 73200 frames were used to create the new dataset.

3.1.2 Feature Selection

The initial feature vector was 31-dimensional; hence, to remove the irrelevant or partially relevant features, a p-value was obtained; table 5 shows the feature and their corresponding p-value.

Features	Corresponding p-value
1. Left eye gaze - vector in x, y, z directions (3 features) 2. Right eye gaze - vector in x, y, z directions (3 features) 3. Both mean eye gaze - vector in x, y, z directions (3 features) 4. Head pose rotation in x,y,z direction (3 features) 5. AU these include: AU01, AU02, AU04, AU05, AU14, AU45, AU06, AU07,	p-value < 2.2e-16

AU09, AU10, AU12, AU15,AU17, AU20, AU23, AU25, AU26 (17 feature)	
Head pose location in x-axis with respect to the camera in millimetres	0.05406
Head pose location in y-axis with respect to the camera in millimetres	0.05409
Head pose location in z-axis with respect to the camera in millimetres	0.05407

Table 5- shows the feature and their corresponding p-value.

Since $p > 0.05$ is statistically insignificant [42], head pose location in x,y,z with respect to the camera in millimetres were discarded, leaving 29 features for machine learning.

3.1.3 Hyperparameter tuning

Seven different binary classification algorithms were considered, they were as follows:

1. Logistic Regression
2. Ridge Classifier
3. KNN
4. SVM
5. Bagging
6. Random Forest
7. Stochastic Gradient Boosting

The randomized search algorithm was performed on the dataset to find the most optimum hyperparameter. Table 6 shows the algorithms, the hyperparameter that was considered and the most optimum hyperparameter for the given algorithm.

Algorithms	Hyperparameters that were considered	Optimum hyperparameter
Logistic Regression	solvers = ['newton-cg', 'lbfgs', 'liblinear'] penalty = ['l2'] c_values = [100, 10, 1.0, 0.1, 0.01]	Solver = 'lbfgs', penalty = 'l2', C = 100
Ridge Classifier	alpha = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	Alpha = 0.1
KNN	n_neighbors = 1-21 weights = ['uniform', 'distance'] metric = ['euclidean', 'manhattan', 'minkowski']	Metric = 'manhattan', n_neighbors = 2, weights = 'uniform'
SVM	kernel = ['poly', 'rbf', 'sigmoid'] C = [50, 10, 1.0, 0.1, 0.01] gamma = ['scale']	Kernel = 'rbf' , gamma = 'scale', C= 50
Bagging	n_estimators = [10, 100, 1000]	n_estimators = 1000
Random Forest	n_estimators = [10, 100, 1000] max_features = ['sqrt', 'log2']	n_estimators = 1000, max_features = 'sqrt'
Stochastic Gradient Boosting	n_estimators = [10, 100, 1000] learning_rate = [0.001, 0.01, 0.1] subsample = [0.5, 0.7, 1.0] max_depth = [3, 7, 9]	Subsample = 0.7, n_estimators = 100, max_depth = 7, learning_rate = 0.1

Table 6- shows the algorithms, the hyperparameters that were considered and the most optimum hyperparameter for the given algorithm

3.1.4 Best algorithms (Analysis of the Implementation)

After the hyperparameters were tuned, k-fold cross-validation was used to find the best binary classification algorithm for this problem. In this instance, 5-fold cross-validation was used. Table 7 shows the mean accuracy of each algorithm; it shows Stochastic Gradient Boosting is the best algorithms for this problem.

Algorithms	Mean accuracy (to 2 decimal place)
Logistic Regression	93.80%
Ridge Classifier	87.72%
K-Nearest Neighbors (KNN)	94.74%
Support Vector Machine (SVM)	95.16%
Bagged Decision Trees (Bagging)	96.07%
Random Forest	96.21%
Stochastic Gradient Boosting	96.58%

Table 7- shows the list of algorithms and its corresponding mean accuracy.

3.2 Implementation of Eye Blink Rate

To measure the student's engagement by observing their eye-blink rate, a program that can count the number of blinks a student performs within a given time frame is needed. Since there are limited numbers of free software that can identify an eye blinking, the choice is narrowed down to OpenFace. Although OpenFace provides functionality that can identify an eye blinking, upon a conversation with a software creator, it became clear that this functionality has not yet been tested, and its accuracy is unknown [43]. Therefore, not only is there a need to create an eye-blink rate program to measure eye-blink rate for this research and but there is also a need to provide an open-source program that can be used for free by future researchers. This program will be created using OpenCV, python and dlib. While this dissertation is mainly research-based and the program is merely designed to obtain data to answer the research question, it is worth noting the main requirement for the program is to measure eye blink rate in real-time (live videos).

3.2.1 Software created using OpenCV, python and dlib

This eye blink detection software was created using python, OpenCV library, dlib and machine learning code. The software operates by initially extracting frames from the live video in real-time. The extracted frames are then converted into greyscale frames using the function provided by OpenCV; this method of grey scaling frames has been adopted by multiple researchers when observing high-speed vehicles. Since it largely reduces computational complexity for the dlib model and decreases the overall processing time [44] [45], which in this instance is crucial as there is a requirement to track the participant's eyes in real-time. The actual grayscale frame is processed by the pre-trained dlib model in two phases: in phase one, the dlib model locates the participant's face within a given frame; in phase two, the dlib model uses the location of the participant's face to predict the location of the facial landmark of a given interest, which in this case is the participation eyes, this process is shown on figure 1.

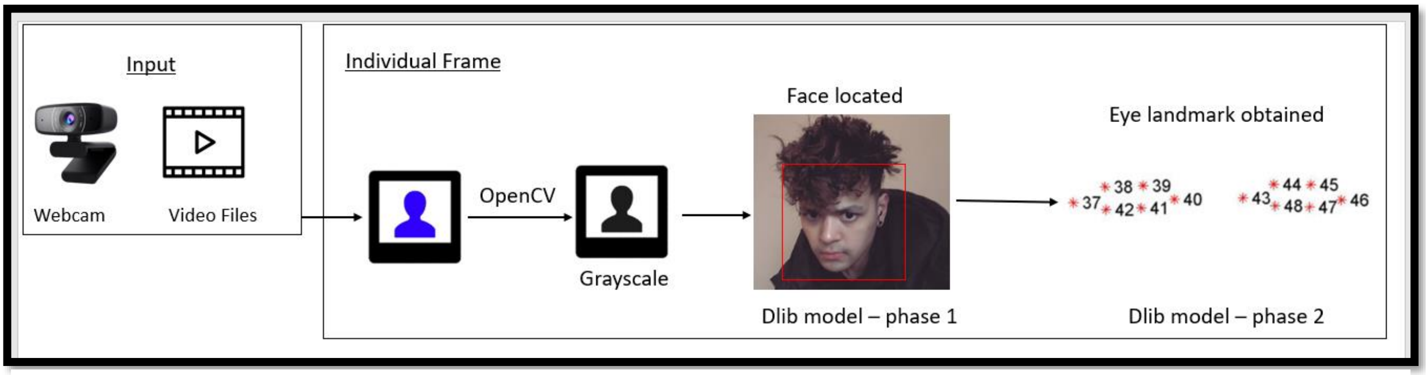

Figure 1- Outlines the core process which are required to otabin eye landmarks.

In phase one, the dlib library provides two methods for detecting and locating faces; the first method involves using a HOG-SVM and the second method involves using a CNN. Since CNN can detect faces at 'odd angles', multiple research has shown Dlib's CNN outperforms Dlib's HOG-SVM method in terms of accuracy when extracting features from an image [46] [47] [48]. While Dlib's CNN requires greater processing power than Dlib's HOG-SVM [49], research had shown Dlib's CNN performs significantly faster than Dlib's HOG-SVM when ran on a system with a dedicated GPU but performs worse when a GPU is not accessible. Due to limited resources available for this research, a system with an available GPU was not accessible for this research; thus, Dlib's HOG-SVM method was used.

In phase two, in order for the dlib model to predict the location of the facial landmark, the model had to be trained. The model was trained using the iBUG 300-W dataset [50]. iBUG 300-W dataset consisted of 600 images with 300 images being taken indoors and 300 taken outdoors, with each image being annotated with 68 landmark points, as shown in figure 2. Since the requirements were to create accurate and 'fast' software, such that it is comparable/superior to OpenFace and can detect eye blink in real-time, only 12 landmark points were considered. All of the 12 landmark points contained the coordinates for the eye. The model was then trained using these 12 landmark points from the dataset. Since it is computationally expensive to tune dlib hyperparameter, a variation of the hyperparameter suggested by the paper [51] was selected.

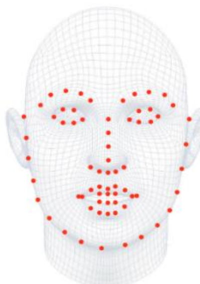

Figure 2- 68 landmark points

3.2.2 Naive approach (Threshold classification)

The 12 landmark points obtained from the dlib model represents both eyes. So one eye is represented by six landmark points on a virtual space, with each landmark points having a unique coordinate, as shown in figure 3. By calculating the distance between the coordinates, a measurement can be obtained to classify the participant's eye state (close or open). A research paper conducted in 2016 [52] proposes finding EAR to determine the state of the participant's eyes. Equation 1 is used to obtain the EAR_{avg}, where pR1, …,pR6 are the right eye landmark points illustrated in figure 3 and pL1, …,pL6 are the left eye landmark points.

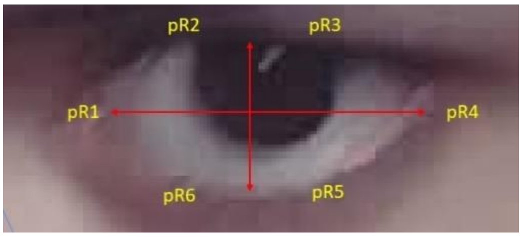

Figure 3- Landmark points for the right eye

$$EAR_{avg} = \frac{\left(\frac{\|pL2-pL6\|+\|pL3-pL5\|}{2\|pL1-pL4\|}\right)\left(\frac{\|pR2-pR6\|+\|pR3-pR5\|}{2\|pR1-pR4\|}\right)}{2} \quad (1)$$

Since EAR_{avg} is the ratio between the vertical distance of the eyelids and the horizontal length of the eye, it is independent of the participant's distance from the camera, making the equation robust. EAR_{avg} is constant when the participant's eyes are open, and EAR_{avg} approaches zero when the participant's eyes are closed. Therefore, using this concept, a manual threshold can be set, such that if the EAR_{avg} is below the given constant, a blink is measured. A research paper [53] suggested that EAR_{avg} for an open eye is 0.339 and 0.141 for a closed eye. Therefore, to provide leeway, a threshold of 0.2 can be set. However, this 'naive approach' of manually setting manual thresholds could potentially be invalid since different participants have different EAR_{avg} thresholds. In addition to this 'naive approach', a 'wise approach' is also proposed.

3.2.3 Wise Approach (SVM classification)

In the wise approach, SVM was used to predict the participant's eye state rather than using a manual threshold. In order to tune the SVM hyperparameter, the EyeBlink8 dataset was used [54]. EyeBlink8 dataset consisted of eight videos with four individuals (1 wearing glasses). While the annotation was provided along with the videos, there were incorrect and unreliable. For instance, one video had an annotation showing that a person had blinked at time = 383 seconds even though the duration of the video was only 374 seconds. Therefore, the dataset had to be manually annotated. Since the dataset has over 70 thousand frames, annotating each frame would be very time-consuming. Therefore, out of the 70 thousand frames, 2 thousand were randomly selected. The randomly selected frames were then manually annotated, and the EAR value was also obtained using a python script; this formed our dataset – one must note that this is indeed a limitation that occurred during this dissertation since only a subset of frames were selected (2 thousand rather than the 70 thousand frames). This dataset was used then to train, test, validate and tune the hyperparameter of the SVM algorithm.

In the SVM algorithm, the 'Gamma' and the 'C' hyperparameter were tuned using grid search along with k-fold cross-validation (where k = 10). 'Gamma' was chosen from the set {"auto",10} such that auto = 1/ n_features, where n_features = 100. 'C' was chosen from the set {0.25 , 0.5 ,1 ,2 ,4 ,8 ,16

,32 ,64 ,128 ,256 ,512 }. The accuracy test was repeated five times, and the overall mean accuracy was reported. Figure 4 shows the average accuracy obtained when 'Gamma' and 'C' were varied.

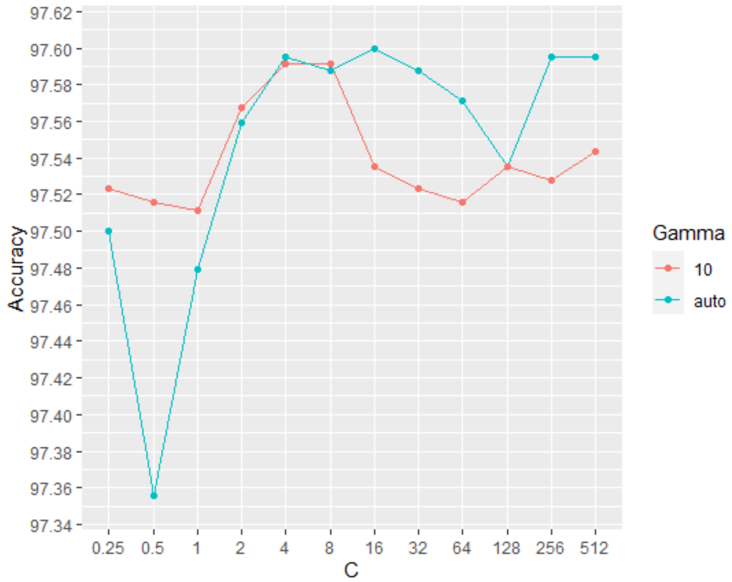

Figure 4- shows the average accuracy obtained when 'Gamma' and 'C' were varied

By observing figure 4, it is apparent that the best accuracy is obtained when 'Gamma' = "auto" and 'C' = 16 (accuracy = 97.60%). However, one must note that varying 'C' had a minimal impact on the accuracy as the difference between the highest and the lowest accuracy was only 0.24%; this shows that the dataset was linearly separable.

3.2.4 Both naïve and wise approaches

In both naïve and wise approaches, OpenCV was also used to highlight the participant's eyes and display the total blink numbers in a window in real-time (i.e show the visual representation of the participant's eyes), as shown in figure 5. False positives can occur in both approaches under certain conditions: for instance, when the participants squint their eyes, a low EAR value will be measured; this can cause the software to misclassify it as a 'blink.' Also, background noises and sudden head movement can all lead to false positive. This was prevented by using the principle of eye blink duration. For example, a blink duration ranges from 100-400 ms [55] , and since the dataset and the recording which will be captured in this research are both 30 FPS, we can state that the blink duration will correspond to 3– 12 frames. Therefore, by discarding any frames successful labelled as 'blink' lasting less than three successive frames or greater than 12 successive frames, the false-positive created by background noises, sudden head movement and participant squinting can be avoided.

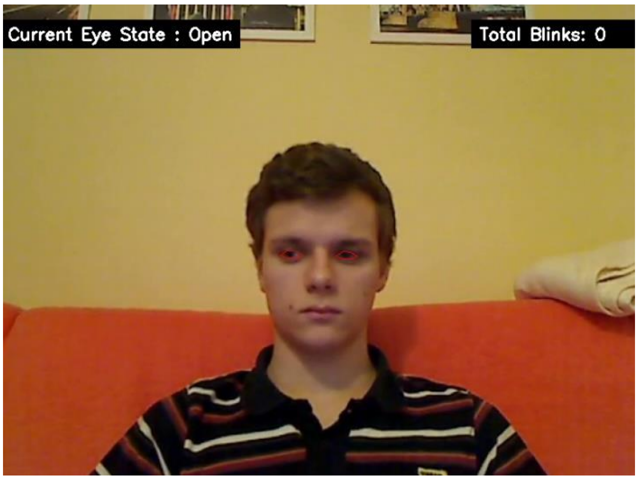

Figure 5- display shown by the software

3.2.5 Comparison (Analysis of the Implementation)

Talking Face dataset [56]was used to test both approaches. It consists of a recording of an individual engaging in a conversation lasting roughly 200 seconds. In this period, the individual performs 61 blinks. 5000 frames were extracted from the dataset and each frame was then manually annotated for this research as either ‘blink’ or ‘open’ to serve as the ground truth. Talking Face dataset [56] was tested using both threshold (naïve) and SVM (wise) classification. The threshold classification detected 54 blinks, whereas the SVM classification only detected 45 blinks. However, this value alone cannot be used to assess the performance of both methods since the blinks could have been measured due to false positives. Thus, a confusion matrix and other performance measuring techniques were used to analyze the data. Figure 6 and 7 shows the confusion matrix for the SVM and the threshold classification respectably.

n= 5000	**Predicted:** Open	**Predicted:** Blink	
Actual: Open	4582 True negative	3 False positive	4585
Actual: Blink	220 False negative	195 True positive	415
	4802	198	

Figure 6- Shows the confusion matrix for SVM classification.

n= 5000	**Predicted:** Open	**Predicted:** Blink	
Actual: Open	4575 True negative	10 False positive	4585
Actual: Blink	105 False negative	310 True positive	415
	4680	320	

Figure 7- Shows the confusion matrix for threshold classification.

Using the figures above, we can obtain further performance measurements, such as accuracy, precision, recall, f1-score. All these measurements are shown in table 8.

Performance measurement	Naïve approach / Threshold classification	Wise approach / SVM classification
Accuracy	97.70%	95.54%
Precision	0.97	0.98
Recall	0.75	0.47
F1 Score	0.85	0.63

Table 8- shows the performance measurement of both classification methods

By observing table 8, it appears that SVM is marginally more precise than the threshold classification. It means for every frame predicted as "blink", SVM classify a greater amount of the frames correctly. However, it also shows that the SVM classification method's recall is significantly worse than the threshold classification method. This means that the SVM classification method classifies a lesser amount of frames as 'blink', i.e. SVM classification is more selective. This could phrase be due to the SVM classification model being trained on a 'strictly' annotated dataset, such that a very low EAR is required for the model to classify a frame as a "blink". The table also shows that the SVM classification method has a significantly lower F1 score than threshold classification due to its low recall value. Overall the SVM classification method performed worse than the threshold classification, this could have potentially occurred since the SVM classification model was trained on a partial dataset - only 2 thousand frames (see section 3.2.3)

Since the threshold classification method is significantly better than SVM classification as it has higher accuracy, recall, and F1 score but a marginally lower precision, the threshold classification method will be used in the experiment.

3.2.6 Classifying engagement using eye blink rate

Initially, to classify engagement level using blink rate, the participant's eye blink rate must be captured to serve as a baseline (see section 4.3). We will be using the data obtained by [34]. Observing their data shows that an individual's eye blink rate decreases by 33% when engaged and increases by 23% when unengaged. Therefore, by using this data, we can form our classification model. This dissertation will provide leeway to account for potential error; hence, a 30% decrease will be classified as engaged. 20% increases will be classified as unengaged. Less than 20% and greater than 30% will be classified as neither engaged nor unengaged. Since the experiments are two minutes long (see section 4.3), the eye blink rate will be measured every 10 seconds; thus, 12 classifications will be obtained, and the average engagement level (in terms of percentage) will be measured.

4 Research Procedure

4.1 Overview of the experiment

Data must be gathered to prove or disprove the hypotheses. Therefore, participants within the age range of 7 to 11 were gathered for the experiment with an incentive of £10 amazon vouchers. Before the participant could begin, they were required to complete the consent form attached in Appendix 1. A total of 9 participants were studied in this experiment, and their age groups are shown in table 9. The overall procedures of the experiment are shown in Figure 8.

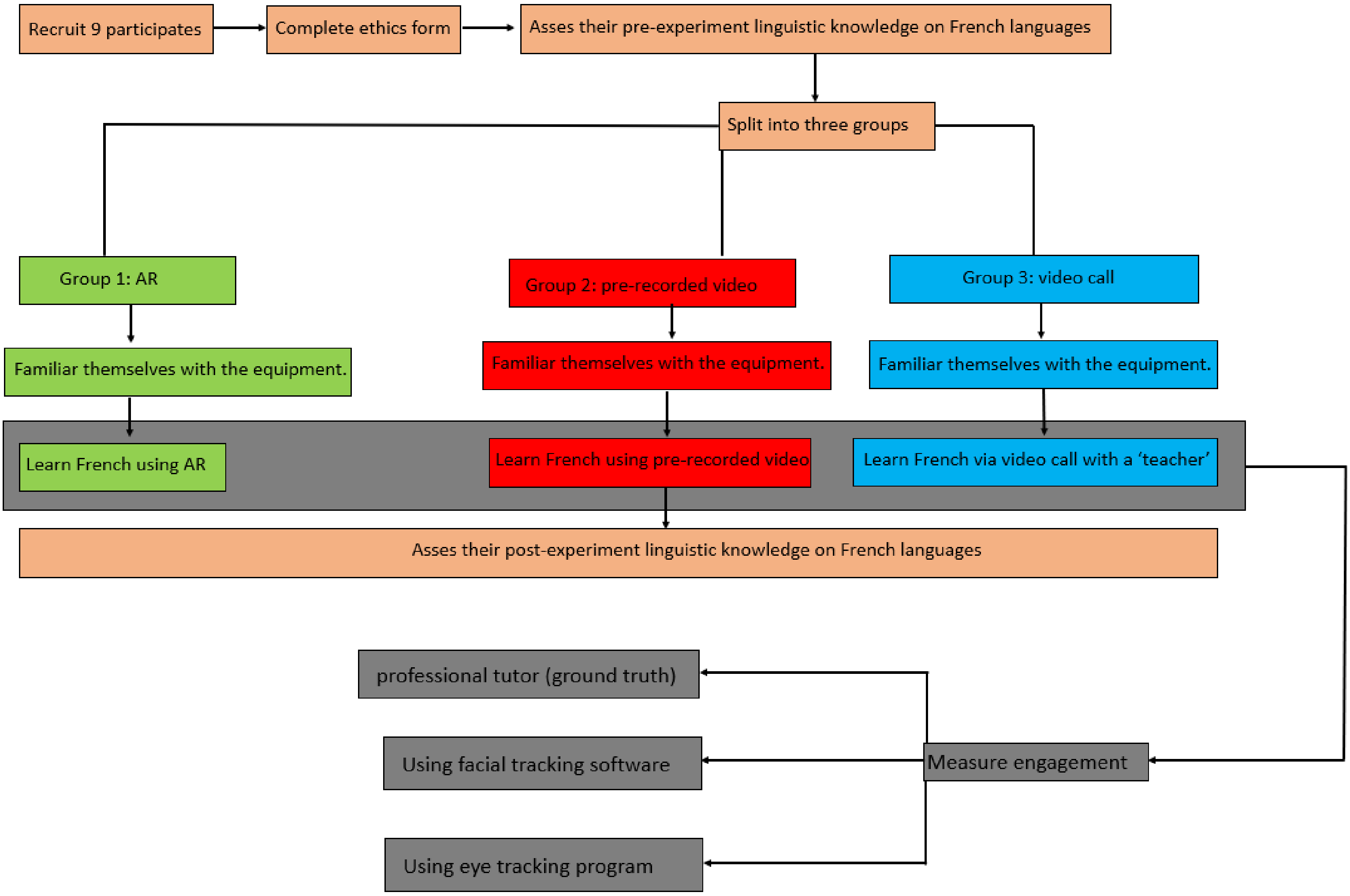

Figure 8- Overall procedure of the experiment

Numbers of children	Age
1	7
1	8
4	9
2	10
2	11

Table 9- shows the list of participants and their age

4.2 Learning

The 9 participants were randomly and equally allocated into three separate groups, with each group learning using different learning methods; the groups were as follows:

- Group 1: Studying using AR.
- Group 2: Studying using pre-recorded videos.
- Group 3: Studying using video conference.

The children were initially taught 12 different French words, all of which were names of farm animals and pets since they are taught to KS2 students in England as a part of their learning curriculum. Table 10 shows the list of the French word and the corresponding English translation. Each word was presented on the screen for exactly 10 seconds to keep the experiment fair for all groups. Usually, when AR applications are used, the users normally walk to another image target in a given space/room. However, to keep the experiment fair and easy to control, the image target was shown on PowerPoint slides, which changed the slides automatically every 10 seconds. For all groups, the word was also pronounced in French. The corresponding English and French translation text were also displayed on an Acer 10” tablet, along with the images. This can be seen in Figures 9 and 10 – a screenshot taken from the pre-recorded video and an image of the AR application in use, respectively. For group 3, a skype video call was used. The ‘teacher’ was in another room and used a ‘share screen’ functionality to show the presentation to the participant. As previously mentioned, in all cases, the display was shown on a tablet. The tablet also had an phone attached at the back, as shown in figure 11, which ran the ‘DroidCam’ application (this allows the phone to operate as a webcam for a computer’).

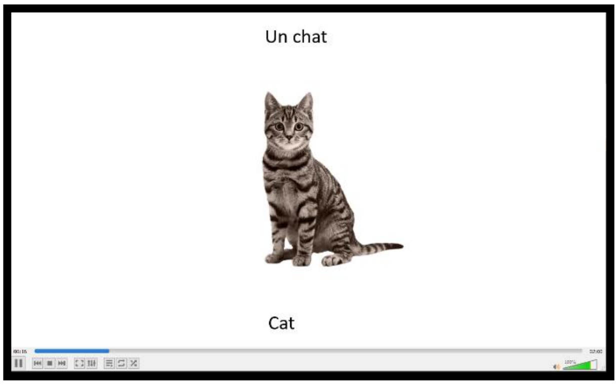

Figure 9- Screenshot of the pre-recoreded video

Figure 10- Image of the AR application in use

French Word	English translation
un singe	Monkey
un chat	Cat
une grenouille	Frog
une araignée	Spider
une poule	Chicken
une vache	Cow
un canard	Duck
un mouton	Sheep
un cochon	Pig
une tortue	Tortoise
une souris	Mouse
un éléphant	Elephant

Table 10- shows the list of farm animals and pet in both French and English

Figure 11-shows the andorid phone (webcam) attached to the tablet

4.3 Obtaining data

The live video from the webcam/phone's camera (via DroidCam' application) was used for two purposes. It was fed into OpenFace to extract the AU, head pose, eye gaze. It was also used by the eye-tracking program to obtain an eye blink rate. One must note that the help of a professional tutor was also used in the experiment to obtain the participant's engagement level to serve as a ground truth. The tutor graded the participant's engagement in terms of a percentage, where 100% refers to a fully engaged and 0% refers to a fully unengaged. The data obtained from both eye and face tracking programs will be tested against this ground truth to grade each engagement tracking method's performance. During the research, the participants were assessed using a simple quiz (see appendix 2) before and after the experiment. The pre-experiment quiz was used to compute their baseline understanding of French words, and the post-experiment quiz was given to compute newly gained knowledge.

5 Results Analysis And Discussion

5.1 Information recall

Figure 12 shows the distribution of the correct answers obtained by each group. For group 1(mean = 80%, SD = 1.53), for group 2 (mean = 53%, SD = 8.19) and group 3 (mean = 64%, SD = 9.29). By observing the data, it is apparent that Group 1 (AR) could recall the majority of the information, followed by group 3 (video calls - *two-way synchronous learning*) and group 2(pre-recorded videos- *one-way synchronous learning*). Contrary to popular belief that 'supervised learning is better than unsupervised learning', this result shows it is not always the case. For instance, the AR group (who were unsupervised) performed 22% better than group 3(who were taught online 1-to-1 and hence were supervised). This can be explained since AR provides two important features required for learning: play and experimentation [57], which in this case has overshadowed even the added benefit that supervised learning can bring. The results also show that a student in group 3 performed significantly worse than 2/3 of students in group 2. This could potentially be an anomaly; however, this cannot be said for certain due to the limited number of participants in the study.

Further, looking into the data collected for each word showed that all groups were able to correctly identify the 'simple' French word such as elephant (éléphant), cat(chat), tortoise(tortue). This is the case since these French and the English corresponding words have similar spellings. However, the majority of the participants failed to correctly identify ‘complex’ words that have entirely different spelling in both French and English i.e. frog (grenouille), spider (araignée), Cow (vache) and Duck (canard), Pig (cochon) and Mouse (souris). With no participants able to identify either frog (grenouille) and Duck (canard) correctly. However, only Group 1 (using AR) were able to identify spider (araignée), Cow (vache) words correctly. This shows learning with AR can help students memorize even ‘complex’ information. Moreover, since Group 1 (using AR) participants were able to correctly recall the most amount of information relative to other groups and were also able to recall even complex words, it proves one of our hypotheses that children retain more information when they learn using AR relative to learning via video calls or pre-recorded videos.

Overall, the results show AR can improve the education sector for children during covid-19 as it can significantly improve student learning/information recall. It also shows that AR does not need ‘complex’ 3D objects/images nor does it require high-tech equipment (such as an AR headset) for it to be a superior form of learning. Instead, it can be simple as a mobile application showing AR objects with text and sounds on a screen via an image target to be more effective than the current e-learning method used in HIC and MIC.

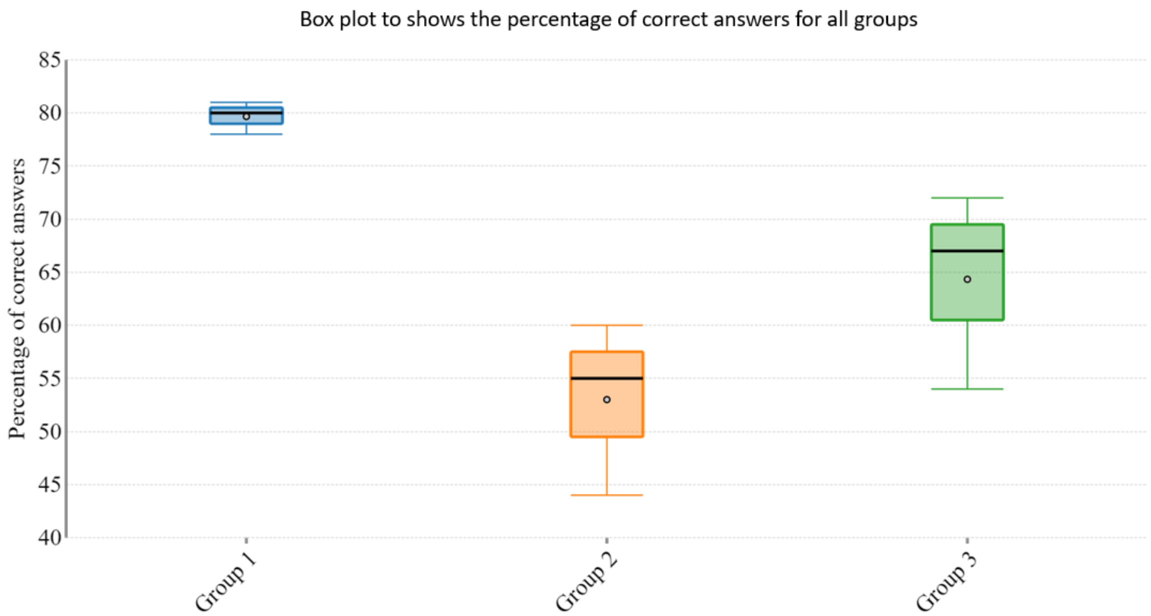

Figure 12- Box plot to shows the percentage of correct answers for all groups

5.2 Relationship between learning and engagement

In this thesis, we will assume/treat *learning* and the number of correct questions answered by the students (their grades) as the same thing. For instance, we can make this assumption since a student who has *learnt* (memorized) all the required information, i.e. all the French words will answer all of the questions correctly. Also, in this section, we will be comparing the ground truth (measured by the tutor) against the student’s grade to find a relationship between learning and engagement. Table 11 shows the student’s grade and their engagement level.

Grade (%)	Engagement level (ground truth) (%)
60	40

54	60
55	40
67	72
81	90
78	95
80	85
72	70
44	50

Table 11- shows the student's grade and their corresponding engagement level.

Pearson correlation coefficient will be used to identify the relationship between learning and engagement. Equation 2 shows the formula which is used to obtain the Pearson correlation coefficient, where *r* is the Pearson correlation coefficient, x_i is the values of the student's grade, $\bar{x}$ is the mean value of all the student's grade, y_i is the values of the student's engagement level, $\bar{y}$ is the mean value of all the student's engagement level.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

For ground truth (r = 0.8543), which shows there is a strong positive correlation between engagement and learning, thus proving our third hypothesis. This contradicts the paper [30], which showed no significant correlation between perceived engagement and learning; since in their data, the value for r was only 0.44. Nevertheless, the fact that our research shows there is a strong correlation between engagement and learning is fascinating as it opens up the possibility of potentially predicting students' grades instantly by simply identifying their engagement level. However as previously stated, only limited numbers of participants were recruited for this experiment hence a further and more thorough investigation should be considered in future research to further this study.

5.3 Measuring engagement using an automated system

Table 12 shows the ground truth and engagement level measured using both automated techniques for all participants.

Participants	Engagement level (ground truth) (%)	Engagement measured using facial tracking (%)	Engagement measured using eye blink rate (%)
Group 1 (AR) – participant 1	90	84.11	67
Group 1 (AR) – participant 2	95	76.31	58
Group 1 (AR) – participant 3	85	76.50	75
Group 2 (pre-recorded video) – participant 1	40	33.39	50
Group 2 (pre-recorded video) – participant 2	50	39.22	42
Group 2 (pre-recorded video) – participant 3	40	37.94	58

Group 3 (video call) – participant 1	60	64.36	42
Group 3 (video call) – participant 2	70	66.56	67
Group 3 (video call) – participant 3	72	56.92	58

Table 12- shows the ground truth and engagement level measured using both techniques for all participants.

To analyze the data in table 12, RMSE is used. Table 12 shows the RMSE for both measuring techniques for each and all groups.

Groups	RMSE – facial tracking	RMSE – eye blink rate
Group 1 (AR)	12.33	22.49
Group 2 (pre-recorded video)	7.40	12.75
Group 3 (video call)	9.28	13.30
All groups (1-3)	9.88	18.30

Table 13- shows the RMSE for both measuring techniques for all groups.

Observing the table shows that both methods performed rather poorly, but it does show that the facial tracking method is significantly better than the eye blink rate method since its RMSE is 60% lesser than that of the eye blink rate method.

The greatest RMSE occurred for group 1; this could have occurred since the participants who were using the AR were surprised by the technology and often made different facial expressions which the machine learning model could have classified incorrectly. The incorrect classification is very probable since the machine learning model was only trained using a subset of the DAISEE dataset, as mentioned in section 3.1.1 and is indeed a limitation of this thesis. The largest RMSE for the eye blink rate method occurred for group 1 (group using AR). This can be explained since during the experiment the participants were rather curious about the AR technology and the AR object/images (farms' animals/pets) displayed on the screen. During the experiment, this curiosity often caused the participants to shift their heads and glance around; this could have potentially led to some blinks not being detected by the program or incorrectly detected due to false positives.

If one were to further classify the engagement data into discrete values – in terms of binary classification, i.e. where ≥ 50% engagement means 'engaged' and < 50% engagement means unengaged, both models performed rather well. To explore this notion, let us convert table 12 into discrete-valued binary classification; the resulting data is shown in table 14.

Participants	Ground truth (%)	Engagement measured using facial tracking (%)	Engagement measured using eye blink rate (%)
Group 1 (AR) – participant 1	Engaged	Engaged	Engaged
Group 1 (AR) – participant 2	Engaged	Engaged	Engaged
Group 1 (AR) – participant 3	Engaged	Engaged	Engaged
Group 2 (pre-recorded video) – participant 1	Unengaged	Unengaged	Engaged
Group 2 (pre-recorded video) – participant 2	Engaged	Unengaged	Unengaged

Group 2 (pre-recorded video) – participant 3	Unengaged	Unengaged	Engaged
Group 3 (video call) – participant 1	Engaged	Engaged	Unengaged
Group 3 (video call) – participant 2	Engaged	Engaged	Engaged
Group 3 (video call) – participant 3	Engaged	Engaged	Engaged

Table 14- shows the ground truth and engagement classification measured using both techniques for all participants.

Confusion matrix and other performance measuring techniques can be used to analyze the data in table 14. Figure 13 and 14 shows the confusion matrix for the facial tracking and eye blink rate methods respectably.

n= 9	**Predicted:** Unengaged	**Predicted:** Engaged	
Actual: Unengaged	2 True negative	0 False positive	2
Actual: Engaged	1 False negative	6 True positive	7
	3	6	

Figure 13- shows the confusion matrix for the facial tarcking method

n= 9	**Predicted:** Unengaged	**Predicted:** Engaged	
Actual: Unengaged	0 True negative	2 False positive	2
Actual: Engaged	2 False negative	5 True positive	7
	2	7	

Figure 14- shows the confusion matrix for the eye blink rate method

Using the figures above, we can obtain further performance measurements, such as accuracy, precision, recall, f1-score. All these measurements are shown in table 16.

Performance measurement	Facial tracking classification	Eye-blink rate classification
Accuracy	88.89%	55.56%
Precision	1	0.71
Recall	0.86	0.71
F1 Score	0.92	0.71

Table 16- shows the performance measurement of both engagement classification methods

By observing table 16, it appears that classifying engagement using facial tracking is more precise than classifying using eye blink rate; in fact, the results show that facial tracking has a precision of 1, meaning it produces no false positives. However, the same cannot be said for false-negative since both methods do not have a recall value of 1. Nevertheless, Table 16 shows that the facial tracking method also has a higher recall value and F1 score and is significantly more accurate than the eye blink rate method, further disproving our fourth hypothesis. One must note that accuracy of 88.89% is significant great considering the limited dataset used. With a larger dataset, the accuracy, precision, recall, etc., can be significantly improved and should be considered in future research. However, the same cannot be said about predicting the actual engagement level (i.e out of 100%) due to the high RMSE.

6 Personal reflection

The workload set out for this project turned out to be larger than I expected due to many reasons. One of the major reasons was due to the second lockdown, this lockdown meant that I was unable to use the AR headset and other AR equipment provided by the mixed-reality department which I was initially anticipating to incorporate into this dissertation. This meant that the experiment had to be adapted to meet the second lockdown rule's requirements. It also meant that I was unable to use devices provided by the computer science department to accurately capture participants faces / their engagement levels. Rather, this meant that the initial plan for the dissertation had to be massively altered such that facial tracking classification models and eye-blink rate programs had to be implemented and their implementation had to be analysed. Both aspects surrounding facial tracking programs and eye-blink rate programs were also rather difficult to implement as I initially lacked knowledge surrounding this topic. The lack of knowledge stems from the fact that I had not enrolled on machine learning, computer vision or mixed reality module as was not aware that it would be required for my dissertation at that given time. Learning about Vuforia, Unity, C# was also a challenge and required more time than expected but was indeed an enjoyable experience and was rather rewarding. Overall, this dissertation has allowed me to explore a wide ranges of topics within computer science and has indeed provided me with great insight on which topics/job prospects I would like to pursue after graduation.

7 Reference

[1] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes and R. F. Garry, "The proximal origin of SARS-CoV-2," *Nature Medicine,* vol. 26, pp. 450-452, 2020.

[2] T. Phillips, "Global report: coronavirus death toll reaches 2 million," The Guardian, 16 January 2021. [Online]. Available: https://www.theguardian.com/world/2021/jan/15/global-coronavirus-death-toll-reaches-2-million-people. [Accessed 21 January 2021].

[3] J. Miks and J. McIlwaine, "Keeping the world's children learning through COVID-19," Unicef, 20 April 2020. [Online]. Available: https://www.unicef.org/coronavirus/keeping-worlds-children-learning-through-covid-19. [Accessed 19 January 2021].

[4] D. Clark, "Psychological myths in e-learning," *Medical Teacher,* vol. 24, no. 6, p. 598–604, 2002.

[5] M. Rostaminezhad, N. Mozayani and M. I. D. Norozi, "Factors Related to E-learner Dropout: Case Study of IUST Elearning Center," *Procedia - Social and Behavioral Sciences,* vol. 83, pp. 522-527, 2013.

[6] M. Sirakaya and B. K. Çakmak, "Investigating Student Attitudes toward Augmented Reality," *Malaysian Online Journal of Educational Technology,* vol. 6, no. 1, pp. 30-44, Technology2018.

[7] S. Cai, F. Chiang, Y. Sun, C. Lin and J. J. Lee, "Applications of augmented reality-based natural interactive learning in magnetic field instruction.," *Interactive Learning Environments,* vol. 25, no. 6, pp. 778-791, 2017.

[8] A. B. Tomi and D. R. A. Rambli, "An Interactive Mobile Augmented Reality Magical Playbook: Learning Number with the Thirsty Crow," *Procedia Computer Science,* vol. 25, pp. 123-130, 2013.

[9] C.-C. Tsai and K.-H. Cheng, "Affordances of Augmented Reality in Science Learning: Suggestions for Future Research," *Journal of Science Education and Technology,* vol. 22, no. 4, p. 449–462, 2012.

[10] S. Yuen, G. Yaotuneyong and E. Johnson, "Augmented reality: An overview and five directions for AR in education," *Journal of Educational Technology Development and Exchange,* vol. 4, no. 1, pp. 119-140, 2011.

[11] S. Cai, F. K. Chiang and X. Wang, "Using the Augmented Reality 3D Technique for a Convex Imaging Experiment in a Physics Course," *International Journal of Engineering Education,* vol. 23, pp. 856-865, 2013.

[12] N. Bosch, "Detecting Student Engagement: Human Versus Machine," *ACM Conference on User Modeling, Adaptation and Personalization,* vol. 24, p. 317–320, 2016.

[13] W. Boykin and P. Noguera, Creating the Opportunity to Learn, Alexandria: Association for Supervision and Curriculum Development, 2011.

[14] E. Cappella, H. Y. Kim, J. W. Neal and D. R. Jackson, "Classroom Peer Relationships and Behavioral Engagement in Elementary School: The Role of Social Network Equity," *American journal of community psychology,* vol. 52, p. 367–379, 2013.

[15] S. Büchele, "Evaluating the link between attendance and performance in higher education: the role of classroom engagement dimensions," *Assessment & Evaluation in Higher Education,* vol. 46, no. 1, pp. 132-150, 2021.

[16] P. Ekman, “Facial Action Coding System,” Paul Ekman Group, 2021 . [Online]. Available: https://www.paulekman.com/facial-action-coding-system/. [Accessed 19 03 2021].

[17] I. E. Buchan, “P Values(calculated probability) hypothesis testing - statsdirect,” tatsdirect, [Online]. Available: https://www.statsdirect.co.uk/help/basics/p_values.htm#:~:text=The%20P%20value%2C%20or%20calculated,the%20hypothesis%20is%20being%20tested.. [Accessed 20 March 2021].

[18] D. E. King, “dlib C++ Library,” Dlib, 28 March 2021. [Online]. Available: http://dlib.net/. [Accessed 20 March 2021].

[19] I. Grahn, “The Vuforia SDK and Unity 3D Game Engine: Evaluating Performance on Android Devices,” 2017.

[20] D. R. R. Syita Ulfah1, U. Fatoni, K. Mukhtar, H. Tajiri and A. Sarbini, “Augmented reality using Natural Feature Tracking(NFT) method for learning media of makharijulhuruf,” *IOP Conference Series: Materials Science and Engineering,* vol. 874, p. 012019, 2020.

[21] “AUGMENTED REALITY MARKER GENERATOR - Brosvision,” Brosvision, 2013. [Online]. Available: https://www.brosvision.com/ar-marker-generator/. [Accessed 01 March 2021].

[22] M. Georgescu, “The Future of E-Learning: Designing Tomorrow’s Education,” *Social Science Research Network,* 2006.

[23] S. S. Noesgaard and R. Ørngreen, “The effectiveness of e-learning: An explorative and integrative review of the definitions, methodologies and factors that promote e-Learning effectiveness,” *Electronic journal of e learning,* vol. 13, no. 4, pp. 278-290, 2015.

[24] L. B. Kiat, M. B. Ali, N. D. A. Halim and H. B. Ibrahim, “Augmented Reality, Virtual Learning Environment and Mobile Learning in education: A comparison,” *IEEE Conference on e-Learning, e-Management and e-Services (IC3e),* pp. 23-28, 2016 .

[25] Á. D. Serio, M. B. Ibáñez and C. D. Kloos, “Impact of an augmented reality system on students’ motivation for a visual art course,” *Computers & Education,* vol. 68, pp. 586-596, 2013.

[26] M. Moore, “Three types of interaction,” *The American Journal of Distance Education,* vol. 3, no. 2, pp. 1-6, 1989.

[27] B. Dalgarno, “A classification scheme for learner-computer interaction,” *Annual conference of the australasian society for computers in learning in tertiary education (ASCILITE),* pp. 240-248, 2004.

[28] U. Neumann and A. E. Majoros, “Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance,” *Proceedings. IEEE 1998 Virtual Reality Annual International Symposium,* pp. 4-11, 1998.

[29] C. Thomas and D. B. Jayagopi, “Predicting Student Engagement in Classrooms Using Facial Behavioral Cues,” *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education,* p. 33–40, 2017.

[30] j. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster and J. R. Movellan, “The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions,” *IEEE Transactions on Affective Computing,* vol. 5, no. 1, pp. 86-98, 2014.

[31] Y. Cho, “Rethinking Eye-Blink: Assessing Task Difficulty through Physiological Representation of Spontaneous Blinking,” *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems ACM,* no. 21, p. 1–12, 2021.

[32] N. Nourbakhsh, Y. Wang and F. Chen, “GSR and Blink Features for Cognitive Load Classification,” *IFIP International Federation for Information Processing,* pp. 159-166, 2013.

[33] R. Paprocki and A. Lenskiy, “What Does Eye-Blink Rate Variability Dynamics Tell Us About Cognitive Performance?,” *Frontiers in Human Neuroscience,* vol. 11, p. 620, 2017.

[34] T. Sakai, H. Tamaki, Y. Ota, R. Egusa, S. Inagaki, F. Kusunoki, M. Sugimoto and H. Mizoguchi, “Eda-BasedEstimation Of Visual Attention By Observation Of Eye Blink Frequency,” *The International Journal on Smart Sensing and Intelligent Systems (S2IS),* vol. 10, p. 296–307, 2017.

[35] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos and M. Pantic, “The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results,” *EEE International Conference on Computer Vision Workshop (ICCVW),* pp. 1003-1011, 2015 .

[36] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1867-1874, 2014.

[37] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang and Y. Sheikh, “Supervision-by-Registration: An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),,* pp. 360-368, 2018.

[38] C. Ranti, W. Jones, A. Klin and S. Shultz, “Blink Rate Patterns Provide a Reliable Measure of Individual Engagement with Scene Content,” *Sci Rep,* vol. 10, p. 8267 , 2020.

[39] “KeenTools FaceTracker,” KeenTools, [Online]. Available: https://keentools.io/products/facetracker. [Accessed 20 March 2021].

[40] A. Asthana and S. Zafeiriou, “Chehra,” Google Sites, 28 September 2016. [Online]. Available: https://sites.google.com/site/chehrahome/. [Accessed 20 March 2021].

[41] G. Hidalgo, “GitHub - CMU-Perceptual-Computing-Lab,” GitHub , 19 April 2021. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose. [Accessed 20 April 2021].

[42] A. Gupta, A. D’Cunha, K. Awasthi and V. Balasubramanian, “DAiSEE: Towards User Engagement Recognition in the Wild,” *Computer Vision and Pattern Recognition,* vol. 14, no. 8, 2016.

[43] G. D. Leo and F. Sardanelli, “Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach,” *European Radiology Experimental,* vol. 4, no. 18, p. 1–8, 2020.

[44] T. Baltrusaitis, “Gaze analysis heatmap and blink rate #229,” Github, 31 July 2017. [Online]. Available: https://github.com/TadasBaltrusaitis/OpenFace/issues/229. [Accessed 20 March 2020].

[45] S. Kamoji, D. Koshti, A. Dmonte, S. J. George and C. S. Pereira, “Image Processing based Vehicle Identification and Speed Measurement,” *2020 International Conference on Inventive Computation Technologies (ICICT),* pp. 523-527, 2020.

[46] S. S. Harsha and K. R. Anne, “A Highly Robust Vehicle Detection Tracking and Speed Measurement Model for Intelligent Transport Systems,” *International Journal of Applied Engineering Research,* vol. 11, no. 5, pp. 3733-3742, 2016.

[47] M. F. Aslan, A. Durdu, K. Sabanci and M. A. Mutluer, “CNN and HOG based comparison study for complete occlusion handling in human tracking,” *Measurement,* vol. 158, 2020.

[48] M. P. Jacob, “COMPARISON OF POPULAR FACE DETECTION AND RECOGNITION TECHNIQUES,” *International Research Journal of Modernization in Engineering Technology and Science,* vol. 03, no. 04, pp. 1304-1313, 2021.

[49] R. G. Garcia, B. Beltran, D. Vilarino and R. Martinez, “Facial Recognition using Convolutional NeuralNetworks and Supervised Few-Shot Learning,” *POLIBITS,* vol. 60, p. 67–72, 2019.

[50] Y.-S. Su, H.-Y. Suen and K.-E. Hung, “Predicting behavioral competencies automatically from facial expressions in real-time video-recorded interviews,” *Journal of Real-Time Image Processing,* 2021.

[51] C. Sagonasa, E. Antonakosa, G. Tzimiropoulosb, S. Zafeirioua and M. Pantic, “300 Faces In-The-Wild Challenge: database and results,” *Image and Vision Computing,* vol. 47, pp. 3-18, 2016.

[52] V. Kazemi and J. Sullivan, “One Millisecond Face Alignment with an Ensemble of Regression Trees,” *IEEE Computer Society,* pp. 1867-1874, 2014.

[53] T. Soukupová and J. Cech, “Real-Time Eye Blink Detection using Facial Landmarks,” *21st Computer Vision Winter Workshop,* p. 1–8, 2016.

[54] N. Kamarudin, N. A. Jumadi, N. L. Mun1, N. C. Keat, A. H. K. Ching, W. M. H. W. Mahmud1, M. Morsin and F. Mahmud, “mplementation of Haar Cascade Classifier and Eye Aspect Ratio for Driver Drowsiness Detection Using Raspberry Pi,” *Universal Journal of Electrical and Electronic Engineering,* vol. 6, pp. 67-75, 2019.

[55] A. Fogelton, “eyeblink,” blinkingmatters, [Online]. Available: https://www.blinkingmatters.com/research. [Accessed 21 March 2021].

[56] E. Langbehn, F. Steinicke, M. Lappe, G. F. Welch and G. Bruder, “In the blink of an eye: leveraging blink-induced suppression for imperceptible position and orientation redirection in virtual reality,” *ACM Transactions on Graphics,* vol. 37, no. 4, pp. 66-77, 2018.

[57] "Talking Face Video," [Online]. Available: https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html. [Accessed 20 March 2021].

[58] C. Daiute, "Play as thought: Thinking strategies of young writers," *Harvard Educational Review,* vol. 59, no. 1, pp. 1-23, 1989.

[59] A. Desvars-Larrive, E. Dervic and N. Haug, "A structured open dataset of government interventions in response to COVID-19," *Sci Data,* vol. 7, p. 285 , 2020.

Appendix 1

CONSENT FORM

University of Nottingham
UK | CHINA | MALAYSIA

Date: [Insert date]

Project: [Insert name]

School of Computer Science Ethics Reference: [Insert ref number]

Please tick the appropriate boxes	**Yes**	**No**
1. Taking part in the study		
a) I have read and understood the project information sheet dated [21/02/2021], ☐ or it has been read to me. I have been able to ask questions about the study and my questions have been answered satisfactorily.	☐	
b) I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
c) I understand that taking part in the study requires me to provide data and that this will involve [complete an anonymous quiz before, shortly after and a week after the use of mobile applications]	☐	☐
2. Use of my data in the study		
a) I understand that data which can identify me will not be shared beyond the project team.	☐	☐
b) I agree that the data provided by me may be used for the following purposes:		
– Presentation and discussion of the project and its results in research activities (e.g., in supervision sessions, project meetings, conferences).	☐	☐
– Publications and reports describing the project and its results.	☐	☐
– Dissemination of the project and its results, including publication of data on web pages and databases.	☐	☐
c) I give permission for my words to be quoted for the purposes described above.	☐	☐
d) I give permission for my visual image contained in photos or video gathered during the research to be used for the purposes described above.	☐	☐

Please tick the appropriate boxes	Yes	No

3. Reuse of my data

a) I give permission for the data that I provide to be reused for the sole purposes of future research and learning.	☐	☐
b) I understand and agree that this may involve depositing my data in a data repository, which may be accessed by other researchers	☐	☐

4. Security of my data

a) I understand that safeguards will be put in place to protect my identity and my data during the research, and if my data is kept for future use.	☐	☐
b) I confirm that a written copy of these safeguards has been given to me in the University's privacy notice, and that they have been described to me and are acceptable to me.	☐	☐
c) I understand that no computer system is completely secure and that there is a risk that a third party could obtain a copy of my data.	☐	☐

5. Copyright

a) I give permission for data gathered during this project to be used, copied, excerpted, annotated, displayed and distributed for the purposes to which I have consented.	☐	☐
b) I wish to be publicly identified as the creator of the following works [answers to the quizzes]	☐	☐

6. Signatures (sign as appropriate)

______________________ ______________________ __________

Name of participant (IN CAPITALS) Signature Date

If applicable:

For participants unable to sign their name, mark the box instead of signing

I have witnessed the accurate reading of the consent form with the participant and the individual has had the opportunity to ask questions. I confirm that the individual has given consent freely.

______________________ ______________________ __________

Name of witness (IN CAPITALS) Signature Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

______________________ ______________________ __________

Name of researcher (IN CAPITALS) Signature Date

7. Researcher's contact details

Name: Jiwa Chhetri

Phone: 07767133181

Email: psyjc9@nottingham.ac.uk

Provide the participant with a copy of the completed form either by email or hard copy as they prefer.

Appendix 2

Draw a line to match the English word (on the left) to the French word (on the right)

Monkey	une poule
Cat	un mouton
Frog	un singe
Spider	une tortue
Chicken	une grenouille
Cow	une araignée
Duck	un canard
Sheep	un éléphant
Pig	une souris
Tortoise	un chat
Mouse	un cochon
Elephant	une vache