



THE HONG KONG
POLYTECHNIC UNIVERSITY

香港理工大學

Pao Yue-kong Library

包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

<http://www.lib.polyu.edu.hk>

**UNDERSTANDING USER ENGAGEMENT LEVEL
DURING TASKS VIA FACIAL RESPONSES, EYE GAZE
AND MOUSE MOVEMENTS**

KWOK CHO KI

M.Phil

The Hong Kong Polytechnic University

2018

The Hong Kong Polytechnic University

Department of Computing

**Understanding User Engagement Level during Tasks
via Facial Responses, Eye Gaze and Mouse Movements**

Kwok Cho Ki

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Philosophy

August 2017

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Cho Ki Kwok (Name of student)

Abstract

User engagement refers to the quality of the user experience (UX) on a particular task or interface. It emphasizes the positive aspects of human and computer interaction, and the desire to work on the same task longer and repeatedly [10]. Users spend time, emotion, attention and effort when they interact with technologies, and a successful application or task should be able to engage users, instead of simply being a “job” that needs to be completed. User engagement is therefore a complex phenomenon that encompasses three different dimensions: (1) cognitive engagement, (2) emotional engagement and (3) behavioral engagement. Researchers use different ways to measure user engagement level, such as self-reporting (e.g. questionnaires), observations (e.g. speech analysis, facial expression analysis) and web analytics (e.g. click-through rate, number of site visits, time spent).

Nowadays, computers are equipped with high computational power and different kinds of sensors, which make possible automated human affect and mental state detection in a variety of situations. Using computers to “observe” human behaviors and using the observed information to detect levels of engagement could be useful in many situations, such as getting feedback for interface improvement or assuring quality of work generated by online workers (crowdsourcing) or students (e-learning). Therefore, there has been much previous work in detecting user engagement through various means such as facial expression, mouse movement or gaze movement. However, this work is hampered by three main challenges: (1) the constraints caused by using intrusive devices, (2) limitations of specific tasks (like gaming) which may produce user behavior different from daily computer usage, (3) and incomprehensive ground truth as collected by straightforward and direct survey questionnaires that capture users’ self-reported numeric level of engagement, which may not cover the three dimensions of engagement.

The work presented in this thesis focuses on non-intrusive visual cues, in particular, visual cues from facial expressions, eye gaze, and mouse cursor signals, for understanding users’ level of engagement in human-computer interaction task. Addressing the first two limitations mentioned above, we conducted experiments and studied users’ facial responses, eye gaze and mouse behaviors related to the change of engagement level during doing Language Learning tasks and Web Searching tasks. Non-intrusive devices, such as the mouse, Tobii eye tracker and off-the-shelf webcam, are used to capture users’ behaviors in the experiment. By using Pearson’s Correlation, Paired T-Test and single factor one way ANOVA, we select a useful feature set from the initial feature set. From the investigation, we have a better understanding of the

relationship between engagement level and user behavior. For example, the facial action unit 5 (“upper lid raiser”) is useful in engagement detection. We observed that this feature is indicative as sleepy users try to keep their eyes open to avoid falling asleep.

To address the third constraints, we collected an engagement dataset that includes a multi-dimension measurement of ground truth. It includes the User Engagement Scale (UES) [89], which covers the three dimensions of user engagement, as the self-reporting tool and the average UES scores can reliably represent the engagement level. It also includes the commonly-used NASA Task Load Index (NASA-TLX) annotations for measuring the cognitive work load. We include a further investigation into the correlation between the UES and TLX sub-scale scores.

We analyze facial affect in two ways. First, we measure momentary affect through the facial action units in every frame of the facial response videos. We then move to an overall affect measurement through segment-based facial features to seek more representative features that cover the whole task period.

The facial affect recognition model was extended into a real life application to identify video viewers’ emotion. We developed an asynchronous video-sharing platform with Emotars, which allow users to share their affects and experience with others without disclosing their real facial expressions and/or features. We analyze the user experience of using this platform in four different dimensions, including emotion awareness, engagement, comfortableness and relationship.

For eye gaze and mouse interaction, we make use of non-intrusive devices, i.e. mouse, Tobii eye tracker and off-the-shelf webcam, to collect eye and mouse interaction data. We investigated using mouse features for user intention prediction, or, in other words, predicting the next type of mouse interaction event. Results show that the mouse interaction features are representative of users’ behavior.

Finally, we group the type of features into three different groups according to the means of data collection: (1) webcam-based features, (2) Eye Tracker-Captured features, and (3) mouse cursor-based features. The performances of different combinations of modalities were evaluated. We apply machine learning techniques to build up user-independent models for both Language Learning tasks and Web Searching tasks separately. The findings suggest that the multimodal approach outperforms unimodal approaches in our studies. Evaluation results also demonstrate the versatility of our feature set, as it achieves reasonable performances of engagement detection in different tasks.

List of Publication

- [1] **T. C. K. Kwok**, C. . N. Shum, G. Ngai, H. V. Leong, G. A. Tseng, H. Choi, K. Mak, and C.-W. Do, “Democratizing Optometric Care: A Vision-Based, Data-Driven Approach to Automatic Refractive Error Measurement for Vision Screening.,” in *2015 IEEE International Symposium on Multimedia (ISM)*, 2015, pp. 7–12.
- [2] **T. C. K. Kwok**, M. X. Huang, W. C. Tam, and G. Ngai, “Emotar: Communicating Feelings through Video Sharing,” in *IUI '15 Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015, pp. 374–378.
- [3] **T. C. K. Kwok**, E. Y. Fu, Y. E. Wu, M. X. Huang, G. Ngai, and H. V. Leong, “Ev’ry Little Movement Has a Meaning of Its Own: Using Past Mouse Movements to Predict the Next Interaction,” in *IUI '18 Proceedings of the 23th International Conference on Intelligent User Interfaces*, 2018. – **In Submission**
- [4] **T. C. K. Kwok**, Y. E. Wu, G. Ngai, S. C. F. Chan, H. V. Leong, and C.-W. Do, “Stereotypes, Perceptions and Behavior of Lay Users towards Self-Diagnostic Medical Applications,” in *CHI '18 Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018. – **Accepted**
- [5] M. X. Huang, **T. C. K. Kwok**, G. Ngai, H. V. Leong, and S. C. F. Chan, “Building a Self-Learning Eye Gaze Model from User Interaction Data,” *Proc. ACM Int. Conf. Multimed. - MM '14*, pp. 1017–1020, 2014.
- [6] M. X. Huang, **T. C. K. Kwok**, G. Ngai, S. C. F. Chan, and H. V. Leong, “Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions,” in *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5169–5179. - **Best Paper Award**
- [7] Y. Fu, **T. C. K. Kwok**, Y. E. Wu, H. V. Leong, and G. Ngai, “Your Mouse Reveals Your Next Activity : Towards Predicting User Intention from Mouse Interaction,” in *COMPSAC '17 IEEE Computer Software and Applications Conference*.

Acknowledgements

I am extremely grateful and remained indebted to my supervisor, Dr. Grace Ngai, for her full support and expert guidance. Without the encouragement, constructive criticism and helpful advice from her, my thesis work would have been an overwhelming and frustrating pursuit.

I would also like to thank the professors in my research group: Dr. Hong-Va Leong, and Dr. Stephen C.F Chan, who patiently support my work through instructional discussions, detailed analyses and continuous suggestions. Their constant sources of knowledge and inspiration provide invaluable guidance through my study.

I am also deeply grateful to Dr. Chi-Wai Do for generously sharing the knowledge in optometry and providing supports to the cross-disciplinary research.

I have had great pleasure working with members in CHILab: Dr. Michael Xueling Huang, Jiajia Li, Eugene Fu, Andy Tam and You Wu. The creativity of all my colleagues has been a constant inspiration throughout my time.

Finally, I would like to acknowledge my family who all kept me going and this thesis would not have been possible without them.

Table of Contents

Certificate of Originality.....	iii
Abstract.....	iv
List of Publication.....	vi
Acknowledgements.....	vii
Table of Contents	viii
List of Figures	xi
List of Tables.....	xii
Chapter 1 Introduction.....	1
1.1 Background and Motivation	3
1.1.1 Understanding User Engagement	3
1.1.2 Engagement Detection via Facial Expressions	4
1.1.3 Engagement Detection via Eye Movements	5
1.1.4 Engagement Detection via Mouse Behavior.....	6
1.2 Study Overview	7
1.2.1 Engagement Dataset Collection.....	7
1.2.2 Engagement Detection with Different Modalities	8
1.2.3 Multimodal Engagement Detection	9
1.3 Thesis Aims and Outline.....	9
Chapter 2 Literature Review.....	12
2.1 User Engagement.....	12
2.2 Facial Expression Analysis	14
2.3 Eye Movement Analysis	15
2.4 Mouse Movement Analysis.....	16
Chapter 3 Engagement Datasets	18
3.1 Language Learning Tasks	19
3.1.1 Scenario 1 – Doing Homework.....	20
3.1.2 Scenario 2 – Working under Distracting Situations.....	21
3.2 Web Searching Task.....	22
3.3 Annotating the Gold Standard	23
3.3.1 Metrics and Measurements	23
3.3.2 Objectivity of Gold Standard Labels of the Dataset	24
3.3.3 Between-Metrics Correlations	25
3.3.4 Implications from Analyzing Results of Two Scenarios.....	26
Chapter 4 Understanding Users from Facial Expression.....	29
4.1 Facial Action Units as Features.....	30
4.1.1 Extract Action Units from Landmarks.....	30

4.1.2	Extract Action Units from OpenFace.....	31
4.2	From a Local Snapshot to a Global Time Segment	31
4.2.1	Action Units as Features	32
4.2.2	Features Selection	33
4.3	Engagement Detection with Facial Expression	37
4.4	Extending Facial Affect Recognition into Real Life.....	39
4.4.1	Asynchronous Video-Sharing Platform with Emotars.....	40
4.4.2	Evaluation	41
4.5	Summary	43
Chapter 5	Understanding Users from Eye Movement.....	44
5.1	Introduction.....	44
5.2	Types of Eye Movements.....	45
5.3	Tobii Features.....	49
5.3.1	Eye Gaze Location and AOIs.....	50
5.3.2	Eye Movements Behaviors and Voting Experts.....	51
5.3.3	Statistics of Eye Movements.....	52
5.4	Eye Behaviors from Webcam	53
5.4.1	Webcam Based Eye Gaze Features	54
5.5	Using Eye Gaze Interaction for Engagement Detection	56
5.5.1	Features Selection	56
5.5.2	Results of Eye Tracker-Captured Eye Gaze Features	62
5.5.3	Results of Webcam-based Eye Gaze Features	64
5.5.4	Results on Using All Selected Eye Gaze Features.....	65
5.6	Summary	65
Chapter 6	Understanding Users from Mouse Movements	66
6.1	Features Extraction from Mouse.....	67
6.1.1	Definition of Mouse Movements in a Segment	67
6.1.2	Statistical Mouse Features	68
6.1.3	Mouse and AOIs	69
6.2	Predict User Intention from Mouse Interaction	70
6.2.1	Dataset.....	71
6.2.2	User Intention in the Tasks.....	74
6.2.3	Models and Results	76
6.2.4	Conclusion on User Intention Prediction Work	77
6.3	Mouse Movements for Engagement Detection.....	79
6.3.1	Feature Selection.....	79
6.3.2	Results.....	82
6.4	Summary	83

Chapter 7	Extending Multi-Modality Engagement Detection into Real Life	84
7.1	Results of Multi Modalities	84
7.2	Summary	87
Chapter 8	Limitations and Future Work	89
8.1	Use of Deep-learning Techniques	89
8.2	Process of Features Selection.....	89
8.3	Calibration of Individual Persons Data.....	89
8.4	Recognition of Engagement Level in Three Dimensions	90
8.5	User Engagement in Mobile Contexts	90
8.6	Extension on Video-Sharing Platform	90
Chapter 9	Conclusion	91
9.1	Other Relevant Contributions	93
9.1.1	Using Interaction Data to build Gaze Model	93
9.1.2	PACE – Personalized, Auto-Calibrating Eye Tracker.....	93
9.1.3	Photorefraction.....	94
References	96

List of Figures

Figure 3-1 Flow of Language Learning Language Learning Task Experiment.....	19
Figure 3-2 User Interface for the Language Learning Task.....	20
Figure 3-3 Instruction slide for Language Learning Task - Scenario 1	20
Figure 3-4 Instruction slide for Language Learning Task - Scenario 2	21
Figure 3-5 Interface of Web Searching Task for displaying and answer questions.	22
Figure 3-6 Captured screen of doing Web Searching task by one of the subject.....	22
Figure 3-7 Figure Showing the Correlation between Different Metrics.....	25
Figure 3-8 Figures showing the average of UES scores in different question sets. The questions set are sorted according to the order in experiment flow.	27
Figure 4-1 Flow of Extracting Facial Features	31
Figure 4-2 Flow of the feature selection process.	33
Figure 4-3 Sleepy subjects in low or medium engagement who were waking up themselves.	36
Figure 4-4 User interface of the video sharing platform with Emotar.....	40
Figure 5-1 Illustration of forward, backward and regressive saccade.	48
Figure 5-2 Results of content extraction. Boxes in red are the lines detected and boxes in orange are the words detected.	48
Figure 5-3 Flow of Extracting Eye Tracker-Captured gaze features.	49
Figure 5-4 Regions of the Language Learning Task. Fixations in 0 refers to the interest on other things; Fixations in area 1 refers to information receiving and in area 2 refers to decision-making.	50
Figure 5-5 Flow of Webcam-based Gaze Features Extraction.	53
Figure 5-6 The eye detected by CLM Face (dots in red color) and CLM Eye (dots in green color) models, and the distance that we could calculated from the landmarks.	53
Figure 5-7 (a-f) Summary figures of comparing differences between pairs of individual criteria.....	58
Figure 5-8 (a-h) Summary figures of comparing differences between pairs of individual criteria.....	61
Figure 6-1 Flow of Mouse Feature Extraction.....	67
Figure 6-2 User Interface of the Crowdsourcing Annotation Experiment.....	72
Figure 6-3 User Interface of Web Searching Tasks.	73
Figure 6-4 Regions of Crowdsourcing Annotation Task.	74

List of Tables

Table 4-1 The 24 facial features extracted from the landmarks.	30
Table 4-2 List of Initial Facial Feature Set.	32
Table 4-3 Intermediate facial feature set selected after step 1 and 2.	34
Table 4-4 Result of doing single factor one-way ANOVA test on the intermediate facial feature set. Features in green color are having statistically significant difference under different level of engagement.	34
Table 4-5 Final facial feature set.	35
Table 4-6 Summary table of the results of the models that are using different of facial feature set and dataset.	37
Table 5-1 List of AOIs Related Eye Tracker-Captured Gaze Features.	51
Table 5-2 List of Pattern Related Eye Tracker-Captured Gaze Features.	52
Table 5-3 List Statistical Eye Tracker-Captured Gaze Features.	52
Table 5-4 Initial Webcam-based Gaze Feature Set.	55
Table 5-5 Intermediate Feature Set of Webcam-based Features.	56
Table 5-6 Intermediate Feature Set of Eye Tracker-Captured Features.	57
Table 5-7 Result of doing single factor one-way ANOVA test on the intermediate Webcam-based feature set. Features in green color are having statistically significant difference under different level of engagement.	57
Table 5-8 Result of doing single factor one-way ANOVA test on the intermediate Eye Tracker-Captured feature set. Features in green color are having statistically significant difference under different level of engagement.	60
Table 5-9 List of Eye Tracker-Captured Final Gaze Feature Set.	61
Table 5-10 List of Webcam-Based Final Gaze Feature Set.	62
Table 5-11 Summary table of the results of the models that are using different of Eye Tracker-Captured feature sets and dataset.	62
Table 5-12 Summary table of the results of the models that are using different of Webcam-based feature sets and dataset.	64
Table 5-13 Summary table of the results of the models that are using different of gaze feature sets and dataset.	65
Table 6-1 Initial Mouse Cursor based Statistical Feature Set.	68
Table 6-2 Part of the Initial Mouse Feature set describing the regularity of the 7 attributes.	69
Table 6-3 Part of the Initial Mouse Feature Set that related to mouse transition pattern in AOIs.	70
Table 6-4 Intermediate Mouse Feature Set.	79
Table 6-5 Significant values of the Levene's Test of Homogeneity of Variance.	80

Table 6-6 Result of doing single factor one-way ANOVA test on the intermediate mouse feature set. Features in green color are having statistically significant difference under different level of engagement.....	81
Table 6-7 Result of doing Kruskal Wallis test on the intermediate mouse feature set. Features in green color are having statistically significant difference under different level of engagement.	81
Table 6-8 Final Mouse Feature Set.	82
Table 6-9 Summary table of the results of the models that are using different of Mouse feature sets and dataset.	82
Table 7-1 List of Final Feature Set Obtained from Webcam Signals.	84
Table 7-2 List of Final Feature Set Obtained from Tobii Signals.	85
Table 7-3 List of Final Feature Sets Obtained from Mouse Cursor Signals.	85
Table 7-4 Summary Table of Using Different Combination of Modalities for Engagement Detection in Language Learning Tasks Dataset.	86
Table 7-5 Summary Table of Using Different Combination of Modalities for Engagement Detection in Web-Searching Tasks Dataset.	87
Table 8-1 Photorefraction. Top photo: vertical orientation of mobile device with flash to the left of the eye; bottom photo: horizontal orientation with flash to the top of the eye. (Ray diagram adapted from Chan, Edwards and Brown [19])	95

Chapter 1 Introduction

Because of their role in human expression, communication, and productivity, human affects have become a major topic in Human-Computer Interaction and Affective Computing.

To understand human mental states, researchers have done much promising work. However, many of the investigated methods are not suitable for daily computer interaction tasks. There are various reasons for this, including the cost of the required sensors or equipment, the difference between environmental conditions in the lab and in real use, and so on. For example, to explore the mental states, some researchers may use intrusive and expensive machines to extract features from electroencephalography (EEG) signals. Even though using EEG can achieve promising results in detecting ones' affects and mental states, it is hard to apply the proposed models in daily use due to the expensive cost of getting an EEG device, and the intrusive nature of its use.

Some researchers propose to embed sensors in daily objects to reduce the intrusiveness of using such devices. For example, there has been some work in using a mouse with embedded physiological sensors that is capable of detecting heartbeats [40] and using a cushion with pressure sensors for recognizing sitting posture [48]. The limitation of these approaches is that special sensing devices are required, which limits their usability.

Given these challenges, the work presented in this thesis focuses on non-intrusive visual cues for understanding human mental states. In particular, we focus on the visual cues from facial expressions, eye gaze, and mouse cursor signals for engagement detection in human-computer interaction tasks.

New technologies, such as computers with high computational power and new kinds of sensors, have made automated human affect detection possible in a variety of situations. Emotion (affect) recognition is essential to Human-Human Interaction. By considering one's facial expression, voice, gesture and speech, humans can recognize someone's affects. Likewise, webcams, which function as the "eye" of the computer, could be used to detect human emotion by "looking" at one's facial and body expression.

One of the major directions of affective computing is to focus on the identification of basic emotions, such as amusement, anger, disgust, fear, surprise and sadness, via a variety of modalities which include but are not limited to facial expressions, motions,

and physiological signals. Techniques in basic emotion recognition have now become mature, some of the detection methods may even reach 99.7% accuracy rate [66]. Instead of classifying emotion into discrete classes, some researchers [44] work on detecting valence and arousal with regression classifiers.

These successes in basic emotion detection give us a general understanding of human behavior. However, more information is needed for deeper understanding of real-life behavior, user experience and user intention. Automatic recognition and detection of conflict, disagreement and human cognitive states such as stress, fatigue and engagement has attracted considerable attention in recent years.

An intriguing idea to enhance usability and the user experience is for computers to be able to make adjustments based on recognized human behaviors. For example, computers can control CCTVs and zoom in on areas where fights or conflicts exist and raise alarms or notifications immediately. Together, these works give us a deeper understanding of how computers can detect human behaviors under different situations.

In the past decade, studies on computer interaction have shown the necessity of broadening the scope of User Experience (UX), instead of just focusing on usability metrics such as the effectiveness of an application. Lalmas [70] even describes User Experience as a part of the “third wave” of Human-Computer Interaction, and that the traditional indicators of usability are not sufficient for capturing UX.

Taking the amount of time spent on a task as an example, usability research may find that a shorter time period indicates that a system is able to more efficiently deliver contents to the user. However, we may find that a short time period is not necessarily desirable when we look at engagement. Is the user spending less time because he is engaged in the task and is thus more productive, or are they discouraged or bored, and thus spend less time on the task? User Engagement is complex as it also includes affective, cognitive and behavioral factors.

We therefore focus on engagement detection during tasks and make use of the visual cues and information hidden in facial expression, eye gaze and mouse movement.

1.1 Background and Motivation

1.1.1 Understanding User Engagement

Human cognitive state recognition and assessment such as stress, fatigue and engagement have attracted considerable attention in recent years. Cognitive states could be measured by different approaches, including but not limited to biological [76,93] and physical measures [3,30,57].

At the same time, user experience (UX) is a relatively new field of research and practice. In general terms, UX deals with the study, design and evaluation of the experience users have with a system [111]. However, UX differs from previous related concepts in that it not only focuses on fulfilling a need, but also considers other factors, e.g. a user's internal state, the system's characteristics and the interaction context [54]. The challenges of developing both the Human-Computer Interaction (HCI) and User Experience (UX) communities have been addressed by Sanchez et al [100].

Recently, there has been interest in detecting user's engagement level, which is a measurable, short-term and affective response [72] and is defined as the quality of user experience which emphasizes the phenomena related to the willingness of using a technological resource for a longer time and more frequently [10]. We define user engagement as the quality of the user experience that emphasizes the positive aspects of interacting with the task and the desire to work on the same task longer and repeatedly.

Precise and successful recognition of the engagement levels of users could be highly valued in many situations. For example, the detected user engagement level during game playing could be used to evaluate the quality of playing experiences and to rate games [1]. Engagement level could also help with evaluating online website designs since it reflects whether the interface and contents successfully attract users' attention [114]. Besides, when users are working on Massive Open Online Courses, tracking their engagement levels could contribute to a clearer understanding of when the student is going to be disengaged and timely interventions could be triggered [15]. One of the major restrictions of current research on engagement level detection is that the experiments and results are derived from specific tasks and there is no general model to detect engagement level.

Due to the difficulty of measuring and quantifying engagement levels, various methods have been adopted to describe engagement. For instance, observational

checklists, rating scales [72,115] and users' self-reports [28] are commonly used to serve as ground truths of the engagement level. Self-reports are undoubtedly useful, but they have their own limitations. This is especially true for self-reporting questionnaires that simply require subjects to answer some straight-forward questions and report a few numeric numbers to serve as their engagement level, as these measures are subjective and may be affected by individual scoring preferences.

In addition, engagement is widely believed to consist of 3 dimensions [39]: behavior, emotional and cognitive engagement. Therefore, an incomplete clarification or definition of engagement could lead to biases in evaluation. *Cognitive Engagement* is related to person's cognitive ability, such as focused attention and memory; *Behavioral Engagement* describes or represents the willingness to participate in the process, like staying on task and finishing the required works; and *Emotional Engagement* represents the emotional attitude towards the task, for example, an employee who is working well on the assigned task, but still dislikes the task. This calls for an objective and comprehensive method of engagement measurement to be adopted.

With an appropriate method of measuring engagement level, extracting useful features for precise detection of engagement levels is also important. Unlike other information sources which require professional and expensive devices, using low-cost and non-invasive devices to recognize human mental states is our main objective. Therefore, we use the mouse cursor and the common embedded webcam. From these, we study and analyze the relationships between these easily-accessed signals and engagement level.

1.1.2 Engagement Detection via Facial Expressions

Facial expressions have also been used to detect users' engagement in recent papers in the context of structured writing activities [85]. Some researchers [43] track facial movements and use the most frequent action units (AUs) to predict engagement levels. Students' emotions and how engaged they are have been studied by Bosch et. al. [14] under uncontrolled group settings where students could move around and talk to each other freely.

There are lots of successful previous works using facial response in affect recognition. However, the existing techniques could be problematic when applying in real-use situations due to the differences between individual users. Much work has focused on training a user-independent model to fit the majority of the users and usually

relies on supervised machine learning [120], which needs sufficient numbers of well-annotated data.

To describe facial expressions, there are two main methods: message judgment and sign judgment [23]. Message judgment focuses on identifying the whole expression and it defines facial expression in terms of inferred emotion. On the other hand, sign judgment tries to code small expressions as they happen on the face and measures affects through the coded behavior. A well-known method using sign judgment is the Facial Action Coding System (FACS) [33] which decomposes facial expressions into action units (AUs). The traditional method is to invite experts to watch the video-recorded facial behavior in slow motion or frame by frame, then coding the AUs manually. This is time-consuming and requires large numbers of well-trained AU coders to code the same video for quality control [61,120].

New techniques in face detection make automatically classifying AUs in every frame easier. In 2015, CMU released OpenFace [5], an open source library which can provide useful information for each frame, such as the location of facial landmarks, head pose information and AUs intensity and existence. Publicly available libraries such as these help to reduce the work load on extracting AUs from a video. However, to detect engagement level from a video, we should consider additional features that could represent the whole video instead of only snapshot frame-based features.

1.1.3 Engagement Detection via Eye Movements

Apart from facial expression, we can also extract information cues from human eyes. The location of visual attention and the interaction-related gaze movement pattern are largely related to human affect in Human-Computer Interaction.

Visual attention, which is the gaze point that users are focusing at on the screen, has been used for basic affect recognition [106] and mental state detection like mindless reading detection [96] and attention level detection [97]. There are many kinds of devices used in state-of-the-art research that analyses eye movements. However, many of them are intrusive and may be uncomfortable to users. For example, the head-mounted cameras place a camera in front of users' eyes for tracking eye gaze location. Though head-mounted cameras could achieve high accuracy in predicting eye gaze location, their expensive cost and the complicated process of installing cameras make it inconvenient for pervasive applications [37].

To estimate the gaze point locations precisely, special infrared equipment that make use of cornea reflection features could be used. There are different commercial eye trackers available in the market; some of these eye trackers can give precise estimation, and some others are intended for user interaction, which are reasonably precise but have a less precise estimation of the gaze point than professional but expensive devices. The advantages of using infrared eye trackers for interaction purpose is that even they reach reasonable precision and are non-intrusive. There are a growing number of researchers using non-intrusive eye trackers to analyze users' mental states. For example, Li et. al. [73] use the Tobii eye tracker to analyze eye gaze behaviors of users and detect how well a user comprehends different reading articles. Granka et. al. [45] use Tobii eye trackers for search engines rank result evaluation.

Meanwhile, commercial eye tracking devices are not the only sensors that can be used for obtaining eye gaze signals. Webcams are another kind of sensor that could help. This is especially worth investigating since webcams have become commonly equipped devices on most computers or laptops. There are lots of researchers who have tried to track users' eye movements with webcams and predict where they are gazing at [113,121]. However, to track ones' gaze location with webcams requires a long period for calibration. Therefore, some work forgoes knowing the precise gaze location of the user, and instead focus on the eye movements tracked by the webcam. In general, eye movements could be categorized into saccades, fixation and smooth pursuits [52]. Eye fixations are defined as the eye gaze staying still on a single location for certain period. Eye saccade is defined as a rapid movement of the eye, while smooth pursuit describes the slow movement of the eye. Extracting the different types of eye movement from webcam signals may also help in detecting users' level of engagement.

1.1.4 Engagement Detection via Mouse Behavior

Computer mouse movements serve as vital and helpful features in various contexts and research especially in human-computer interactions. Proposed by Huang et. al. [57], eye gaze and mouse clicks can be mined to understand users' stress level and Li et. al. [74] also investigated using mouse features to detect comprehension attention.

Besides, mouse movements have been proposed to contribute to verification systems [122] in a transparent and natural way for continuous re-authentication. Mouse cursor information has also been used to detect the quality of workers who are responsible for doing crowdsourcing tasks [84].

Mouse features including cursor information have recently been used in detecting the level of user engagement. One early research work found that the ratio of mouse cursor movement to time spent on a webpage was a good indicator of how interested users were in the webpage content [78]. A mouse-cursor based method has been proposed to study within-content engagement using an unsupervised learning method and reaches a good accuracy [9]. However, the study is based on a specific task – news reading – and the news articles involved in the study have been pre-selected to clearly create 2 levels of engagement, which is somewhat artificial. Even though the same group tried to avoid collecting data under artificial contexts by using a more open-ended task and a Likert-type 5-point scale questionnaire was used to ascertain the users’ engagement level [8], their study of engagement focuses mainly on how much attention has been paid by the users and how interested the users are in those contents. There is no doubt that whether the user understands the content is important, however, engagement is also about the interaction between users and the task, including the extent of involvement, and users’ affects during the task.

1.2 Study Overview

Understanding human behavior and their level of engagement is vital as it is related to the user experience during tasks and could help to evaluate applications in different aspects. As technology becomes more pervasive in our daily lives, researchers are further emphasizing the importance of measuring user experience. Since user engagement contains affective, cognitive and behavioral components, we would like to focus on the information cues hidden in facial expression, eye gaze and mouse movement which may related to these three components. This thesis proposes effective methods of detecting users’ level of engagement in the common daily computer interaction tasks. To have a comprehensive understanding of the relation between engagement level and behaviors, we conduct systematic analysis based on multiple modalities by using non-intrusive devices.

1.2.1 Engagement Dataset Collection

There has been much previous work in detecting user engagement through various means, but the work is hampered by three main challenges: (1) the constraints caused by using intrusive devices, (2) limitations of specific tasks (like gaming) which may produce user behavior different from daily computer usage, (3) and incomprehensive ground truth as collected by straightforward and direct survey questionnaires that

capture users' self-reported numeric level of engagement which may not cover the three dimensions of engagement. We therefore designed our experiments to cover two very different computer tasks for a more generalizable discussion and measure the ground truth with User Engagement Scales (UES) [89], which covers three different dimensions of engagement.

1.2.2 Engagement Detection with Different Modalities

Although there has been a certain degree of success in research on engagement detection, there still exist significant challenges as previously mentioned. This thesis attempts to address these challenges by introducing user-independent models for engagement level detection during daily computer interaction through facial, eye and mouse analysis. We carry out two main experiments: simulating the scenario of doing homework (which is a Language Learning task), and searching for information on the web. During the experiments, subjects' eye movement, facial response and mouse cursor movement are recorded with an off-the-shelf webcam and a non-intrusive remote eye tracker. This approach does not require our subjects to wear any special devices or sensors. We designed different scenarios during the simulation of doing Language Learning task so as to simulate daily activities that covers different levels of engagement.

For each data instance, we have the facial response videos, eye movement records, mouse movement records, recorded screen videos and users' self-report. The Facial Action Units and head movement in each frame of the facial response video is first extracted. The extracted per-frame attributes are then analyzed to construct features that will be used for representing the whole video. The recorded eye gaze data is classified into five types of eye gaze behaviors, including eye fixations, saccades, smooth pursuits, blinks and failures. We then extract features that describe these behaviors. The recorded screen video is used to obtain the Area of Interest (AOI) of the users through their eye gaze positions. Features related to AOI are then extracted. Apart from behavior-based features, we also use gaze-based features to capture the information from unfiltered eye gaze positions, such as the statistical descriptors of gaze movements in a segment. The recorded mouse movement data is denoised for processing and all mouse movements period is found in that instance. Statistical information is then extracted, such as the average travel distance of mouse, and the pattern of mouse transition between different AOI is also considered.

A three-step feature selection is then performed on the initial set of features. In the

first step, we make use of the Pearson's correlation to indicate the relation between features and class label. We further use the selected features from step 1 to build a base model and test on the other features one by one, by doing Paired T-Test on the results of 10 times 10-fold cross validation, to see if each feature could bring significant improvement to the model. After step 1 and 2, single factor one-way ANOVA is adopted to determine whether the interaction features perform differently under different level of engagement. If the feature is statistically significantly different under different level of engagement, it is selected for the final set of features.

The set of features that contribute to the detection of engagement level in Language Learning tasks may change in different contexts. We therefore do feature engineering and feature selection on the Language Learning tasks dataset and further use the same features set to test on the Web Searching Dataset. Our findings show the consistency of the contribution of the final features set used for engagement detection.

1.2.3 Multimodal Engagement Detection

Apart from detecting engagement level with different modalities separately, we would like to know how well the models could perform if we use multiple modalities at the same time. Considering different combinations of the modalities allows us to know how well the model would perform even we if cannot obtain signals from all modalities in real-life usage. For example, suppose we have a user who wants to detect his/her level of engagement but he/she does not have a Tobii Eye Tracker. In this case, without the eye gaze position data, how well can the model perform? Thus, in Chapter 7, we investigate model performance under different situations.

1.3 Thesis Aims and Outline

The aims of this thesis, as outlined in the study overview, are as follows:

- To collect an engagement dataset, which contains two computer interaction tasks, with non-intrusive devices, i.e. webcam, Tobii Eye Tracker and mouse.
- To investigate the detection of engagement level based on two common daily computer interaction tasks, i.e. Language Learning tasks and Web Searching task, by investigating the eye gaze, facial and mouse behavior with non-intrusive devices.

- To identify indicative features that are effective in describing specific eye gaze, facial and mouse behaviors and build user-independent models to detect the engagement level.
- To propose models with multiple modalities to detect engagement level via facial features, eye gaze features and mouse features with off-the-shelf devices.
- To compare the performance of models with different combinations of modalities for engagement level detection for a deeper analysis of the different modalities.

The remaining chapters of this thesis will cover the following:

Chapter 2 provides the literature reviews on the facial affect recognition, the gaze and mouse analysis research work. More specifically, research efforts related to user experience and user engagement detection, eye gaze behavior analysis, mouse behavior analysis and human affects recognition in daily computer interaction tasks.

Chapter 3 describes how the two engagement datasets were collected and mentions about the details of the experiments flow. Meanwhile, the results of self-reports were analyzed and discussed in this chapter.

Chapter 4 explores engagement level detection based on facial expression. We first explore frame-based emotion detection and further analysis user experience on the application that developed with the results of the model. Then, moving from frame-based model to segment-based model, ways of extracting the possible features set and final features set are described. The performance of unimodal classification using facial features is discussed. We further extend the facial affect recognition model into a real-life application and developed a video-sharing platform with Emotars.

Chapter 5 introduces engagement level detection based on eye gaze behaviors. The type and definition of eye movements, extraction of the initial features set and the final features set are discussed. We further did a pilot study on using webcam to identify eye behaviors for detecting engagement. The performance of unimodal classification using eye gaze features is discussed.

Chapter 6 depicts the techniques to extract mouse movement features for detecting

engagement level. We first explore the possibilities of predicting user next interaction event using mouse features. We add a few more mouse features in to the features set used in next interaction event prediction and form the initial features set for engagement level detection. The final feature set was selected after feature selection and used for detecting engagement. Finally, the performance of the model is discussed.

Chapter 7 explores the performance of the models with different combination of modalities. The modalities are grouped by the way of features collection and three different modalities, which are webcam, Tobii Eye Tracker and mouse.

Chapter 8 summarizes the contributions and limitations of this thesis and the potential future work. This chapter also introduces other contributions we have made that are related to or beyond the scope of this thesis.

Chapter 2 Literature Review

2.1 User Engagement

Engagement level detection is critical for many different tasks. For example, whether workers are focusing on their tasks, or whether students are doing an e-learning task seriously, or even whether workers are working on a crowdsourcing task properly. Grinberg et al. [46] are interested in detecting users' short-term engagement levels since it helps with understanding of online contributors and improvement of designs of social network sites. Whitehill et al. [115] investigated in detecting learning engagement level automatically using features extracted from users' facial expressions.

More systematically, user engagement is composed of three dimensions [39]: (1) emotional, (2) cognitive and (3) behavioral. There are some popular characteristics associated with engagement where focused attention is one of them. Focused attention evaluates the level of excluding other people or other things while users are involved in specific tasks [87]. It relates to the difference between users' perceived time and real time spent on tasks [88] and it is thought to be indicative of cognitive involvement [11]. Positive affect [60] and aesthetics [87] have also been proposed as vital characteristics related to users' engagement level. Endurability, which indicates the likelihood of repeating the same tasks, and novelty, that represents surprising and unexpected things, are also characteristics of user engagement [95,107].

Considering the comprehensiveness of user engagement, measurement and evaluation are challenging. Obtaining the engagement level is not easy and is usually done via self-reporting [12] - through the use of questionnaires or surveys. It is known that self-reporting provides useful feedback, but it relies on reasoned self-reflection and does not provide information about spontaneous, instantaneous or even unconscious reactions from the users. Self-reporting may also distract users' attention from the task that they are focusing on. One frequently used scale based on self-report is UES [89], which contains a systematical and standardized structure for eliciting user engagement assessment. They use questionnaires to collect online shoppers' engagement and categorize information into key characteristics of UES by using factor analyses. Some researchers have also validated UES for other fields. For example, Wiebe et al. [116] investigate whether UES could be used to measure engagement during video game-play and compare it with other scales. Based on the original UES, they proposed a revised version which is more predictive in video game play context.

Besides subjective methods for user engagement, there are some other methods which make use of devices for an objective measurement of engagement. Andujar and Gilbert [6] proposed to measure one's engagement level via a non-invasive EEG emotive EPOC device and they implemented a prototype that uses a “physiological reading method” approach for evaluating engagement in current and future educational tools. Mathur et al. [76] conducted studies to compare engagement scores generated from EEGs and found a high correlation with UES.

In order to understand more about human behavior, some recent works tried to recognize human actions and predict user intention. For example, Kato et. al. [62] predict the next step of human behaviors by modeling human body movement and gestures for human robot communication. Wacharamanatham [112] try to detect whether a gesture or a movement of user is intended to control or not by modeling and classifying user's gesture type. Similarly, Frank et al. [38] make use of multimodal information from 2D and 3D imaging and sound to evaluate and detect the engagement level at group level, in a sensor-based environment where personal body motion, gesture, facial expressions, voice and other biometric signals could be processed.

Various types of features have been used in engagement detection, and facial expression is one of the most frequently used features. Whitehill et al. [115] make use of facial features which are proposed by human observers in a pilot experiment to automatically detect students' engagement using machine learning. They confirm the reliability of these facial features in discriminating low and high levels of engagement and indicate the relationship between student's performance and their automatically detected engagement level. Alyuz et al. [4] also use appearance information together with context-related features to build a semi-supervised model adaption method to achieve accurate emotional engagement detectors.

Besides facial expressions, eye gaze features are also helpful to engagement detection. Nakano and Ishii [86] analyzed speakers' eye-gaze patterns in conversational contexts and extract disengaged patterns based on a proposed engagement estimation method. Then they build a conversational agent which is able to detect users' engagement level and probe out questions for attraction.

To have a more comprehensive understanding of these useful features, it is necessary to do more literature review concerning the extracting and implementation of features including facial expressions, eye gaze and mouse movement in a broader discipline.

2.2 Facial Expression Analysis

In order to recognize someone's affects or stands, humans usually focus on facial expression, gesture, speech and voice. Traditional Human-Computing Interaction (HCI) designs rarely pay attention to the implicit information of user [120], such as the affective state of the user. If we make use of user affective states, we could enhance users' experience by giving suitable reactions or responses. New technologies have improved computational power and provided lots of sensors and equipment that could be used to produce inputs for HCI. Detecting human affects can become possible with these inputs.

Therefore, many researchers focus on recognizing affects using facial expression detection [23], speech detection [58], motion detection [92] and physiological signals detection [18]. Basic affects like happy, sad, anger and surprise etc. have been successfully detected with these methods.

Facial expression recognition has been investigated for many years, Ekman and Friesen [108] proposed a Facial Action Coding System (FACS) that systematically divided facial motion into action units. Facial expressions are then defined by different combination of facial action units (AUs) and used for affects recognition [13]. Essa and Petland [34] also proposed the FACS+ model which extend FACS model to combine temporal and spatial modelling of facial expressions.

There are two main methods used for detecting AUs [102]: Geometric-based methods and Appearance-based methods. Geometric-based methods obtain facial features points and the shape of each component from the face. AUs could be detected through the "motion unit" from the face [24,101]. In contrast, Appearance-based methods extract features from facial textures such as wrinkles for analysis. With the extracted facial features, different kinds of models could be used for predicting affects such as Hidden Markov Model (HMM) [119], Naive Bayesian fusion (NBF) [2] and Support Vector Machines (SVMs) [66].

Besides human affects, facial responses could also be used for detecting human interaction relationships in activities. Kim and Vinciarelli [64] proposed an approach for conflict detection and defined conflict in dimensional instead of categorical terms. They use a Bayesian approach for Automatic Relevance Determination (ARD), which weighs the features according to their influence on the regression output. They used Gaussian Process Regression (GPR) to build the model and got results that show the

correlation between actual and predicted conflict level is between 0.7 and 0.8.

McDuff et. al. [80] use facial responses for detecting viewer agreement and preference using a dimensional approach. During the 3rd Obama-Romney presidential debate, viewer facial responses were collected to detect voters' preference by recognizing smiles and smirks from facial expressions. Meanwhile, they [79] proposed to use crowdsourcing for collecting responses from different people who are watching commercials and advertisements, using the same technique to determine if the advertisements are effective or not.

2.3 Eye Movement Analysis

Besides facial expressions, users' eye behavior is another fundamental and significant information source since it reflects the progress of acquiring various knowledge. In recent years, there has been much research making use of eye behaviors to investigate users' interaction with computers. For example, Li [73,74] studied the relationship between user's comprehension level as well as attention level with eye gaze patterns in reading tasks. They used eye gaze features extracted by commercial eye trackers to detect the level of users' comprehension and achieved a performance improvement of over 30% above baseline. Slanzi et. al. [105] studied human eye gaze behavior, pupil dilation and EEG signal for predicting web users click intention. Another research [123] investigated the use of user eye gaze movement for input and applied it to build a gaze-controlled game for user authentication. There is another work [90] focusing on detecting drivers' focus attention by tracking their eye gaze behaviors and authors proposed an eye-gaze-based model for focusing attention detection.

In these research efforts, eye movement patterns are frequently used as representative eye behavior features. There are four types of eye movement defined by Hansen and Ji [52]: smooth pursuit, saccades, eye blinks and fixation. Smooth pursuit defines relatively slow motions of eyes and saccades represent relatively rapid motion toward a stationary object. Fixation implies a stationary gaze: Crouzet [26] proposed that staring at a point for 180ms could be considered as a fixation of eye movement. Besides eye movement, eye gaze locations and patterns are also frequently used.

To extract eye gaze features, some commercial devices are used frequently, e.g. Tobii, SMI, EyeLink, and Smart Eye etc. However, commercial eye tracking devices are expensive and not (yet) home equipment. They are not the only modules for getting eye gaze signals and webcam is an alternative module that could help. Considering that

webcams have become a standard device on most computers or laptops, it could be easily used to capture users' head including their movements. Gaze point could be estimated by features extracted from eye regions [121] and also by images of eye regions without extracting specific features [35]. However, many models require a complex procedure of calibration. Williams et al. [117] require users to follow specific spots on the screen and Lu et al. [36] require users to adjust head position while fixating on one calibration point. However, unlike eye gaze point estimation which requires a complex calibration process, estimating eye behaviors is more feasible without calibration since images of eyes could be processed clearly and directly for their movement pattern. In this case, using webcam to detect users' eye movement behaviors is promising and useful.

2.4 Mouse Movement Analysis

Traditionally, together with the keyboard, the mouse serves as a major and vital role in the interaction between human and computer in various occasions, like making decisions, assisting reading, selecting texts, etc. Thus, a large volume of interactions gets generated while using mouse in tasks and some research suggests the dominant role of mouse movement in daily computer interactions [20,83].

Considering the significance of mouse movement, there are many researchers interested in extracting and investigating mouse interaction data in various situations and tasks. Proposed by Lalle et.al [69], mouse interactions have been used to predict user's intentions in searching tasks and cloud-sourcing tasks. Li et. al. [57] investigated features extracted from mouse movement for detecting users' level of stress in mathematics tasks. Mouse movements have also been proposed for verification systems [122] in a transparent and natural way for continuous re-authentication. Mouse cursor information has also been used to detect the quality of workers who are responsible for doing crowdsourcing tasks [84]. A linear regression model is proposed by Asano et al. [3] to apply mouse endpoint prediction and a probabilistic model is proposed by Ziebart et al. [116] to predict points' movement using inverse optimal control techniques.

Mouse features including cursor information has recently been used to detect the level of user engagement and one early research found that the ratio of mouse cursor movement to time spent on a webpage was a good indicator of how interested users were in the webpage content [78]. A mouse-cursor based method has been proposed to study within-content engagement using an unsupervised learning method and achieves a good accuracy [9]. However, the study is based on a specific task – news reading –

and the news articles involved in the study have been pre-selected to clearly create 2 levels of engagement, which is somewhat artificial.

Chapter 3 Engagement Datasets

In order to investigate how users' behaviors correlate to engagement, it is necessary to build a dataset that contains user behaviors under different levels of engagement. Our work focuses on understanding users' behaviors when they are engaged in daily computer tasks involving mouse interactions. We therefore design our dataset to contain user behavior in two widely different computer tasks: a multiple-choice practice system referred as Language Learning task and Web Searching task. As these two tasks are quite different from each other, we believe that this will lead to a more generalizable result and analysis.

It is imperative that our dataset contains adequate samples of users working under different levels of engagement. In order to ensure this, each task was segmented into three phases (Figure 3-1). The first phase is simply a warm-up phase to allow the experiment subjects to get familiar with the task and the interface. The second phase, or Scenario 1, asks the users to work on the same task. Presumably, since the users are still fresh and alert, this phase has a higher likelihood of generating user behavior that exhibits high levels of engagement. The third phase, or Scenario 2, asks the users to work on the same task but with a distractor (fatigue, or fatigue plus noise). The rationale is that during the third phase, the users are less fresh and therefore likely to be more easily distracted (especially with the distractor) and hence not very highly engaged. After each phase, the user self-reports his/her level of engagement via three commonly-used metrics (Section 3.3). This self-report is used as the *gold standard*, or the "correct" engagement level.

Our experiment setting logs all user behaviors that might be obtainable unintrusively using standard computer equipment. This includes the facial expression, as captured by a standard webcam, the mouse movement, as logged by the system, and the eye gaze movement, both captured by the webcam as well as detected by the commercial Tobii EyeX Eyetracker. The engagement datasets described in this chapter will be used and mentioned in Chapter 4 to 7.

The experiment participants were introduced to the purpose of our experiment and their permission obtained to collect data. Since our goal was to collect user behavior under differing levels of engagement, we explicitly told the subjects to behave naturally and that the quality of their work or their self-report "answer" would not affect their compensation.

3.1 Language Learning Tasks

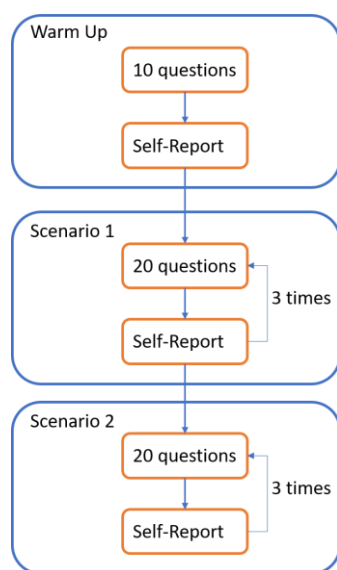


Figure 3-1 Flow of Language Learning Language Learning Task Experiment.

This task is representative of common tasks in educational contexts, such as online learning systems or MOOCs (Massive Open Online Courses). To simulate real daily experiences of doing tests, we designed 2 common scenarios drawn from daily life.

For each scenario, as shown in Figure 3-1, participants are required to finish 3 sets, 20 questions each, of basic English questions that are selected from an ESL (English as a Second Language) tests. The difficulty of the questions is at primary to junior secondary school level in Hong Kong. Before starting, a warm up section containing 10 questions followed by a set of self-report is conducted to ensure that participants are familiar with procedures and operations in later scenarios. To avoid biases brought by the order of English questions, the questions are randomly selected from the database and get shuffled in each experiment. The user interface of this task was shown in Figure 3-2.

In pilot experiments, it was noted that since the experiment was conducted in the laboratory, it naturally made our subjects more tense and focused, and it is harder to obtain low / medium engagement samples. Therefore, our distractor in Scenario 2 included noise on top of the fatigue factor.

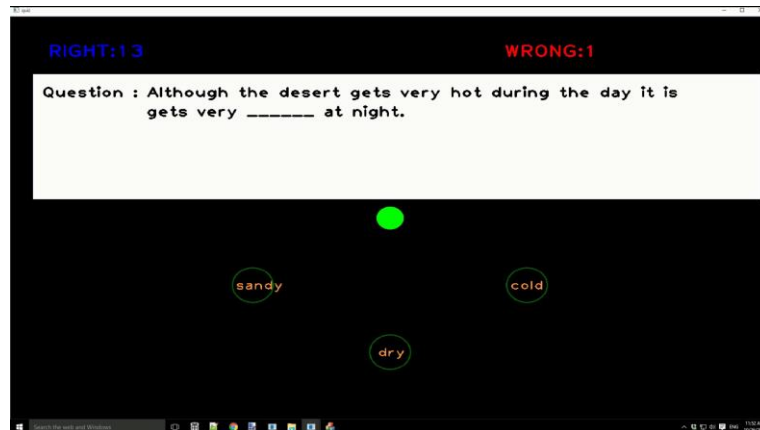


Figure 3-2 User Interface for the Language Learning Task

A total of 20 subjects (14 F) aged 18 – 29, participated. All subjects are familiar with computer usage. In total, 119 instances are collected, where each instance corresponds to one participant answering one set of 20 questions with a successful self-report annotation. Each instance lasts for four to six minutes. There are 7 instances for low engagement, 70 instances for medium engagement and 42 for high engagement.

3.1.1 Scenario 1 – Doing Homework

Scenario 1

Scenario description

- You are doing a **home work** which is an English test;
- You **do not have time limitation**;

Details

- In this scenario, there are **3 rounds** of tests.




Figure 3-3 Instruction slide for Language Learning Task - Scenario 1

In this scenario, participants were told that they are doing their homework without any time limitation. The instructions are shown in Figure 3-3. For each question, they needed to choose 1 answer out of 3 potential answers. Once they clicked and chose the answer, the next question was then shown on the screen. In case he/she clicked on the wrong answer, the background of the interface will flash red. For each set of questions, his/her marks were counted and shown at the top of the interface. After the subject finished 1 set of questions, i.e. 20 questions, they were asked to self-report their

engagement level status using a questionnaire embedded into the same user interface. To ensure the consistency of reported scores, the interface also showed the reported scores from the previous round as a reminder to the participants.

3.1.2 Scenario 2 – Working under Distracting Situations

Scenario 2

Scenario description

- You are doing a home work which is an English test, while **your family members are watching TV outside** your room with large volume;
- You **do not have time limitation**;

Details

- In this scenario, there are **2 rounds** of tests.




Figure 3-4 Instruction slide for Language Learning Task - Scenario 2

Similar to scenario 1, participants were told to do their homework and follow the same working process. The only difference is that Scenario 2, shown in Figure 3-4, simulates a situation whereby the participant's family is "sitting" outside the room watching TV, which is set at a loud volume (not uncommon in Hong Kong homes with limited space). To simulate this scenario, the participant wears headphones, pre-set at a fixed volume, over which a talk show is playing. Since our participants have different native languages, we pre-recorded 3 talk shows in different languages, and chose the one in the participant's native language for this scenario. All talk shows are long enough to make sure that it will keep playing for the length of time that the subjects are working on Scenario 2.

In order to further reduce the engagement level, we intentionally mislead the subjects and tell them that there are only 2 sets of questions in Scenario 2, when in fact the scenario consists of 3 sets of questions.

3.2 Web Searching Task

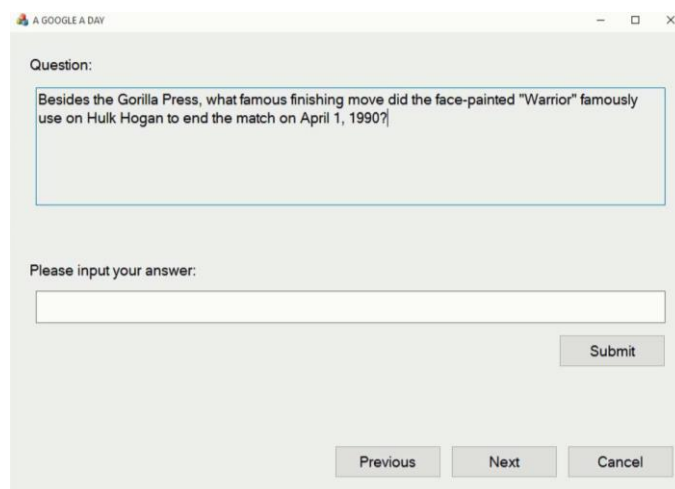


Figure 3-5 Interface of Web Searching Task for displaying and answer questions.

Web Search is a very common task nowadays. Unlike the Language Learning task, the Web Searching task is more open-ended. There are no pre-determined display pages and interactions between users and computers are more complicated and unpredictable.

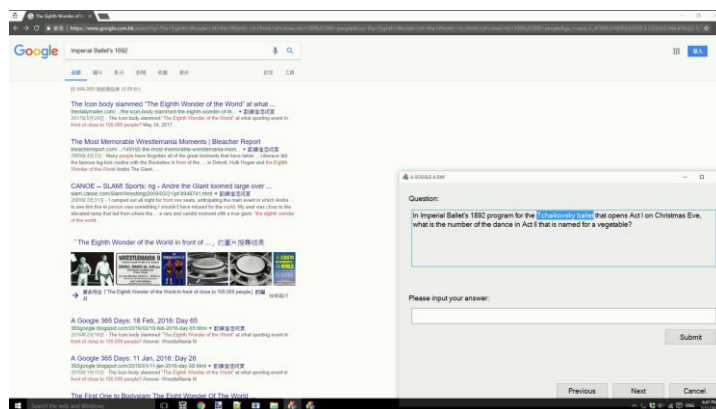


Figure 3-6 Captured screen of doing Web Searching task by one of the subject.

The interface for this task is shown in Figure 3-5. It is based on the Web Searching game developed by Google, called “A Google a Day” [42]. The computer presents a question in natural language, and participants are expected to find the correct answer based on information that they can find online, as shown in Figure 3-6. The questions are written to avoid cases where answers could be found by simply using basic web search, therefore users are required to rephrase their search queries, often over multiple iterations. Every 5-10 mins, participants were prompted to self-report their engagement

level during the last question. There is no limit on the number of queries to be answered at each sitting, and the number of times that a participant can do the experiment is also unlimited.

The experiment lasted a total of two weeks, during which a total of 12 subjects (8 F) aged 18 – 27 participated. All subjects are familiar with computer. 77 instances were collected. Each instance lasts for five to ten minutes. Among all instances, there are 7 instances for low engagement, 43 instances for medium engagement and 27 for high engagement.

3.3 Annotating the Gold Standard

3.3.1 Metrics and Measurements

Self-Assessment Manikin (SAM)

There are several measuring tools used for self-reporting. First of all, we would like to record the affective status of our user, therefore, Self-Assessment Manikin (SAM) [17] was used for self-reporting the arousal and valence. Psychologists define arousal as an intensity that ranges between quiet to active and valence is defined as the direction of the affect, which ranges from feeling pleasant to unpleasant [67]. SAM is an effective non-verbal method for quickly accessing person's emotional reaction. This could help our subjects, where none of them are native English speaker, to report their affective states in an effective way.

User Engagement Scale (UES)

Meanwhile, the user engagement level was measured with the User Engagement Scale (UES)[89]. UES is a 7-point Likert scale designed for measuring user engagement level with 31 questions that cover 6 different areas. After discussing with the author, the following 5 dimensions are used in our study: Endurability (EN), which measures the perception of the activity as worthwhile; Focused Attention (FA) for measuring the perception of time passing; Felt Involvement (FI), which measures the perception of involvement with the session; Perceived Usability (PU), which measures users' affective, (e.g. frustration), and cognitive, (e.g. effort), responses to the task; and Novelty (NO), which measures users' level of interest in the task and curiosity evoked by the system and its contents.

As the original UES is designed for a web browsing task, we made some minimal changes on wording to fit our scenarios. The averaged UES score will be used as the ground truth of the user engagement level.

NASA Task Load Index (NASA-TLX)

The NASA Task Load Index (NASA-TLX) [53] is commonly used to measure the subjective perception of the workload of a task. It measures 7 different areas, such as mental demand, physical demand, temporal demand, overall performance, frustration level and effort.

3.3.2 Objectivity of Gold Standard Labels of the Dataset

Obtaining the gold standard labels is not an easy task for any dataset, and is especially difficult for affective computing. In our study, we use the subjective measurement, i.e. average UES scores, as the annotated user engagement level. The reason is because objective measurements may be intrusive, and we tried to avoid having to invite expert observers to sit in the experiments. Even though objective measurement was not used in our data collection, investigating how well humans can recognize subjects' engagement level from the collected facial response videos could help us to understand the advantages or limitations of using objective measurement.

We randomly selected 50 instances from the Language Learning Task dataset, and had two observers annotate the videos into three levels of engagement (low, medium and high). Cohen's κ was run to determine if there was agreement between two observers' judgement on whether the 50 instances were low, medium or high engagement. There was poor agreement between the two observers' judgements, $\kappa = .173$, $p < .005$. The result of poor agreement shows that it is difficult for human observers to judge subjects' engagement level. Meanwhile, there was a moderate positive correlation between the two observers' judgements with $r = .411$, $p < .001$. The moderate positive correlation indicated that even though observers may have some ideas of what is engaged and what is disengaged, there are still some inconsistency between the two observers' judgements.

3.3.3 Between-Metrics Correlations

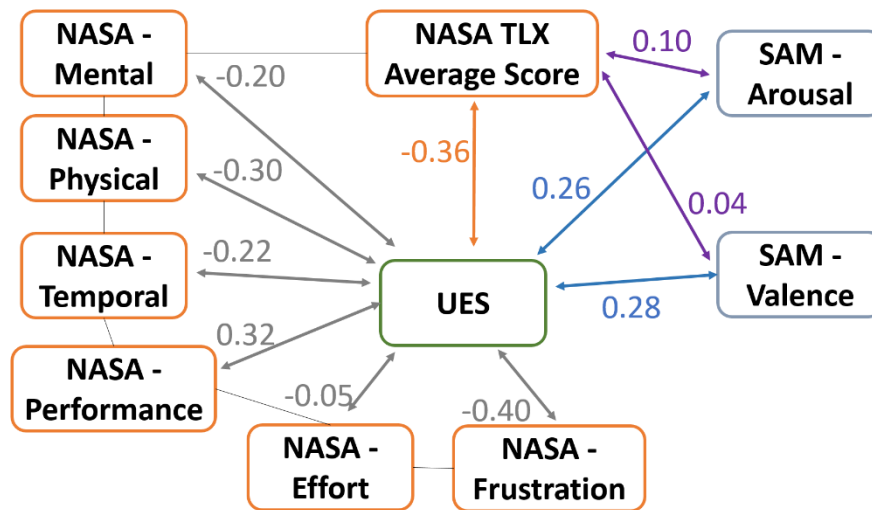


Figure 3-7 Figure Showing the Correlation between Different Metrics.

Since the subjects are required to self-report with different metrics, it would be interesting to know the correlation between the metrics. The results are shown in Figure 3-7. First of all, we can see that the highest correlation is between UES and the frustration subscale of NASA-TLX. It is a negative correlation because the higher the UES score, the higher the engagement level and at the same time, the lower the NASA-TLX score, the lower the task load in the task. Such high negative correlation between “NASA-Frustration” and UES may help to explain that when our subjects feel less frustrated by the task, their engagement level tends to be higher. However, this is just a speculation and further study is needed to confirm if this is the case.

Meanwhile, the correlation between UES and other NASA-TLX subscales exhibit low to moderate correlation. In particular, the correlation between UES and “NASA-Effort” is only 0.05. Such low correlation may imply that the UES score is not affected by the effort that the user is required to spend. This appears to be further corroborated by a Kruskal Wallis Test using NASA-Effort as the variable and the UES as the criteria. The result ($\chi^2(2) = 1.000, p = 0.606$) shows that there is no significant difference in NASA-Effort between high, medium, or low engagement levels.

At the same time, we observe that UES has a moderate correlation with SAM-Valence and SAM-Arousal, which measure user affect. This shows that UES does not only measure cognitive engagement, but also affective engagement. It is not surprising that the NASA-TLX has low correlation with SAM-Valence and SAM-Arousal as NASA-TLX only contains one question related to affective feeling.

From these results, we conclude that UES is a comprehensive measurement which covers both the cognitive and emotional engagement.

3.3.4 Implications from Analyzing Results of Two Scenarios

The Task Load in Scenario 2 is Significantly Higher

Furthermore, for each subject and scenario in the Language Learning task, we averaged the responses to the 6 items from the NASA-TLX questionnaire and normalized the scales between 0 to 1. We use one way ANOVA to analyze the data, Welch' ANOVA will be used if the attribute violates the assumption of homogeneity of variances. One of the items, "Performance" is in reverse order and therefore we reverse the scores of this item. The average scores are 0.27 (SD: 0.148) for scenario 1 and 0.37 (SD: 0.176) for scenario 2. The results from Welch' $F(1,113.098) = 11.499$, $p = .001 < .01$) showed a significant difference between the task load of the 2 scenarios of Language Learning task.

We further run single-factor one-way ANOVA on each item of the questionnaire so as to understand how task load is induced. The main differences between scenario 1 and 2 are that the latter will (1) requires the subjects to listen to a noise distractor in the form of a talk show broadcast, and (2) contains an expectation discrepancy in that we intentionally misinform the user that there will be only 2 sets of questions, when there are actually 3 sets. We want to find out if these experiment designs will significantly increase subjects' task load.

The results show that there is no significant difference in the scores for "Physical Demand", "Temporal Demand" and "Performance" between scenario 1 and scenario 2. However, there are statistically significant differences for "Mental Demand" ($F(1,117) = 6.378$, $p = .013 < .01$), "Effort" (Welch' $F(1,103.769) = 11.834$, $p = .001 < .01$) and "Frustration" ($F(1,117) = 8.552$, $p = .004 < .01$) between scenario 1 and 2. These results suggest that forcing the subjects to listen to the talk show and researchers intentionally giving wrong information may help to increase the subjects' task load. The significant differences in frustration index between different scenarios, which measures how frustrated, discouraged, insecure or annoyed the subjects are, also indicated that our experiment design does indeed succeed in increasing the level of discouragement.

“Intentionally providing wrong information” will not Affect the Perceived Task Load

To further understand the effect of “intentionally providing wrong information to participants”, we run the test to investigate if there the user reports any differences in the overall task load index between the first 2 sets of questions (which were expected) and the last set of questions (which was unexpected) in scenario 2. The average NASA-TLX scores are 0.371 (SD: 0.174) for the first 2 sets of questions and 0.379 (SD: 0.186) for the last set of questions. The results from one-way ANOVA ($F(1,57) = 0.025, p = .875 > .01$) showed that there is no significant difference between these two criteria. This implies that our misinformation design, telling the subjects that they only need to do 2 sets of questions, when in fact they need to complete 3 sets, is not useful in reducing the engagement level. This may be due to the fact that participants forgot what they were told, or do not care about how many sets of questions they have to work on.

Experiment Design helps to Lower Subjects’ Engagement Level

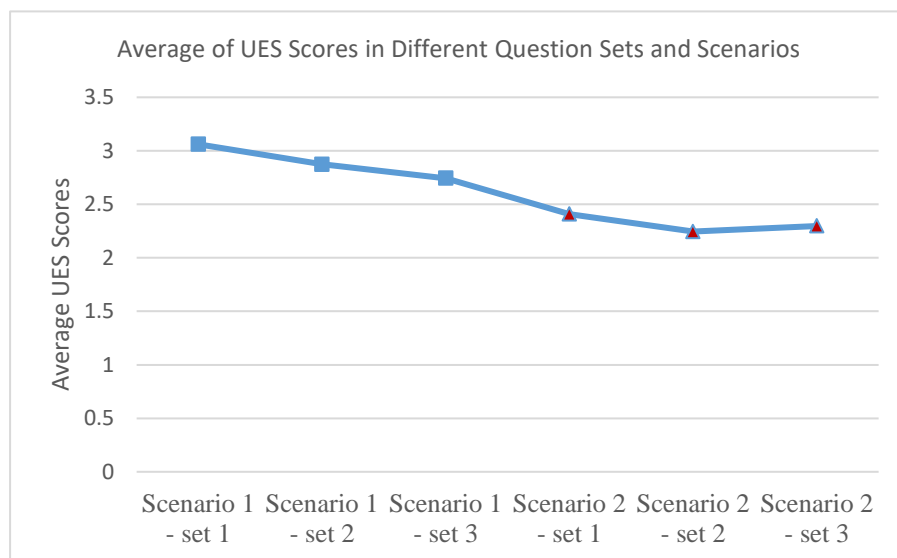


Figure 3-8 Figures showing the average of UES scores in different question sets. The questions set are sorted according to the order in experiment flow.

We also want to confirm if the experiment design successfully lowers the level of engagement. we calculate the average UES scores for each set of questions in the 2 scenarios. Figure 3-8 shows the UES scores in different question sets sorted in order of experiment flow. The results show that the level of engagement drops in the later part of the experiment. We further run a single-factor one-way ANOVA, with the null hypothesis being that there will be no difference in the average UES scores under different scenarios. The result ($F(1,117) = 16.120, p = .000 < .01$) shows that the null

hypothesis can be rejected and implies that the average UES scores are significantly different between the two scenarios. This drop may be caused by (1) sound distraction and/or (2) fatigue. Since there are two possible factors causing the low engagement level, at this point it is not possible to separate their individual contribution, and further studies will need to be done if we want to identify which one is the major cause.

Chapter 4 Understanding Users from Facial Expression

Facial expressions, gesture, speech and voice are frequently used to recognize someone's affects or stands, by understanding which users' experience on using computer like online shopping could be enhanced by giving suitable react or response. New technologies improved computational power of devices and provided lots of sensors and equipment that could be used to process the above sources including facial expression for HCI.

In this chapter, we investigate automatically detecting the level of engagement on different tasks. We specifically focus on facial expressions in this chapter. Facial expressions have also been used to detect users' engagement in recent papers in the context of structured writing activities [85]. Some researchers [43] track facial movements and use the most frequent action units (AUs) to predict engagement levels. Students' emotions and how engaged they are have been studied by Bosch et. al. [14] under uncontrolled group settings where students could move around and talk to each other freely.

We invited subjects to carry out experiments in Language Learning tasks and Web Searching tasks. Feature selection was applied to extract useful features. For cross-task generalization, we only use Language Learning task data during the feature selection process.

We construct our models for user-independence. Our models are tested in 2 different ways. The first is 10 times 10-fold cross validation on Language Learning task data. The second one performs the same validation but on Web Searching task data. We find that our selected facial features achieve a performance improvement of 9.3 % above baseline for testing and training on Language Learning tasks, and 7.8% above baseline for testing and training on Web Searching tasks.

The rest of this chapter is organized as follows. In Section 4.1, methods of extracting facial action units are discussed. Then we investigate how to extract features over a global time segment instead of local (frame-based) snapshot in Section 4.2. Results of the user-independent models are disused in Section 4.3. In Section 4.4, we describe the development of Emotar, an extension of facial affect recognition to real life applications. Section 4.5 concludes this chapter.

4.1 Facial Action Units as Features

Human facial expressions can be represented by different combinations of Facial Action Units (AUs) [33], which are movements of different facial features. For example, AU12 is “Lip Corner Puller” which is usually activated when we smile. To identify the existence of an AU, face recognition and tracking is needed. There has been much work in face recognition and tracking algorithms.

4.1.1 Extract Action Units from Landmarks

Our system uses the Constrained Local Model (CLM) [22], which was trained with CMU Multi-PIE Face database [47], to obtain the locations of 66 facial landmarks and the head pose information. CLM has been proven to achieve satisfactory performance in tracking user-independent facial landmarks with robustness against occlusion and generalizability across different individuals [22].

However, because of the nature of the training image data and the influence of local maxima in optimization, CLM sometimes fails to robustly track Asian subjects under poor lighting conditions and when exhibiting certain kinds of expressions, such as the mouth corner depression. Therefore, Supervised Descent Method (SDM) [118] was used to optimize the 48 landmarks in the central face area. Because SDM tracks landmarks based on both the textures of the landmarks’ local patches as well as their geometric inter-dependency, it is less susceptible to influences from illumination and can provide reliable landmarks that describe the center face region. For each frame, the SDM fitting results was used as the initial state for CLM fitting, which improves the localization accuracy and reduces the convergence time.

With the 66 landmarks located, we further calculate the distance between different landmarks so that 24 different facial features and head pose information could be extracted, as shown in Table 4-1.

(1 – 4) Inner and outer brow movement (Left & Right Eye)	(14) Lip pucker
(5-6) Eye brow movement (Left & Right Eye)	(15) Lip stretcher
(7-8) Eye lid movement (Left & Right Eye)	(16) Lip tightener
(9) Upper lip movement	(17) Lip depressor
(10-11) Lip corner puller	(18) Cheek raiser
(12) Eye brow gatherer	(19-21) Head orientation
(13) Lower lip depressor	(22-24) Head position

Table 4-1 The 24 facial features extracted from the landmarks.

4.1.2 Extract Action Units from OpenFace

In 2015, CMU released OpenFace [5] as an open source library which could provide useful information such as the facial landmarks location, head pose information and AU intensity and existence. As OpenFace achieves competitive accuracy and performance results and it is able to provide frame-based AUs for a given video, we choose to extract our features from OpenFace results.

4.2 From a Local Snapshot to a Global Time Segment

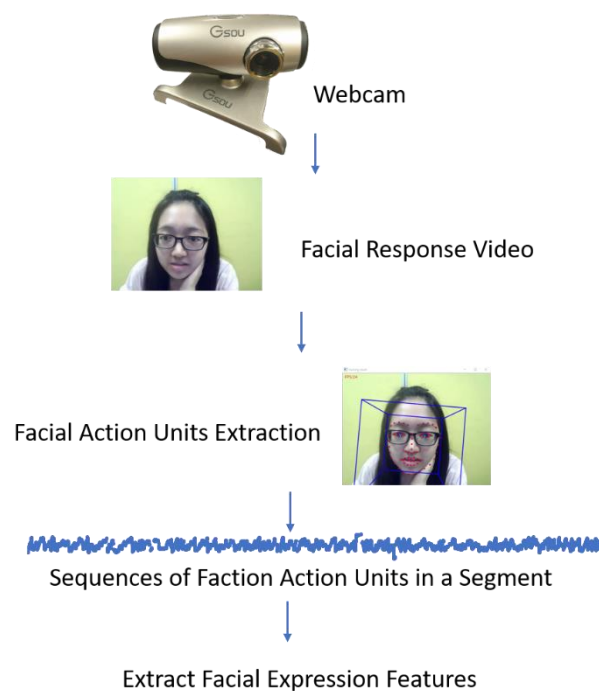


Figure 4-1 Flow of Extracting Facial Features

To represent the whole segment of the video, the sequences of each frame-based feature will be considered and used to extract the segment-based feature. In the following, we will describe the features used for representing an entire video segment.

We extracted and selected the features with the Engagement Dataset - Language Learning Tasks which was described in Section 3.1.

4.2.1 Action Units as Features

We first perform face detection and attribute extraction using the publicly available OpenFace [5] which extracts different features, including but not limited to the head positions in x, y, z dimensions and 18 Action Units (AUs) classification results of each frame in the facial response videos. Therefore, we further process these results to abstract some potential segment-based features from these sequences, as shown in Table 4-2.

Feature	Meaning	Formulation
f 1-6	Location of Head in x, y and z	(1) Standard Deviation and (2) Cumulated Delta of the location of the head with respect to camera in millimeter
f 7-12	Rotation of head in radians around X,Y,Z axes	(1) Standard Deviation and (2) Cumulated Delta of the rotation of the head
f 13-48	18 Facial Action Units	(1) Count of the number of appearance (2) Presence of that AU
f 49	Confidence on Face Detection	Percentage of frames with low confidence (≤ 0.8) in face detection

Table 4-2 List of Initial Facial Feature Set.

The design rationale behind these features is as follows: The standard deviation of head position signals represent and quantify the amount of variation of head movements, such as up and down movements, left and right movements and forward and backward movements. The cumulated delta of head position signals imply the frequency of the subjects' head movement. In cases where subjects move frequently with small variation in position, this will lead to a higher value for cumulated delta and lower value for standard deviation. The presence of AUs is a Boolean value that indicates if an AU exists in the video or not. The count of the number of appearance of AUs implies how often an AU exists in the segment. In total, there are 49 potential features extracted from facial response videos.

4.2.2 Features Selection

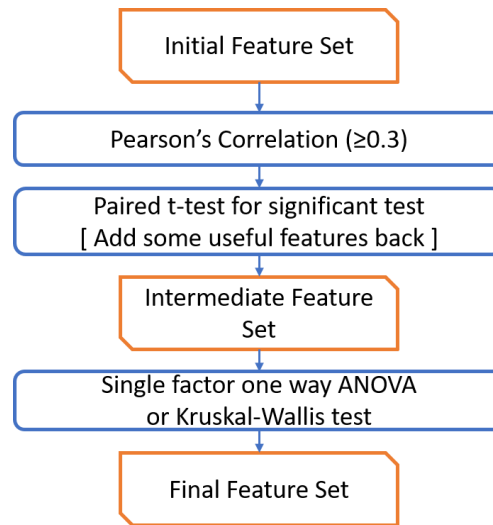


Figure 4-2 Flow of the feature selection process.

Step 1 : Select by Pearson's Correlation

Feature selection was then performed with the 49 potential features extracted from facial responses. Our feature selection is a multi-step process. We first filter the features by calculating the Pearson's correlation coefficient between each potential feature and the label, we use absolute correlation larger than or equal to 0.3, which implies they are having high correlation, as the threshold.

Step 2 : Select by Significant Improvement Brought to the Base Model

The second step uses the features with correlation larger than or equal to 0.3 to build a base model and adds in features one by one, selecting those that bring significant improvement to the base model.

In order to choose between two learning algorithms, Bouckaert [16] suggested using 10 times repeated 10-fold cross validation, where all 100 individual accuracies are used, with 10 degrees of freedom Paired T-Test. Therefore, we run 10 times repeated 10-fold cross validation, with different random seed each time, on the base model and the model with added feature. We then run a Paired T-Test and check if each feature could bring statistically significant improvement to the model. Table 4-3 shows the 6 selected facial features after this step.

Feature	Descriptions
f 2	Standard Deviation of the location of the head - y
f 3	Standard Deviation of the location of the head - z
f 13	Count of the Existence of AU 1 (Inner Brow Raiser)
f 14	Count of the Existence of AU 5 (Upper Lid Raiser)
f 30	Count of the Existence of AU 45 (Blink)
f 49	Percentage of Frame with Low Confidence in Face Detection

Table 4-3 Intermediate facial feature set selected after step 1 and 2.

Step 3 : Select by Performing Single Factor One-Way ANOVA

We finally performed a single factor one-way ANOVA to determine whether the interaction features behave differently under different level of engagement. Before we run the single factor one-way ANOVA test, we also checked the required assumptions, including but not limited to homogeneity of variances, no significant outliers and consisting of two or more independent groups, etc., and confirmed that our data of facial features fits the assumption of using single factor one-way ANOVA. If the assumption of homogeneity of variances is violated, the Kruskal-Wallis test will be used instead. We set the null hypothesis as ***H₀: The behavior (e.g. raw value) of the feature does not change differently when subjects are in low, medium or high engagement level.*** If the result of the significance tests shows a 95% likelihood (i.e. $p < .05$) that the results do not fit the null hypothesis, meaning that the null hypothesis can be rejected, that feature is chosen to be in the final feature set.

Feature	Descriptions	p value of one-way ANOVA
f 2	Standard Deviation of the location of the head - y	.053
f 3	Standard Deviation of the location of the head - z	.028
f 13	Count of the Existence of AU 1 (Inner Brow Raiser)	.099
f 14	Count of the Existence of AU 5 (Upper Lid Raiser)	.003
f 30	Count of the Existence of AU 45 (Blink)	.143
f 47	Presence of AU 28 (Lip Suck)	.008
f 49	Percentage of Frame with Low Confidence in Face Detection	.010

Table 4-4 Result of doing single factor one-way ANOVA test on the intermediate facial feature set.

Features in green color are having statistically significant difference under different level of engagement.

The results suggest that there are significant differences in half of the facial responses features, including f_3 , f_{14} , f_{47} and f_{49} . Table 4-4 shows whether the features behave significantly difference in each class. Table 4-5 lists the final set of facial response features.

Feature	Descriptions
f 3	Standard Deviation of the location of the head - z
f 14	Count of the Existence of AU 5 (Upper Lid Raiser)
f 47	Presence of AU 28 (Lip Suck)
f 49	Percentage of Frame with Low Confidence in Face Detection

Table 4-5 Final facial feature set.

For f_3 “standard deviation of the location of the head – z”, there was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,116) = 3.690$, $p = .028 < .05$). A Tukey post hoc test revealed that the standard deviation of the location of the z coordinate of the head was statistically significantly lower in median engagement level ($2.29 \pm 2.62\%$, $p = .027$) compared to the high engagement level ($5.11 \pm 8.64\%$). There was no statistically significant difference between the low and medium groups ($p = .965$) and low and high groups ($p = .553$). From the facial response videos, we observed that this feature is caused when subjects in high engagement tend to move closer to the monitor when they read questions.

For f_{14} “count of the existence of AU 5 (upper lid raiser)”, there was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,116) = 6.193$, $p = .003 < .05$). A Tukey post hoc test revealed that the number of AU5 exist in a segment was statistically significantly higher in median engagement level (4355.6 ± 2038.1 times, $p = .002$) compared to the high engagement level (3141.8 ± 1384.3). There was no statistically significant difference between the low and medium groups ($p = .991$) and low and high groups ($p = .187$). Our results imply that subjects raise their upper lid more frequent when they are under medium engagement than high engagement. Based on our observation, this feature is caused by two possible behaviors. The first one is illustrated in Figure 4-3: sleepy users, who belong to low or medium levels of engagement, tend to force their eyes open and try not to fall asleep.



Figure 4-3 Sleepy subjects in low or medium engagement who were trying to wake up themselves.

The second type of behavior was from users under low or medium engagement, who read the question and answer choices repeatedly. As the questions are usually 2 to 3 lines long and the answers are shown after the questions, subjects' eye balls will move upward and downward, which causes the upper eye lid to move together.

For f 47 “presence of AU 28 (lip suck)”, there was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,116) = 5.061$, $p = .008 < .05$). A Tukey post hoc test revealed that the average number of instances that contains AU 28 was statistically significantly higher in median engagement level ($.657 \pm .478$, $p = .005$) compared to the high engagement level ($.357 \pm .485$). There was no statistically significant difference between the low and medium groups ($p = .896$) and low and high groups ($p = .525$).

For f 49 “percentage of frame with low confidence in face detection”, there was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,116) = 4.784$, $p = .010 < .05$). A Tukey post hoc test revealed that percentage of frame with low confidence in face detection was statistically significantly lower in median engagement level ($2.27 \pm 4.85\%$, $p = .011$) compared to the high engagement level ($11.34 \pm 25.8\%$). There was no statistically significant difference between the low and medium groups ($p = .993$) and low and high groups ($p = .189$).

Unexpectedly, the percentage of frames with low confidence in face detection in high engagement group is on average higher than that of the medium engagement group. Inspecting the data suggests that this phenomenon is caused when users with high engagement move closer to the monitor so as to read the questions more carefully, which raises the probability that the face area does not completely fall within the view of the webcam. Even though subjects with medium level engagement tends to move around and sit back, their full facial area is still within the camera viewport. Therefore,

detecting the distance between subject and monitor may help to represent this behavior as a feature to improve model's performance.

4.3 Engagement Detection with Facial Expression

In order to detect the level of engagement, we used a 3 class Support Vector Machine (SVM), with RBF (radial basis function) kernel using C (C=1.0) as the default parameter, to train models. All prediction models are constructed using the SMO method implemented in the Weka data mining tool.

As summarized in Table 4-6, we focus on comparing the performance of models with different combination of test (Language Learning task vs Web Searching task) and feature sets (*FeatureSet* without ANOVA vs *FeatureSet* with ANOVA). Here, *FeatureSet* without ANOVA indicates the feature set generated by the first two steps of feature selection, i.e. the intermediate feature set in Table 4-3. *FeatureSet* with ANOVA is the final feature set, listed in Table 4-5, that is generated after using single factor one-way ANOVA to analyse and select features from *FeatureSet* without ANOVA.

	Dataset	Feature Set	\overline{CCR}	Baseline	ΔCCR	F1
M_{Face1}	Language Learning Task	FeatureSet without ANOVA	63.9%	58.8%	5.1%	.609
M_{Face2}		FeatureSet with ANOVA	68.1%		9.3%	.589
M_{Face3}	Web Searching Task	FeatureSet without ANOVA	62.3%	55.8%	6.5%	.570
M_{Face4}		FeatureSet with ANOVA	63.6%		7.8%	.595

Table 4-6 Summary table of the results of the models that are using different of facial feature set and dataset.

Since the features were selected from the Language Learning Task dataset, we run a 10 times 10-fold cross validation with 119 data instances to understand the performance of the model. We use 9 folds of Language Learning data as the training set and test on the remaining fold. The process is repeated until all folds have been tested. The average performance for the 10 experiments is reported.

As a baseline, we use the initial facial feature set to train the same SVM model under 10 times 10-fold cross validation evaluation. This gives us a baseline of 58.8% correctly classified rate.

We first use *FeatureSet* without ANOVA, which contains 7 facial features, to build a model (M_{Face1}) and test on Language Learning data. M_{Face1} reaches a correctly

classified rate (CCR) of 63.9%, which is 5.1% higher than the baseline. We then built a model (M_{Face2}) with *FeatureSet* *with ANOVA* that reaches 68.1%, 9.3% above baseline.

To test the generalizability and representativeness of our features, the feature sets are further tested on the Web Searching dataset. Therefore, we built two models and evaluate with 10 times 10-fold cross validation on the Web Searching dataset.

Similar to the Language Learning Task, the initial facial feature set was used to train the SVM model and test on Web Searching Task data to get a baseline. 10 times 10-fold cross validation gave us an average correctly classified rate of 55.8%.

The third model (M_{Face3}) using *FeatureSet* *without ANOVA* was built and tested on the Web Searching dataset. The model is able to classify 61.0% of instances correctly, which is 5.2% higher than the baseline. The fourth model (M_{Face4}) using *FeatureSet* *with ANOVA*, tested on Web Searching tasks dataset, achieved 63.6% CCR, which is 7.8% higher than the baseline.

We run the 1-tailed Paired T-Test with 10 degrees of freedom on the 10 times 10-fold cross validation. The results of the test for models (M_{Face1} and M_{Face2}) [$t(10) = 2.5545$, $p = 0.014 < 0.05$], and models (M_{Face3} and M_{Face4}) [$t(10) = 2.4891$, $p = 0.016 < 0.05$] show that the classification results of the listed groups of models are significantly different. These results are encouraging as they reveal that models using *FeatureSet* *with ANOVA* outperform the models using *FeatureSet* *without ANOVA* in both datasets.

4.4 Extending Facial Affect Recognition into Real Life

Affect exchange is essential for healthy physical and social development [32], and friends and family communicate their emotions to each other instinctively. In particular, watching movies has always been a popular mode of socialization and video sharing is increasingly viewed as an effective way to facilitate communication of feelings and affects, even when the parties are not in the same location.

Current state-of-the-art work in video sharing generally involves users explicitly and intentionally giving feedback on the video, either in the form of votes or comments. This allows for the sharing of very detailed and precise information, but it also requires explicit user effort as well as a certain degree of comfort with the computer.

We extend on previous work by investigating the potential of automated affect capture in facilitating implicit sharing of experiences. We use facial affect recognition techniques to detect the users' emotions, which are then used to label specific frame sequences in the video. This allows a spontaneous reaction that is precisely linked to point in the movie that triggered it. It also allows for users who may not be comfortable or literate with the computer.

In this sense, Cui et al.'s work [27] is similar to ours. They investigated emotional communication between close-knit individuals via photo-sharing by embedding viewer's facial response (an image) into the shared photo (i.e. the media). Their finding suggests that the shared response can help to convey feelings and create a sense of co-presence. However, conveying feelings through video involves more complicated issues.

To visualize human emotional states while encouraging people to share their emotional response, we introduce the Emotar [68], which is a combination of emoticon and avatar. Aoki et al. [7] pointed out that emoticons could express emotions that cannot be adequately communicated in words. Janssen et al. [59] suggested that emoticon used in communication increases the perceived intimacy. Derks et al. [29] considered emoticons as a visually salient representation of expression, which has the potential to be as rich as expressions in face-to-face interaction.

4.4.1 Asynchronous Video-Sharing Platform with Emotars

As an emotional avatar, Emotars allow viewers to share their affects and experience with others without disclosing their real facial expressions and/or features. This has implications on user comfort and privacy, and we believe that using emoticons instead of photos of the user's facial response helps to increase the level of comfort and security, which encourages the users to be more willing to share with others. From the communication point of view, emoticons provide an efficient abstraction of viewers' emotional states. It takes less time to process, hence causing less distraction to receivers, which in turn results in higher overall efficiency in communication.

Figure 4-4 shows the system interface, which is implemented in HTML5 and can be accessed through the Internet as a web page. The movie plays in the center and Emotars are displayed on both sides of the movie window. Each Emotar represents one person. To communicate the affect more effectively, we designed the Emotar with a background color that changes according to the emotion currently represented [31].



Figure 4-4 User interface of the video sharing platform with Emotar.

We investigate the efficacy of our system with two videos: one that is commonly regarded to be funny, and another that is commonly considered to be sad. The two movies are therefore designed to induce happiness and sadness, which are considered to be two primary, universal feelings. The viewers' affect was classified into seven types: ecstatic, pleased, neutral, down, depressed, interested, and disgusted. Each Emotar therefore has seven forms (or appearances) that correspond to these types.

To accurately represent the temporal change of viewers' facial expressions, we trained user-dependent facial affect models for individual subjects. 10 images were

captured per subject for each type of affect. We then used the 24 facial features extracted from these calibration images with Support Vector Machines (SVMs) to train a facial affect model on the calibration data for each user [82]. Our facial affect model therefore gives us the frame level affect sequence corresponding to the viewers' facial response, which then drives the Emotar.

Since our objective is to explore the communication of affects and not to build a facial affect recognition system, we manually validated the automated facial affect recognition model and made necessary corrections to compensate for large head pose variance. This ensures that the Emotar correctly represents the viewers' emotions. In total, about 22% of the frames needed to be corrected.

4.4.2 Evaluation

We recruited 32 unpaid participants (21 female), aged between 19 to 47, to use the system. We used a Tobii EyeX Controller to track the gaze points of the subjects. This allowed us to analyze the gaze behavior, including the moments when the subjects' gaze switches from the video to the Emotars. Facial responses of each subject were recorded down and were used to share to their friends who did the experiment later. After the experiment, we conducted a face-to-face focus interview with each subject.

In order to evaluate the system, we are interested in the following issue:

- **Emotion awareness:** do the subjects have a clear and correct awareness of other viewers' affects while they watch the video? How is such awareness influenced by the Emotar?
- **Engagement:** how do the Emotars and emotion awareness influence the subjects' engagement during video watching?
- **Comfortableness:** how comfortable are users with our interface? Are they willing to share their own feelings and responses if they are represented in such a manner?
- **Relationship:** how do differences in factors such as relationship and gender between the viewer and the person represented by the Emotar impact the above issues?

Emotion Awareness and Relationship

We have selected 2 males' and 2 females' Emotars to be shown on the screen and within this 4 Emotars, 1 male and 1 female were the person that the subject knew.

We calculate the probability of the user's emotion intensity reaching a peak within 3 seconds before and after a gaze saccade to the Emotars. The figure suggests that the user is more likely to have experienced an increase in emotional intensity before the gaze saccades occurs. In other words, users are more likely to look at the Emotar after something has triggered an intense emotion.

Not surprisingly, subjects were more interested in the Emotar of a person that they knew. 70% of the Emotar glances were directed towards an Emotar of a close friend or family member.

Interestingly, we observed that people were more likely to look at the female Emotars. 53% of the gaze saccades from female subjects were directed to female Emotars. This is even more pronounced with male subjects: 59% of their saccades were to female Emotars.

Engagement during Video Watching

Encouragingly, the experiment results suggest that the increased emotional awareness may enhance the level of engagement. We replayed one part of the happy video, which is not commonly considered funny, to the interviewees, and asked whether they considered that part to be funny. Most of them answered "No" or "Not really". We then reviewed their facial features during that time period and found that some were actually smiling despite reporting that they did not find that part funny. When asked the follow-up question "Then why are you smiling?", some users pointed out that they smiled because they saw that their friends' Emotars were in the highest "happy" state.

Comfortableness

When people are engaged in a video sharing activity, privacy and security are extremely important. People may not object to communicating feelings with friends, but they are not comfortable posting actual videos of themselves (as a means of communicating emotions). Using an Emotar to represent a person could help to communicate one's facial expressions while still maintaining privacy and anonymity.

Among all 32 subjects, 21 (66%) stated that they would be willing to use our system to share their movie-watching experience with their friends. However, many of them were not willing to share a video of their face. They also point out that a system that uses their actual images instead of Emotars would make them wary. Since people consider the risks and weigh them against the benefits when they decide to disclose information [81], and trust is the key to disclosure in online relationships, this may be due to a lack of trust in the social platform [55].

4.5 Summary

In this chapter, we propose using facial expression as a modality to recognize different engagement levels during daily computer interaction tasks. Our method uses signals captured by an off-the-shelf webcam with the resolution of 640 x 480 pixels. We extracted facial features and filter them with a 3- steps feature selection process. User-independent models are then built and evaluated on two different datasets and results show that (1) the features selection method helps to remove useless features from the initial feature set without bring harms to the models, and (2) the selected features are general enough to apply in another type of task and still perform reasonably.

We also developed the asynchronous video-sharing platform with Emotars that apply the results of affect detection from facial expression for video-sharing and evaluated the user experience and their behaviors of using such platform.

Chapter 5 Understanding Users from Eye Movement

Affective computing is a major area for research. Research focusing on enabling computers to understand human emotions, attentions or interactions have the potential to allow human intention to be predicted and better service to be provided.

In this chapter, we investigate detecting the level of engagement on tasks, which is one of the form of human affect. We specifically focus on eye gaze behaviors in this chapter. Eye gaze behaviors could be captured by different methods; and we focus on eye gaze signals captured by a commercial eye tracker (Tobii EyeX Controller) and a standard webcam with resolution of 640 x 480 px.

We invited subjects to carry out experiments in Language Learning tasks and Web Searching tasks to induce different levels of engagement. Feature selection was applied to extract useful features. Similar to the process for facial features, we only use Language Learning data for feature selection. We then construct user-independent models for engagement level detection and test the models in 2 different tasks, which are (1) Language Learning task and (2) Web Searching task.

We find that our selected webcam-based gaze features achieve a performance improvement of 4.2 % above baseline for testing and training on Language Learning tasks, and 10.4% above baseline for testing and training on Web Searching tasks. Meanwhile, the selected Eye Tracker-Captured gaze features achieve a performance improvement of 8.4% above baseline for testing and training on Language Learning tasks, and 9.1% above baseline for testing and training on Web Searching tasks.

When we combine both webcam-based and Eye Tracker-Captured gaze features for engagement detection, it could even reach higher performance. It reaches a performance improvement of 12.6 % above baseline for testing and training on Language Learning tasks, and 14.3% above baseline for testing and training on Web Searching tasks. These results show that the selected gaze features are useful for detecting level of engagement even with different tasks contents.

5.1 Introduction

Doing homework and Web Searching are parts of the common computer interaction activities which happen in daily life. In the same way of teachers might

observe their students' level of engagement in class, it is rational to believe that having good understanding of user level of engagement during tasks and successfully detecting the engagement level would be helpful to provide in time assistance or suggestions for taking a rest or providing something to raise their attention.

Recent advances in hardware and sensors created the opportunity of using new modality for getting interaction signals for public. One of opportunities is to use the commercial eye tracking sensors for understanding users gaze behaviors. There are games and applications built based on these devices.

Moreover, commercial eye tracking devices are not the only modules for getting eye gaze signals, webcam is another promising module. Since webcam is equipped on most of the computers or laptops, they can easily and ubiquitously be used to capture users' head features, including their movements. There has been much research on eye movement tracking and gaze point prediction with webcam signals [121]. However, to track ones' gaze location with webcam require long period for calibration. Instead of knowing the gaze location of the user, the type of eye movements (fixation, saccades, smooth pursuits) can be more easily identified from webcam signals. Therefore, in this chapter, we also investigate the possibilities of extracting eye movement features from webcam eye information.

The rest of this chapter is organized as follows. In Section 5.2, the types of eye movements and our definition of each type of eye movements are described. Section 5.3 describes the features extracted from commercial eye trackers (Tobii) and the webcam. Section 5.4 describes the features extraction and selection process of webcam eye movement data. The results of the models using different gaze features would be discussed in Section 5.5. Finally, we conclude this chapter briefly in Section 5.6.

5.2 Types of Eye Movements

In this section, we describe the types of eye movements and their definitions. Human gaze behaviors can be categorized into four behaviors: fixation, smooth pursuit, saccade and blink [52]. To analyze the gaze behavior, we used a Tobii EyeX Controller to track gaze points of the subjects, and use the gaze x- and y- coordinate for gaze behaviors extraction. We use dispersion-threshold identification [99] to identify fixations and use velocity-threshold identification [99] to identify saccades. Meanwhile, the Tobii EyeX Controller outputs the x- and y- coordinate of the gaze points which are normalized to [0,1].

There are a few challenges in working with the raw eye gaze data extracted from Tobii EyeX Controller. Firstly, the eye tracker may fail to detect one eye or even both eyes on occasion. The eye tracker reports a normal pair of (x,y) coordinates if only one eye is detected, and (-1,-1) if it fails to detect both eyes. Failure in detecting both eyes usually happens when the user blinks, or moves their head rapidly or move out of the effective recording area. Secondly, the Tobii EyeX is a basic commercial version, which unlike the research-grade version, does not report its precise accuracy or sampling rate. Therefore we do not know how accurate the predicted eye gaze locations are.

All subjects are required to calibrate the eye tracker before the experiment starts, and there is an interface for testing the calibration results. In this interface, shown in Figure x, there are 9 circles with dots inside each circle. Subjects are required to gaze at the small dot and the experiment starts if all predicted gaze locations are within the circle of the dot they are gazing at. Therefore, we use the size of the circle as the accuracy of the eye tracker.

Eye Blink

Given an eye gaze sequence with normalized x and y coordinates of gaze location, eye blinks can be detected by locating the moments of (-1,-1) data points.

Some previous work [110] has suggested that the duration of a blink would be on average 100-150 milliseconds. Other work suggests a higher upper bound: between 100-400 ms [94].

We therefore label eye closures between 100 ms to 400 ms as eye blinks. If the eye is closed or not tracked for less than 100 ms or more than 400 ms, we define it as a failure or error situation. We observed that eye closures that are less than 100 ms are usually due to noise and that more than 400 ms are due to failures from the eye tracker, or when the user looks outside the screen or consciously closes their eyes (e.g. rest).

Eye Fixation

Fixation means a stationary gaze on a single location. During the act of fixating, the eyes are relatively stationary. Crouzet [26] mentioned that “If a dependence on task status is needed to be able to infer that a particular neural response is related to high-level processing, it would be natural to conclude that no influence of such high-level factors is visible until around 180 ms in this task.”. Meanwhile, Dahal’s [103] analysis result shows that fixations usually last for 180 ms.

At the same time, the error present in the eye tracker makes fixation detection from the gaze signal more than just simply finding periods in which the eye coordinates stay still. Since the Tobii EyeX Controller, which we used for eye tracking, did not provide any official performance value, we use the size of circle shown in the calibration interface as its performance. We measured the radius of the circle in pixels, which is 80 pixels for the monitor used in the experiment. This radius was then used as the one of the rules for detecting fixation.

We therefore define fixation as a period of longer than or equal to 180 ms during which the variation of gaze location is less than or equal to 80 pixels of radius.

Eye Saccades

After eye fixations and blinks are identified from the eye gaze sequence, the sequences in between fixations may consist of eye saccades or smooth pursuit. Thus, we further consider those sequences and try to identify the other two types of eye movement.

Eye saccades are behaviors that rapidly move towards a target. It usually last for around 20 to 40 ms. Erkelens [25] defined saccades as gaze movements with speed larger than or equal to 40 degree of visual angle per second. Thus, we define eye saccades as any gaze movement with speed larger than or equal to 40 degrees of visual angle per second.

In addition to identifying eye saccades, we also identify saccades during reading texts as regressive or non-regressive saccades. Given the nature of the tasks, we first classify the eye saccades during reading lines of words into two types of saccades: forward saccades and backward saccades. Forward saccades imply the eye follow the direction of the text during normal reading, in our case, it is from left to right as the context are in English. In contrast, backward saccades indicate those eye gaze locations move in the direction against the flow of text, i.e. from right to left. With the information of backward and forward saccades, we could identify regressive saccades from the sequence. Within a sequence of eye gaze location in between two eye fixations, if both backward saccades and forward saccades exist for more than one time, it will be classified as regressive saccades. Figure 5-1 illustrate the examples of forward, backward and regressive saccades.

Forward saccade:



Backward saccade:



Examples of regressive saccade:

1 :



2 :

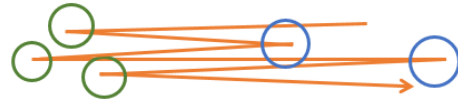


Figure 5-1 Illustration of forward, backward and regressive saccade.

To check whether the eyes are gazing at words or not, we need to first identify the location of the paragraphs, lines and words. However, extracting the regions of the contents is not an easy task. Shen et al. [21] successfully extracted two major components of web content text and image by applying image processing to each screen shot of the web page. They applied low pass filter to eliminate text regions and extracted image regions from screen shots, while applied high pass filter to obtained text regions. Instead of looking at the spatial frequency, we decided to use another image processing approaches to extract content information. We dilate and erode the screen image with different scale such that we could identify the location of paragraphs, lines and words.

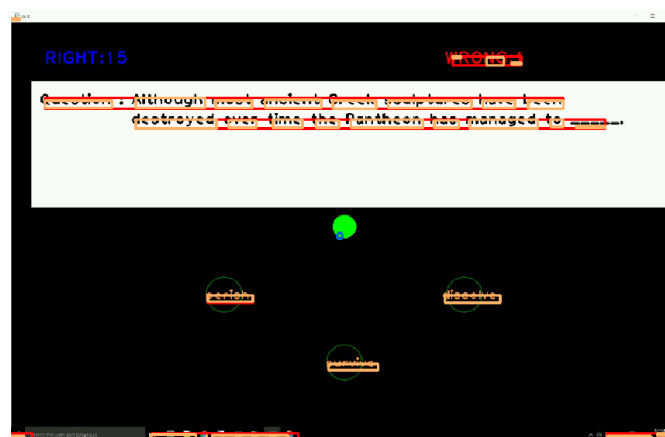


Figure 5-2 Results of content extraction. Boxes in red are the lines detected and boxes in orange are the words detected.

Figure 5-2 showed the example of content extraction. After obtaining content information, we can identify the content that the user is gazing at. If the subject is gazing

at words, we will check if regressive saccades are happening.

As we would like to reduce the probability of labeling a saccade caused by line changing as regressive saccades, we not only check the existences of both forward and backward saccades within a sequence, but also check for backward saccades, defined as moving back to a previous word or a word in next line. This was done by considering the line information extracted by image processing the captured screen data.

Smooth Pursuit

Smooth pursuit usually happens when a subject is reading the text and slowly move from one word to another word. Compare to eye saccades, smooth pursuit has relatively slow gaze movements normally with speed of 30 degree of visual angle per second. There are also some related works which suggest that the upper limit of smooth pursuit could be 100 degree per second [50]. By considering the speed of smooth pursuit to be between 30 to 100 degrees per second and the definition of saccades used in our work, we define smooth pursuit as gaze movement with speed less than 40 degree of visual angle per second, which essentially are eye movements that are not classified as eye saccades.

5.3 Tobii Features

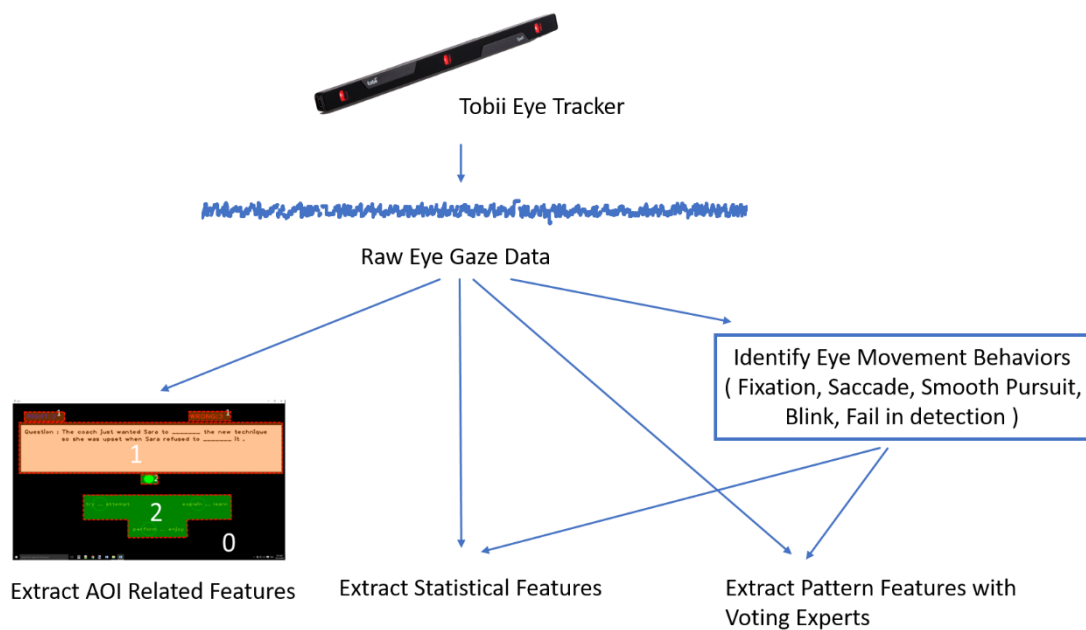


Figure 5-3 Flow of Extracting Eye Tracker-Captured gaze features.

We used the Tobii EyeX device to track the gaze points of the subjects and obtained the gaze x- and y- coordinate sequences for each instance. Within each sequence, we classify the 5 types of eye movement: blink, fixation, smooth pursuit, saccade and fail in tracking the eyes; and a new sequence containing information of types of eye movement was generated.

The aim of our work is to automatically detect the engagement level of user when they are doing tasks, such as searching task or homework. Given this objective, we consider the eye gaze behaviors within a given period of working on a task.

We tried to extract eye movements features by three different methods, as shown in Figure 5-3. The first one is to make use of the eye gaze location and the area of interest (AOIs). The second one is to recognize some special eye movement patterns. Last but not least, we use the basic statistical information of the sequences as the features.

5.3.1 Eye Gaze Location and AOIs



Figure 5-4 Regions of the Language Learning Task. Fixations in 0 refers to the interest on other things; Fixations in area 1 refers to information receiving and in area 2 refers to decision-making.

Considering the nature of the tasks, including Web Searching tasks and Language Learning tasks, it is reasonable to segment the screen into 3 different Areas of Interest (AOIs), shown in Figure 5-4. These correspond to the type of actions that the user is taking, which include information receiving, decision-making and others. We then locate the gaze data for every fixation with respect to an AOI. This allows us to build a histogram of AOIs transition for each segment, each histogram essentially acting as a summary of the distribution of different user behavior transitions.

To find the distribution of different combination of transition, we consider the last

3 fixation AOIs . For example, a subject may fixate on a word in the question, then fixate again on another word and finally fixate on one of the possible answers in the decision-making area. In this case, we will get a sequence of eye gaze transition starting from information receiving to information receiving to decision-making. We count the frequencies of all combinations and then normalized the values by the total number of transition. All the values of the histogram are the potential features that could be used. In total, there are 30 potential AOIs related features, as shown in Table 5-1.

Feature	Meaning	Formulation
e AOI 1-27	Gaze transition pattern in AOIs	Probabilities each of the eye gaze transition pattern exists
e AOI 28-30	Time fixated in each AOI	Percentage of time fixated in each AOI

Table 5-1 List of AOIs Related Eye Tracker-Captured Gaze Features.

5.3.2 Eye Movements Behaviors and Voting Experts

We previous described the process of processing a sequence of eye gaze data into a sequence of eye movement behavior consisting of 5 types of eye movement: blink, fixation, smooth pursuit, saccade and fail in tracking the eyes in each gaze sequence. With the eye movement type sequence and eye gaze transition sequence, we would like to automatically extract some common and special patterns as features.

We used “Voting Experts” [91], which is an unsupervised algorithm for segmenting sequences, for segmenting the data automatically. An N gram tree was built and normalized frequency and boundary entropy were then calculated. Afterward, a segmentation score with 2 experts, with 1 expert vote based on the frequency and 1 expert vote based on entropy, is calculated. With the segmentation score, we select the locations to segment according to the zero crossing rule, window size and vote threshold.

With the segmented sequences, we further use k-means clustering to cluster the segments into clusters. Kocyan et. al. [109] suggested to get cluster representative of the segmented sequence using Dynamic Time Warping (DTW). Thus, we ran the k-mean clustering (k = 20) with DTW to choose the representative for 10 iterations. After 10 iterations of clustering, the cluster representatives will be chosen if it contains more than 1 cluster member and 80% of the members belong to the same class.

After clustering, as shown in Table 5-2, 11 patterns were selected from the eye behavior sequence and 6 patterns were selected from the AOIs transition sequence. With the selected patterns, we use the frequency of each existing patterns and whether the pattern happens or exists within that segment. In total, we obtain 34 potential features from the eye movement behaviors.

Feature	Meaning	Formulation
e VE 1-22	Eye Behavior Type Sequence (Fixation, Smooth Pursuit, Saccade, Blink and Detect Failure)	(e VE 1-11) Number of existence, (e VE 12-22) Presence of eye behavior pattern existences
e VE 23-34	Gaze based AOIs Transition Sequence	(e VE 23-28) Number of existence, (e VE 29-34) Presence of gaze based AOI pattern

Table 5-2 List of Pattern Related Eye Tracker-Captured Gaze Features.

5.3.3 Statistics of Eye Movements

Apart from finding the eye movement pattern, some statistical descriptions were extracted to represent eye movement behaviors within a segment. For both type and movement sequence, we calculate some descriptive statistics as features. In total, 34 potential features are extracted, which are listed in Table 5-3.

Feature	Meaning	Formulation
e stat1-3	Fixation	(1) Mean of Attribute Duration, (2) Standard deviation of Attribute Duration, and (3) Percentage of time covered by the attribute
e stat4-6	Smooth pursuit	
e stat7-9	Saccade	
e stat10-12	Detect Failure	
e stat13	Blink	(1) Count number of Eye Blinks
e stat14	Regressive Saccade	(1) Count of regressive saccades
e stat15-19	Travel distance	(1) Mean, (2) Maximum, (3) Minimum, (4) Median and (5) Standard Deviation of all eye movements within one segment
e stat20-24	Straight distance	
e stat25-29	Straight to Travel Distance Ratio	
e stat30-34	Movement speed	

Table 5-3 List Statistical Eye Tracker-Captured Gaze Features.

5.4 Eye Behaviors from Webcam

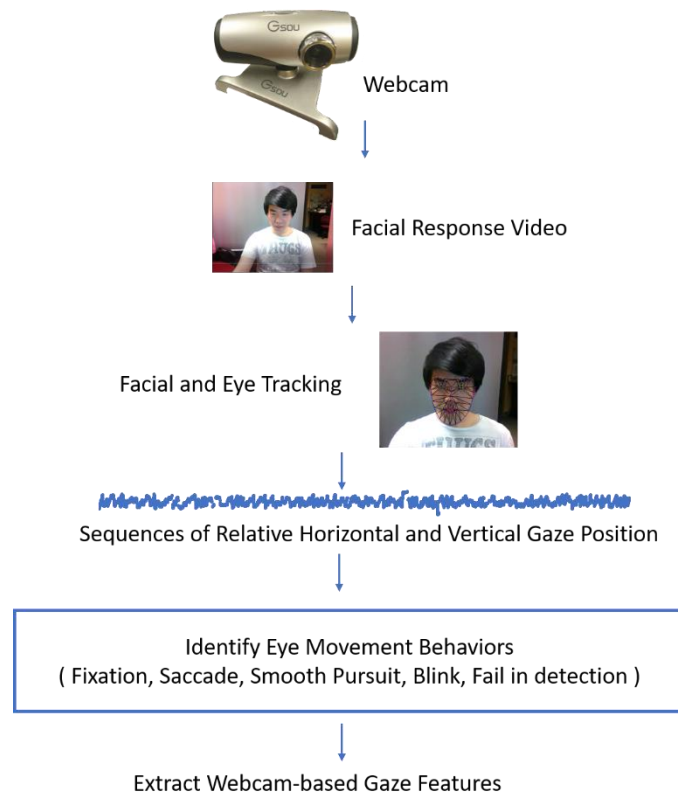


Figure 5-5 Flow of Webcam-based Gaze Features Extraction.

We also extract eye gaze features from the facial response videos using eye behavior recognition. In this section, we analyze and classify the types of eye movement, eye fixations, eye saccades and smooth pursuits, from the video. We do not consider the information of eye blinks and failure of detecting eyes as they are already represented by the facial features mentioned in Chapter 4. We still identify eye blinks and failure of detecting eyes from the videos but we do not extract features related to these two types of movements.

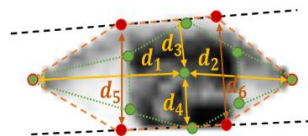


Figure 5-6 The eye detected by CLM Face (dots in red color) and CLM Eye (dots in green color) models, and the distance that we could calculated from the landmarks.

First of all, we need to identify users' eyes in the videos and extract the eye

landmarks. As presented in Section 4.1.1, we use SDM fitting results with 48 facial landmarks as the initial states of CLM fitting and finally it located 66 facial landmarks in each frame in the video. Out of the 66 facial landmarks, we identified 6 landmarks for each eye. Figure 5-6 shows the 6 landmarks in red that located on upper and lower eye lid and eye counter. To describe the eye gaze behavior accurately, the pupil center need to be located for the purpose of estimating gaze direction. However, unconstrained situations lead to the limitations of appearance information of eye in the responses videos fails to provide clear cue for identifying pupil center. Meanwhile, in real-use scenarios, low video resolution and reflections on cornea and glasses usually marks pupil unobservable. Thus, techniques that based on edge detection often fail to track the pupil center reliably.

To solve this, we use eye geometry to estimate the pupil center by tracking landmarks on the iris contour and eye lid corners deduce pupil center locations based on their geometric interdependency. CLM based on eye [56] is applied to track the 9 eye landmarks including the pupil center. Based on the 9 landmarks identified by the eye CLM and 6 landmarks from face tracking model, we use the landmark distances for each eye to compute 2 sequences: the relative horizontal and vertical eye gaze positions. As shown in Figure 5-6, $d1$ and $d2$ are the distance between two eye corners to the pupil center, $d3$ and $d4$ are the vertical distance between upper/lower eye lid and the pupil center. From this, the relative horizontal eye gaze positions $(\frac{d1}{d1+d2})$ and relative vertical eye gaze positions $(\frac{d3}{d3+d4})$ could be calculated. The temporal changes in relative horizontal and vertical eye gaze positions are then adopted in eye fixations, saccades and smooth pursuit.

5.4.1 Webcam Based Eye Gaze Features

We compute the relative eye gaze positions within each eye, which is independent of the actual coordinates of the eye. The eye movements over a temporal period can be analyzed from the eye gaze relative position sequence. Considering that most humans move left and right eyes together in normal situations, we simplify the eye gaze position by calculating the average value of both eyes. This gives us the eye gaze position sequence consisting of the information of relative averaged horizontal and averaged vertical eye gaze positions.

Eye fixations are defined to be periods during which the eye gaze stays still in a

location. However, because of error caused by head movement and landmark jittering of the eye CLM, fixation detection from signal becomes more than finding periods in which the gaze position does not change. We ran a pilot study for the purpose of determining the extent of noises during fixation detection. In this study, we asked 2 subjects to fixate on the mouse cursor as it is pointed at different locations on the screen, and manually remove the frames in which the mouse cursor is moving and the eye is following the mouse cursor. The eye gaze positions sequences are extracted. The eye gaze positions difference detected by the eye CLM model between successive frames are used to compute the standard deviation and mean value of eye movement during fixations. To filter the noises in the eye gaze positions sequences, we define the fixation threshold as $(\text{Mean fixations' gaze movement} + \text{Standard deviation of fixations' gaze movement})$. If the difference in eye gaze position between successive frames is less than the defined threshold, we will mark it as stationary. Considering that eye fixations normally last for around 180 ms (previously mentioned in Section 5.2), we label continuous stationary sequences that last for more than 180ms as fixations.

We also ran a pilot study to obtain the thresholds of saccades. The 2 subjects are asked to first fixate on one point and then look quickly at another point. The periods of eye saccades are manually segmented and used to compute the mean value and standard deviation of eye movements between successive frames. We then defined the saccade threshold as $(\text{Mean saccades' gaze movement} - \text{Standard deviation of saccades' gaze movement})$. Any eye gaze movements located in between the fixations with movement larger than the saccades threshold will be classified as eye saccades, others will be classified as smooth pursuit. We calculate some descriptive statistics as features, which are listed in Table 5-4.

Feature	Meaning	Formulation
e webcam 1-4	Fixations	(1) Mean Duration, (2) Standard Deviation of Duration, (3) Count, and (4) Percentage of times exist in the segment of the eye movement behaviors.
e webcam 5-8	Smooth Pursuit	
e webcam 9-12	Saccades	
e webcam 13-15	Ratios Between Numbers of Different Eye Movement	

Table 5-4 Initial Webcam-based Gaze Feature Set.

5.5 Using Eye Gaze Interaction for Engagement

Detection

5.5.1 Features Selection

There are in total 113 potential features extracted from users' eye movement data, but not all of them are useful. Therefore, feature selection is needed to identify relevant features. The 3-step feature selection shown in Figure 4-2 was applied to the eye interaction features.

Features Left After Features Selection Step 1 & 2

We adopt the correlation attribute evaluation for the feature selection, which considers the Pearson' correlation between an attribute and the class label. We use CorrelationAttributeEval with ranker search method provided by Weka [51] for feature selections and filter out all features with less than 0.3 correlation.

Step 2 uses the features selected in Step 1 to build a base model and adds other features one by one to check if it could bring significant increase in performance. The significance test was done by running 10 times 10-fold cross validation with 10 degrees of freedom Paired T-Test. After Steps 1 and 2, the following features are left. Table 5-5 shows the selected webcam-based eye gaze features and Table 5-6 shows the selected Eye Tracker-Captured eye gaze features.

Feature	Descriptions
e webcam 1	Mean Duration of Fixations
e webcam 3	Number of Fixations
e webcam 11	Number of Saccades
e webcam 12	Percentage of Saccades Times Exist in the Segment
e webcam 13	Ratio between Count Fixations and Count Pursuit
e webcam 14	Ratio between Count Fixations and Count Saccades
e webcam 15	Ratio between Count Pursuit and Count Saccades

Table 5-5 Intermediate Feature Set of Webcam-based Features.

Feature	Descriptions
e AOI 15	Probabilities of eye gaze transition pattern [1>1>2]
e AOI 18	Probabilities of eye gaze transition pattern [2>0>0]
e AOI 23	Probabilities of eye gaze transition pattern [2>1>1]
e AOI 24	Probabilities of eye gaze transition pattern [2>1>2]
e AOI 26	Probabilities of eye gaze transition pattern [2>2>1]
e AOI 27	Probabilities of eye gaze transition pattern [2>2>2]
e AOI 30	Time fixated in decision making AOI
e stat 13	Count of eye blinks
e stat 14	Count of regressive saccades

Table 5-6 Intermediate Feature Set of Eye Tracker-Captured Features.

Final Feature Set selected after Step 3

We then performed a single factor one-way ANOVA to determine whether the eye gaze interaction features exhibit differently under different levels of engagement. Before we run the test, we also checked the required assumptions, including but not limited to homogeneity of variances, no significant outliers and consist two or more independent group etc., and confirmed that our data fits the assumption of using single factor one-way ANOVA.

Webcam-based Feature Set

Table 5-7 shows the significance values of the webcam based eye gaze features set generated from step 1 and 2. The results suggest that there are statistically significant differences in most of the webcam-based eye gaze features as exhibited across different engagement levels, except e webcam 1 “the mean duration of eye fixation”. These results indicate that the behavior of features “e webcam 3” and “e webcam 11-15” is significantly different between different engagement levels.

Feature	Descriptions	p value of one-way ANOVA
e webcam 1	Mean Duration of Fixation	.166
e webcam 3	Number of Fixations	.036
e webcam 11	Number of Saccades	.000
e webcam 12	Percentage of Saccades Times Exist in the Segment	.000
e webcam 13	Ratio between Count Fixation and Count Pursuit	.000
e webcam 14	Ratio between Count Fixation and Count Saccade	.001
e webcam 15	Ratio between Count Pursuit and Count Saccade	.000

Table 5-7 Result of doing single factor one-way ANOVA test on the intermediate Webcam-based feature set. The behavior of the features in green color are statistically significantly different under different levels of engagement.

To understand more about the differences between pairs of individual criteria, we analyse the data with the Tukey-Kramer method. The Graph shown in Figure 5-7 (a-f) is the summary of the mean values of each feature in each class. If there is significant difference between pairs of individual criteria for that attribute, the significant values and the pairs are indicated.

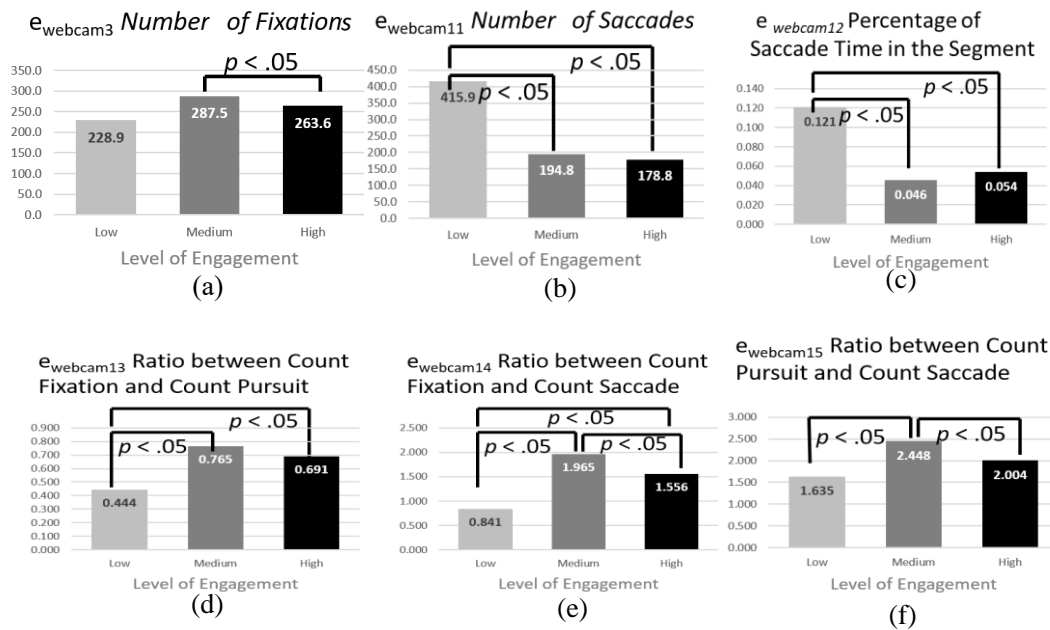


Figure 5-7 (a-f) Summary figures of comparing differences between pairs of individual criteria.

Results for $e_{\text{webcam}3}$ “Number of Fixations”, shown in Figure 5-7 (a), suggested significant differences exist between medium and high engagement ($p < .05$). This reveals that subjects tend to fixate more when they are in medium engagement than in high engagement.

Results for $e_{\text{webcam}11}$ “Number of Saccades” and $e_{\text{webcam}12}$ “Percentage of Saccades Times Exist in the Segment”, shown in Figure 5-7 (b & c), suggested significant differences exist between low and high engagement ($p < .01$) and low and medium engagement ($p < .01$). This reveals that subjects tend to have more saccades and spend more time on saccades when they are in low engagement than in medium and high engagement.

Results for $e_{\text{webcam}13}$ “Ratio between Count Fixation and Count Pursuit”, shown in Figure 5-7 (d), suggested significant differences exist between low and high engagement ($p < .01$) and low and medium engagement ($p < .01$). The mean value of $e_{\text{webcam}13}$ in low engagement is significantly lower than that in medium and high

engagement, which may cause by (a) less number of fixation or (b) more number of pursuit exist in low engagement instances.

Inspecting the data shows that subjects with low engagement tend to “read” the questions without processing/fixating on each word in the question. They then tend to have many saccades, presumably to re-read the question, and jumping between the question and answers. Subjects with medium and high engagement level, however, read by fixating on a word, followed by a smooth pursuit to another word and fixating again.

Results for e webcam 14 “Ratio between Count Fixations and Count Saccades” and e webcam 15 “Ratio between Count Pursuit and Count Saccades”, shown in Figure 5-7 (e & f), suggest significant differences exist between low and medium engagement ($p < .05$) and medium and high engagement ($p < .05$). This reveals that subjects tend to make more saccades and spend more time on saccades when they are in low engagement than in medium and high engagement.

Eye Tracker-Captured Feature Set

Table 5-8 shows the significant values of the Eye Tracker-Captured eye gaze features set generated from step 1 and 2. The results suggest that there are statistically significant differences in most of the Tobii-based eye gaze features, except e stat 14 “count of regressive saccades”. (The indexing of the AOIs is as follows: AOI 1 is the area for information receiving, AOI 2 is the area of decision making and AOI 0 is any area other than AOI 1 and 2. The AOIs are also shown in Figure 5-4.)

Feature	Descriptions	p value of one-way ANOVA
e AOI 15	Probabilities of eye gaze transition pattern [1>1>2]	.000
e AOI 18	Probabilities of eye gaze transition pattern [2>0>0]	.000
e AOI 23	Probabilities of eye gaze transition pattern [2>1>1]	.000
e AOI 24	Probabilities of eye gaze transition pattern [2>1>2]	.000
e AOI 26	Probabilities of eye gaze transition pattern [2>2>1]	.000
e AOI 27	Probabilities of eye gaze transition pattern [2>2>2]	.000
e AOI 30	Time fixated in decision making AOI	.000
e stat 13	Count of eye blinks	.015
e stat 14	Count of regressive saccades	.325

Table 5-8 Result of doing single factor one-way ANOVA test on the intermediate Eye Tracker-Captured feature set. Features in green behave statistically significantly differently under different levels of engagement.

The Tukey-Kramer method is used afterwards and Figure 5-8 (a-h) summarizes the mean values of each feature in each class. If there is a significant difference between pairs of individual criteria for that attribute, the significant values and the pairs are indicated.

For all selected AOI-related features, { e AOI 15 , e AOI 18 , e AOI 23 , e AOI 24 , e AOI 26 , e AOI 27 , e AOI 30 }, shown in Figure 5-8 (b-h), our analysis shows that they behave significantly differently between medium and high engagement ($p < .05$). The mean values of these features under medium engagement is smaller than that under high engagement, implying that the probabilities of those transition patterns in medium engagement level is less than that in high engagement level.

Meanwhile, the results for e stat 13 “count of eye blinks” show that there is a significant difference in number of eye blinks between medium and high engagement

level ($p < .05$), where on average, users under medium engagement level have fewer eye blinks than those under high engagement levels.

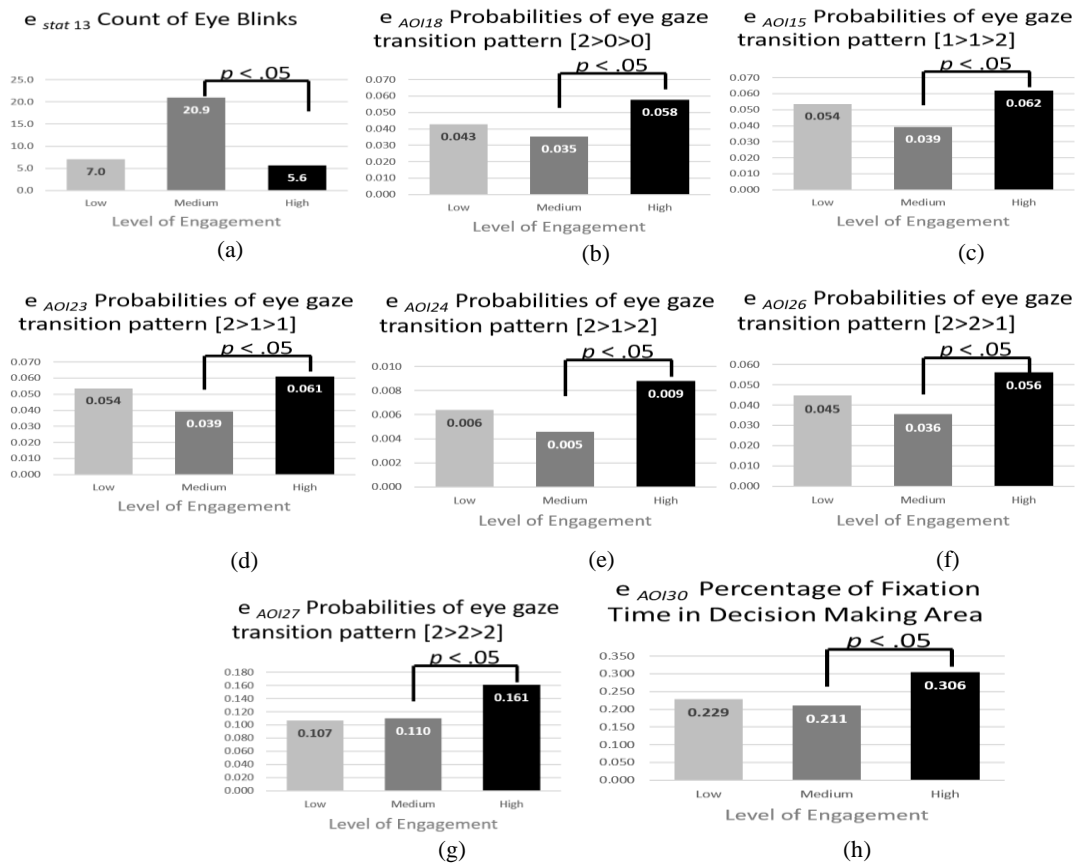


Figure 5-8 (a-h) Summary figures of comparing differences between pairs of individual criteria.

Table 5-9 and Table 5-10 are the final feature sets of Eye Tracker-Captured and Webcam-Based eye gaze features.

Feature	Descriptions
e AOI 15	Probabilities of eye gaze transition pattern [1>1>2]
e AOI 18	Probabilities of eye gaze transition pattern [2>0>0]
e AOI 23	Probabilities of eye gaze transition pattern [2>1>1]
e AOI 24	Probabilities of eye gaze transition pattern [2>1>2]
e AOI 26	Probabilities of eye gaze transition pattern [2>2>1]
e AOI 27	Probabilities of eye gaze transition pattern [2>2>2]
e AOI 30	Time fixated in decision making AOI
e stat 13	Count of eye blinks

Table 5-9 List of Eye Tracker-Captured Final Gaze Feature Set.

Feature	Descriptions
e webcam 3	Number of Fixations
e webcam 11	Number of Saccades
e webcam 12	Percentage of Saccades Times Exist in the Segment
e webcam 13	Ratio between Count Fixations and Count Pursuit
e webcam 14	Ratio between Count Fixations and Count Saccades
e webcam 15	Ratio between Count Pursuit and Count Saccades

Table 5-10 List of Webcam-Based Final Gaze Feature Set.

5.5.2 Results of Eye Tracker-Captured Eye Gaze Features

In order to detect the level of engagement, we used a 3-class Support Vector Machine (SVM) to train models, with RBF (radial basis function) kernel using C (C=1.0) as the default parameter. All prediction models are constructed using the SMO [63] method implemented in the Weka data mining tool [51].

In this subsection, we evaluate models built with the Eye Tracker-Captured Gaze Features. As summarized in Table 5-11, we focus on comparing the performance of models with different combinations of different datasets (Language Learning task vs Web Searching task) and feature sets (*FeatureSet* *without ANOVA* vs *FeatureSet* *with ANOVA*). Here, *FeatureSet* *without ANOVA* implies the intermediate feature set, in Table 5-6, extracted after the first two steps of feature selection. *FeatureSet* *with ANOVA*, which is the final Eye Tracker-Captured eye gaze feature set, listed in Table 5-9, is obtained after using single factor one-way ANOVA to analyse and select features.

	Dataset	Feature Set	\overline{CCR}	Baseline	ΔCCR	F1
M_{Tobii1}	Language Learning Task	FeatureSet <i>without ANOVA</i>	66.4%	58.8%	7.6%	.640
M_{Tobii2}		FeatureSet <i>with ANOVA</i>	67.2%		8.4%	.650
M_{Tobii3}	Web Searching Task	FeatureSet <i>without ANOVA</i>	64.9%	55.8%	9.1%	.603
M_{Tobii4}		FeatureSet <i>with ANOVA</i>	64.9%		9.1%	.599

Table 5-11 Summary table of the results of the models that are using different of Eye Tracker-Captured feature sets and dataset.

Firstly, we test on the Language Learning Dataset with 119 instances. 10 times 10-fold cross validation was used to evaluate the models. The average performance for the 10 experiments is reported.

We use the initial facial feature set to train the SVM model and use the results of 10 times 10-fold cross validation of the model as our baseline (correctly classified rate: CCR of 58.8%).

We first use *FeatureSet without ANOVA*, which contains 8 Eye Tracker-Captured Gaze features, to build a model (M_{Tobii1}) and test on Language Learning tasks data. M_{Tobii1} reaches 66.4% CCR, 7.6% higher than the baseline. We then built a model (M_{Tobii2}) with *FeatureSet with ANOVA*, i.e. using the intermediate Eye Tracker-Captured feature set. M_{Tobii2} reaches 67.2%, which is around 8.4% above baseline.

Same as Language Learning Task, the initial facial feature set was used to train the SVM model and test on Web Searching data. 10 times 10-fold cross validation of the model was carried out. This gives a baseline CCR of 55.8%.

The third model (M_{Tobii3}) using the *FeatureSet without ANOVA* was built and tested with the Web Searching dataset. The model is able to detect 64.9% of instance correctly, which is 9.1% higher than the baseline. The fourth model (M_{Tobii4}) using the *FeatureSet with ANOVA* performs similarly with M_{Tobii3} except that the F1 measure results is 0.4% lower.

The results show that the Tobii feature sets selected from Language Learning tasks can be generalized for engagement detection, as it can still reach 9.1% improvement over the baseline when we applied to train and test on an unseen dataset.

Unlike facial features, however, the performance of Eye Tracker-Captured models using *FeatureSet with ANOVA* is not significantly different from the models using *FeatureSet without ANOVA*. *FeatureSet with ANOVA* does not have the feature “Count of regressive saccades” that appears in the earlier feature sets. The results show that even though using signal factor one way ANOVA to select features in Eye Tracker-Captured feature sets does not bring significant improvement in model performance, it is still able to remove some non-useful (and also non-harmful) features.

5.5.3 Results of Webcam-based Eye Gaze Features

	Dataset	Feature Set	\overline{CCR}	Baseline	ΔCCR	F1
$M_{\text{Webcam_Gaze1}}$	Language Learning Task	FeatureSet _{without ANOVA}	60.5%	58.8%	1.7%	.573
$M_{\text{Webcam_Gaze2}}$		FeatureSet _{with ANOVA}	63.0%		4.2%	.589
$M_{\text{Webcam_Gaze3}}$	Web Searching Task	FeatureSet _{without ANOVA}	66.2%	55.8%	10.4%	.603
$M_{\text{Webcam_Gaze4}}$		FeatureSet _{with ANOVA}	66.2%		10.4%	.596

Table 5-12 Summary table of the results of the models that are using different of Webcam-based feature sets and dataset.

Since the Tobii eye tracker is not commonly found in normal consumer settings, it makes sense to investigate the possibility of using only the off-the-shelf webcam to extract eye gaze features for engagement prediction, for generalizability.

From the results summarized in Table 5-12 , we see that the webcam-based feature set is general enough to apply in the Web Searching dataset ($M_{\text{Webcam_Gaze3}}$ and $M_{\text{Webcam_Gaze4}}$), which gives 10.4% of improvement over baseline. Meanwhile, the difference between using different webcam-based feature sets (**FeatureSet**_{without ANOVA} and **FeatureSet**_{with ANOVA}) do not bring any significant difference on the Web Searching dataset, but achieve a little improvement (around 2.5%) for the Language Learning task dataset. We carry out a 10 times 10-fold cross validation and run a 1 tailed Paired T-Test with 10 degrees of freedom, as suggested in previous work [16] for model comparison, on the models $M_{\text{Webcam_Gaze1}}$ and $M_{\text{Webcam_Gaze2}}$. The result $t(10) = 3.2324$, $p = 0.0045 < 0.05$ shows that there are statistically significant differences between the results of $M_{\text{Webcam_Gaze1}}$ and $M_{\text{Webcam_Gaze2}}$. That shows that using the webcam-based feature sets (**FeatureSet**_{without ANOVA} and **FeatureSet**_{with ANOVA}) with single factor one way ANOVA for feature selection helps to improve the performance of the models.

5.5.4 Results on Using All Selected Eye Gaze Features

	Dataset	Feature Set	\overline{CCR}	Baseline	ΔCCR	F1
M_{Gaze1}	Language Learning Task	FeatureSet _{without ANOVA}	72.3%	58.8%	13.5%	.674
M_{Gaze2}		FeatureSet _{with ANOVA}	71.4%		12.6%	.679
M_{Gaze3}	Web Searching Task	FeatureSet _{without ANOVA}	67.5%	55.8%	11.7%	.620
M_{Gaze4}		FeatureSet _{with ANOVA}	70.1%		14.3%	.625

Table 5-13 Summary table of the results of the models that are using different of gaze feature sets and dataset.

Table 5-13 shows the results of combining both Eye Tracker-Captured and webcam-based eye gaze features. We observe that this combination can achieve more than 10% improvement over the baseline. The change from using single factor one way ANOVA for feature selection is not significant. Running the 1-tailed Paired T-Test with 10 degrees of freedom on the 10 times 10-fold cross validation results of models M_{Gaze1} and M_{Gaze2} gives the result $t(10) = 0.3318$, $p = 0.3734 > 0.05$, which shows there is no significant difference between two models. This shows that using the single factor one-way ANOVA for feature selection on Tobii and webcam based eye gaze features does not harm the models and is able to remove some useless features.

5.6 Summary

In this chapter, we propose using webcam-based eye gaze features and Eye Tracker-Captured gaze features to recognize human's engagement level in daily tasks – Language Learning and Web Searching. Our method identifies the eye movement behaviors from the eye gaze locations captured by the eye tracker and the videos captured by the webcam. We then extract and select useful features for building user-independent models. The models are evaluated with two different datasets and results show that (1) both webcam-based eye gaze features and Eye Tracker-Captured eye gaze features are useful for detecting users' engagement level, (2) the feature selection method helps to remove useless features from the initial feature set without hurting the performance of the models, and (3) the selected features are general enough to apply in another very different task and still perform reasonably well.

Chapter 6 Understanding Users from Mouse Movements

Traditionally, human-computer interaction involves three devices: the keyboard and mouse for input, and the screen for output. These are collectively known as the KVM interaction devices. Users use the mouse to achieve a lot of tasks, such as implementing inputs, selecting text, clicking buttons to trigger events, etc. These actions generate a good volume of the total amount of interaction from human to computer. Previous studies have suggested that mouse activities dominate interaction in daily computer use [20].

As mouse interaction is one of the most common interaction between human and computer, it makes sense that the user's mouse movement behavior, such as for aiding reading, selecting texts, and clicking buttons etc., may help to indicate his/her engagement level on task. It is also highly likely that there is a strong temporal pattern to mouse movements, and it may be possible to anticipate the user's next mouse activity from past behavior.

We invited subjects to carry out experiments in Language Learning tasks and Web Searching tasks. Feature selection was applied to extract useful features. Again, we only use Language Learning data for feature selection.

We construct user-independent models to identify the level of engagement and we test the models in a similar manner as the previous facial and gaze models. Specifically, we do 10 times 10-fold cross validation on Language Learning data and Web Searching data.

We find that our selected mouse features are able to achieve a performance improvement of 5.9% above baseline for testing and training on Language Learning tasks, and 5.2% above baseline for testing and training on Web Searching tasks.

The rest of this chapter is organized as follow. In Section 6.1, methods of extracting mouse movement features are discussed. Then we investigate the possibilities of using mouse interaction signals for user intention prediction in Section 6.2. Results of the user-independent models and feature selections are disused in Section 6.3. Finally, we conclude this chapter briefly in Section 6.4.

6.1 Features Extraction from Mouse

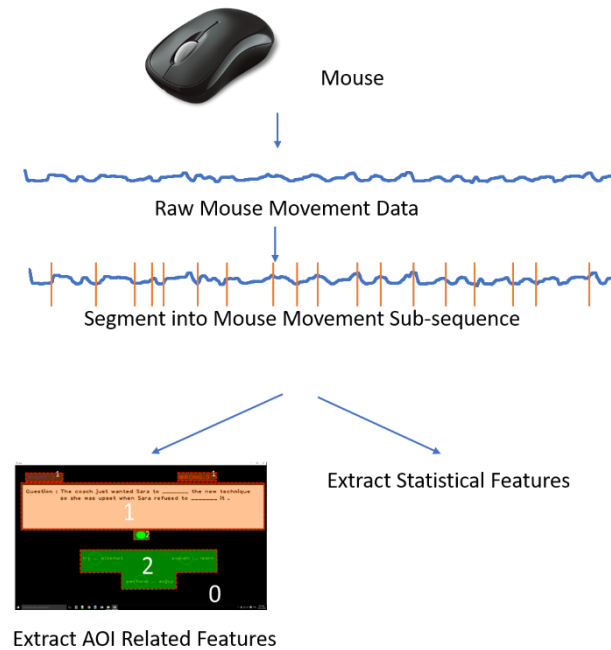


Figure 6-1 Flow of Mouse Feature Extraction.

In daily computer usage, a user moves the mouse cursor for different reasons. Even in reading tasks, which do not seem to require a lot of user input, the mouse is commonly used to highlight text, select relevant link, aid reading etc. We hypothesize that users' engagement level may be hidden behind mouse behaviors as the mouse serves as one of the main input methods. Mouse interaction signals contain a lot of information and may be represented with various methods. We use the basic statistical information of mouse movement and the mouse transitions between different areas of interest (AOI).

6.1.1 Definition of Mouse Movements in a Segment

There are many ways to define a mouse movement. In this work, we segment mouse movements according to the moments when the mouse movement speed drops to zero for more than 0.1 secs. Mouse movement sequences can then be represented as $MM = \langle MM_1, \dots, MM_k \rangle$ of k mouse movements, where $MM_i = [mc_1, \dots, mc_n]$ and $mc_j = [mc_{xj}, mc_{yj}]$. mc_{xj} is the mouse cursor x coordinate and mc_{yj} is the mouse cursor y coordinate of the j^{th} item in the sub-sequence of MM . MM_i is the component of the i^{th} item in mouse movement sequence MM and it consists of mouse

cursor sequences *mc*.

Each mouse movement sub-sequence will then be used to generate the following basic attribute of mouse movements.

- **Shortest-path distance:** the distance between the start and end points of the mouse movement;
- **Travel distance:** the total traversed distance of the mouse;
- **Mouse movement angle:** the angle between the shortest path defined above and the horizontal x-axis;
- **Sum Angle of Curvature:** the sum total of all angle changes in each movement;
- **Ratio of Shortest-path Distance to Travel Distance**
- **Movement speed**
- **Acceleration**

6.1.2 Statistical Mouse Features

For each sub-sequence in mouse movement sequence, we computed the 7 attributes mentioned in the previous paragraph. In each mouse movement sequence, we calculate the descriptive statistics including kurtosis, skewness, mean, median and standard deviation, as shown in Table 6-1.

Feature	Meaning	Formulation
m stat 1-5	Travel Distance	(1) Kurtosis
m stat 6-10	Movement Speed	(2) Skewness
m stat 11-15	Shortest-path Distance	(3) Mean
m stat 16-20	Acceleration	(4) Median
m stat 21-25	Movement Angle	(5) Standard deviation
m stat 26-30	Sum Angle of curvature	of the mouse movements in the sequence
m stat 31-35	Ratio of Shortest-path Distance to Travel Distance	

Table 6-1 Initial Mouse Cursor based Statistical Feature Set.

In addition to features extracted from basic mouse movement, we include features that capture information related to time and control stability. Fractal analysis, which uses power spectral density to conduct monofractal analysis, can estimate the degree of long-range correlations across a period of time. A scaling exponent α is calculated, an

α that is close to 1, pink noise, indicate that the data series contains substantial long-range correlations. Kloos et. al. [65] pointed out that the variation of pink noise in behavior reflect an optimal combination of stability and flexibility in control.

Thus, after we have extracted the 7 attributes, we try to calculate the scaling exponent for each attribute with 10-second moving windows and 50 % overlap. These sequences represent the change of the stability and flexibility in control over time, and going one more step further, it represents the regularity of the behavior over time. This generates 7 sequences of scaling exponents and the descriptive statistics are then calculated, as shown in Table 6-2.

Feature	Meaning	Formulation
m alpha 1-5	Regularity of Travel Distance	(1) Kurtosis
m alpha 6-10	Regularity of Movement Speed	(2) Skewness
m alpha 11-15	Regularity of Shortest-path Distance	(3) Mean
m alpha 16-20	Regularity of Acceleration	(4) Median
m alpha 21-25	Regularity of Movement Angle	(5) Standard deviation
m alpha 26-30	Regularity of Sum Angle of curvature	of the Scaling exponent sequence of each
m alpha 31-35	Regularity of Ratio of Shortest-path Distance to Travel Distance	instance

Table 6-2 Part of the Initial Mouse Feature set describing the regularity of the 7 attributes.

Combining features extracted from the descriptive statistics of basic mouse movement information and the descriptive statistics of the scaling exponent sequences, gives us 70 mouse movement features.

6.1.3 Mouse and AOIs

In addition to basic mouse movement features, the transitions of the mouse pointer between different areas of interest (AOIs) of an interface could be useful for predicting user's intention and attention.

To identify the AOI automatically from the screen, we first dilate and erode the screen image with different scale such that we could identify the location of paragraphs, lines and words, as mentioned in Section 5.3.1. To distinguish the functionality of the different areas: for information receiving, decision making and other areas in the Language Learning tasks, we make use of the fact that the questions and answer areas

are separate, which is a common interface design in online homework systems. For example, if the cursor is pointing at a word located in the question area, we classify it as being located in the information receiving area. However, if the cursor is pointing at a location in the question area but it is far away from the words, we classify it as located in other areas.

To find the distribution of different combination of transition, we consider the last 3 AOIs that the mouse cursor has been located in. We count the frequencies of all combinations and then normalized the values by the total number of transition. All the values of the histogram are the potential features that could be used. In total, there are 27 potential features, in Table 6-3, extracted from facial response videos.

Feature	Meaning	Formulation
m AOI 1-27	Mouse transition pattern in AOIs	The probability that a given mouse cursor transition pattern exists

Table 6-3 Part of the Initial Mouse Feature Set that related to mouse transition pattern in AOIs.

6.2 Predict User Intention from Mouse Interaction

We hypothesize that cues for users' mental states, such as engagement level and intentions, may be hidden behind mouse behaviors as the mouse serves as one of the main input methods. We therefore first use the mouse features to predict human intention and further use part of the mouse feature set for engagement detection.

We investigate user intention prediction along two dimensions: *understanding the type of the interaction*, and *the time that it will occur*.

Understanding the type of the interaction refers to determining the intended nature of an interaction activity just when it is about to occur. Guo and Agichtein [49] use the mouse movements, scrolling and other content information from the screen for detecting the user's search goal and propose a practical application of predicting ad clicks for a given search session. They pointed out that since some users use a mouse as a reading aid but others may not, one possible solution is to classify users into 2 different groups according to their mouse usage pattern and another solution is to use the eye gaze features such that the computer could know where the user is gazing at.

Apart from finding out "what happened?", some researches focus on "what will

happen?" This research problem is related to user intention prediction, such as giving the computer the ability to know what a user is going to do. Understanding the time would involve predicting not only the next event, but also the time interval within which it would occur. Laufer and Nemeth [71] use physiological data to detect and predict user action 2 seconds before it was carried out with the trained artificial neural networks model.

According to these previous works, user's physiological signal, gaze movement as well as the interaction patterns can indicate user intention. However, to our best knowledge, only a few works have been done to predict user intention by jointly modeling multi-channels of signals of user.

We investigate user intention prediction in two common web-based tasks: crowdsourcing annotation and Web Searching task. We use mouse interaction features to predict the next type of interaction that the users intended to carry out and the time interval within which it would occur.

6.2.1 Dataset

Crowdsourcing Annotation Task

Crowdsourcing has recently gained ground as a method of gathering training data for supervised machine learning methods. Even though it is easier and faster to collect data via crowdsourcing, in return label quality may be lower than employing experts, especially for complicated tasks. Therefore, different methods have been employed to assure label quality. One of them is to gather more than one answer from different annotators [104]. Previous work [41] proposed to mutually reinforce the aggregated crowd label by assessing the worker confidence level, worker's history and how well the worker agrees with the aggregated label.

Most studies on crowd-sourcing have the ground truth annotation labels and therefore they can compare the annotated results with the ground truth and calculate the quality of the answer. However, given the lack of ground-truth, the crowd-sourcing system is only useful if the average labeler is more or less trustworthy. In cases where the majority of the annotators produce the wrong label, the label generated by majority vote should not be trusted. The question then is how to tell whether the label should be trusted.

Liu and Liu [75] pointed out the same limitation of crowd sourcing task and they developed and analyzed an online learning algorithm that can differentiate high and low quality labelers over time and select the best set for labeling tasks. However, their approach starts to eliminate the worst labeler after a certain number of steps, for example they mentioned that it happens around step 90 in the real case. This implies that the problem still exists for the first 90 annotations. Therefore, given the lack of ground truth, we propose to use facial expression, gaze movement signals and interaction signals for detecting user's engagement level, which could be an indicator of the quality of the answer.

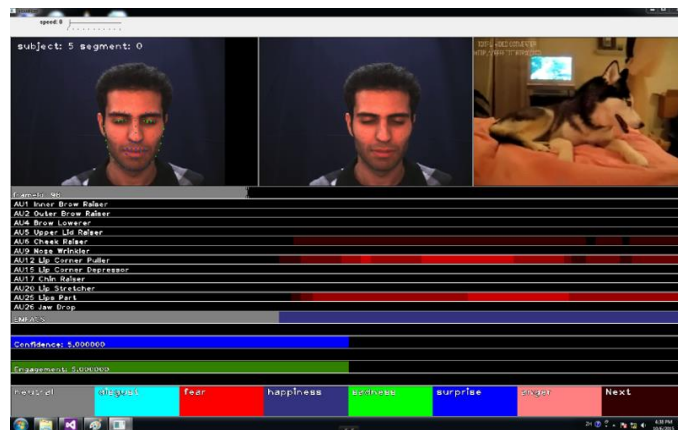


Figure 6-2 User Interface of the Crowdsourcing Annotation Experiment.

Figure 6-2 shows our crowdsourcing experiment interface. We ran an experiment which works on a crowdsourcing annotation task, which requires subjects to watch the videos of a person who was watching a movie clip from DISFA dataset [77]. The task given to the subjects is to annotate the emotion(s) that he/she feels that the viewer is exhibiting based on the facial expressions. The movie clip that is being viewed is also shown for reference. Each of the subjects need to annotate 243 video clips and report their confidence level (0-10) and engagement level (0-10), to reinforce the quality of the returned results. On average, each subject will take 2 hours to finish the task.

Web Searching Task

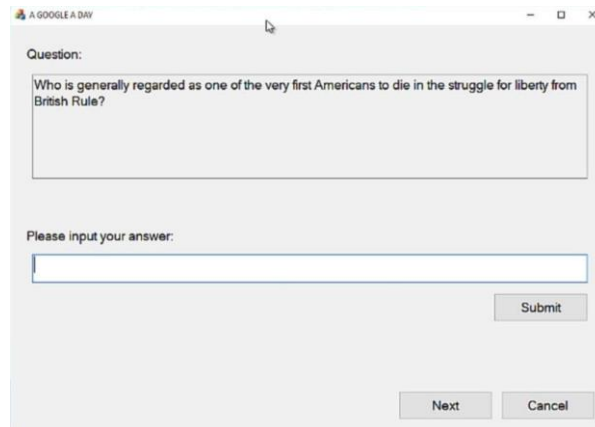
A screenshot of a web browser window titled "A GOOGLE A DAY". The window contains a "Question:" section with a text box containing the question: "Who is generally regarded as one of the very first Americans to die in the struggle for liberty from British Rule?". Below the question is a "Please input your answer:" section with a text input field. To the right of the input field is a "Submit" button. At the bottom of the window are "Next" and "Cancel" buttons.

Figure 6-3 User Interface of Web Searching Tasks.

Web search is another important daily application that often uses mouse and keyboard for interaction. For the Web Searching task, a user interface (Figure 6-3) was designed for the purposes of repeatability and controllability in the questions that will be presented. This interface is used only as an interface to display the search question and to input the answer. This user interface is the same as the one used in the Web Searching task when collecting the engagement dataset. However, these two Web Searching task datasets are separate from each other.

Our experiments involved 15 subjects, aged 21 to 31, with 7 females and 8 males, all comfortable with computer usage. They are all familiar with Chrome and Google search. For the crowdsourcing task, each subject was asked to annotate 243 video clips and report their confidence level (0-10) before they annotate the next video clip. On average, each subject took about 1.5 hours to finish the annotation task.

For the web search task, subjects are required to use the web browser and Google search engine. Each session of the web search task contained 6 questions, and on average each question took about 3.7 minutes to finish. Altogether there are 46 sessions.

In total, we collected about 1100 minutes of video and mouse logs for the crowdsourcing task, and around 1020 minutes for the web search task.

6.2.2 User Intention in the Tasks

When a user triggered an interaction, we wish to know what he/ she intended, or, we wish to be able to predict the type of the next upcoming activity. We also hope to know the approximate time window within the interaction occur.

User Intention in Crowdsourcing Task

Usually, a mouse click interaction activity occurs when a user is going to click or select on the screen, swap between different windows or trigger other events, etc. In a general crowdsourcing annotation task, users click to mark their annotation and confirm their answers. This allows users' intent to be modeled by their mouse click events.

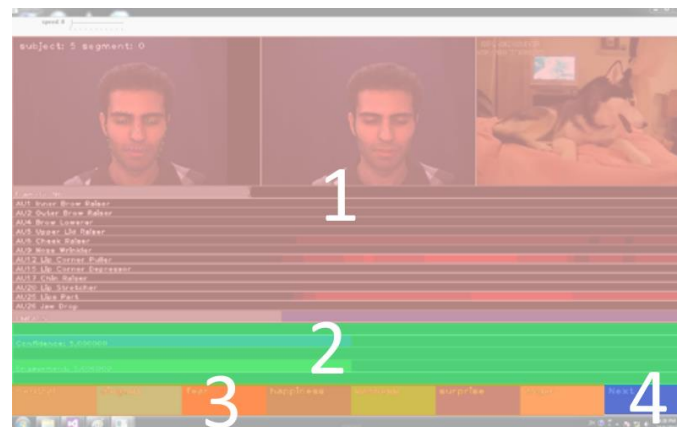


Figure 6-4 Regions of Crowdsourcing Annotation Task.

The user interface of the crowdsourcing annotation task contains 4 basic components, shown in Figure 6-4, which occupy 4 spatial regions: (1) the video and information panels, (2) self-report bar, (3) annotation buttons, and (4) Next button. Hence, we have 4 basic activities and each of them corresponds to a user intention. In addition to the 4 basic activities, we noticed that users sometimes click outside of the annotation window which gives us 1 more type of activity. The 5 types of activities are as follows:

- **Activity type 0:** Click outside the interface window
- **Activity type 1:** Click in region 1, the video and information panel, that will trigger jumps forward or backward in video.

- **Activity type 2:** Click in region 2, the self-report bar, that registers users' confidence level on annotating the video.
- **Activity type 3:** Click in region3, the annotation buttons, that allow user to annotate the video with one or more of 7 emotions that the user feels the viewer is exhibiting. Click on the button for select/deselect the emotion.
- **Activity type 4:** Click in region4, the Next button, for confirming the annotation and progress to the next screen.

User Intention in Web Searching Task

In the Web Searching task, a question in natural language is presented to the user, and he/ she has to find the answer via searching on the web. The question is chosen such that the answer could not be found via a simple web search. Since there is no control on what web page might be returned from a web search, we cannot classify user intention based on the mouse click location. Instead, the currently-viewed page is used to classify the user intention. The 5 types of activities for Web Searching tasks are listed as follows:

- **Activity type 0:** Question page – that the user is either reading the question or extracting keywords from the page
- **Activity type 1:** New tab in browser – that the user is starting a new search
- **Activity type 2:** Search result page – that contains multiple pieces of retrieved information and user is deciding among the information
- **Activity type 3:** Other webpage – user is finding the potential answer or useful information.
- **Activity type 4:** Clicking Submit button – user has come up with an potential answer

6.2.3 Models and Results

Our first approach for user’s intention prediction is to make use of user’ current and past activities, i.e. their historical activity sequence. We refer to this as the “probability model”. The second approach for user’s intention prediction is using “classification model” with mouse movement pattern.

Probability Model

The probability model relies on a classical n-gram model over the user activity sequence. We consider the most recent k activities and the conditional probability given the previous k activities could be calculated. The probabilities are computed from the training dataset.

Meanwhile, we also consider the time spent on each activity. We use the time duration spend on the activities to build a second prediction model. We quantize the continuous duration sequence into 3 discrete levels {long, medium, short} so that the duration level sequence could be generated. The second prediction model uses the previous k activities and time durations to calculate the probabilities.

We first evaluate how the value of k impacts the performance of the two conditional probability models. For crowdsourcing annotation task, the best performance of the two probability models is achieved when $k = 3$ (CCR: around 68%, Baseline: around 47%) which reveals that the past 3 activities are helpful for predicting the next activity. When k increased to 4 and 5, the performance drops rapidly.

At the same time, the best performance for probability models in Web Searching is achieved when $k=4$ (activity probability model with CCR: 54%, Baseline: 32%) and $k=2$ (activity and duration probability model with CCR: 57%, Baseline: 32%).

Generally, results show that we can detect the user’s next activity type from the historical activities and duration of the interaction sequence. It also implies that it is not informative to go too far back: user activities from more than 4 events back in time are unlikely to be useful.

Classification Model

For each instance, we extract the mouse movement sequence *MM*, which contains lots of sub-sequences of mouse movement, constructed by mouse *xy* location. We then extract the descriptive statistics, i.e. the mean, maximum, minimum, median and standard deviation, for the attribute values, as mentioned in Section 6.1.2, with the exception of the Ratio of Shortest-path Distance to Travel Distance. We further compute the features related to the area of interest, including the duration that the mouse has stayed in a particular AOI and the transition count between all pairs of AOIs.

For leave-one-subject-out setting, the best performance, achieved by the multimodal model that integrates user activity information and mouse interaction data, reaches around 69% CCR (Baseline: 47%) for crowdsourcing annotation, and around 70% CCR (Baseline: 32%) for web search.

Apart from predicting what is going to happen, we also want to know when it is going to happen. This is essential if we want to develop an intelligent system or agent that can prepare in advance to assist users, for example, by zooming in on an area when it knows that the user is about to click it. We therefore evaluate our intention prediction model in predicting a user's next activity *x* seconds, varying from 1 to 5 seconds, ahead before it actually occurs.

As expected, the performance drops as the window size is enlarged (e.g. as we try to predict further into the future). Our results suggest that prediction of time is difficult; However, the results are still reasonable for web search task, we can beat the baseline by around 30% (CCR: 62.1%, Baseline: 32%) when we try to predict 1 second ahead and around 29% (CCR:61.5%, Baseline: 32%) when we try to predict 2 seconds ahead. For the crowdsourcing task, the model beat the baseline by around 7% (CCR: 54%, Baseline: 47%) when predicting 1 second ahead and around 7.5% (CCR:54.6%, Baseline: 47%) when predicting 2 seconds ahead.

6.2.4 Conclusion on User Intention Prediction Work

We proposed two prediction models for user intention prediction. One model considered only the historical activity sequence, while the other applied mouse interaction signals and features extracted from mouse trajectory and clicking events. The results show that information cues of users' mental states, such as intentions, is

hidden behind mouse behaviors and could be predicted or detected via mouse movement statistical and AOI related features. We therefore also extracted the descriptive statistic of mouse movements (**m stat 1-35**) and AOI related features (**m AOI 1-27**) for engagement detection. Unlike user intention, user engagement level also considers the affective and focus attention states. Thus, we also consider the regularity of the mouse movements (**m alpha 1-35**) as features.

6.3 Mouse Movements for Engagement Detection

6.3.1 Feature Selection

In total, the initial feature set for mouse interaction have 97 features but not all of them are useful. Therefore, feature selection is used to select the relevant features. Same as the selection process for facial response and eye gaze features, the 3-step features selection process was applied in the mouse interaction features selection.

Features Left After Features Selection Step 1 & 2

Correlation attribute evaluation (CorrelationAttributeEval) with ranker search method provided by Weka [51] is used for feature selection. We consider the Pearson's correlation between an attribute and the class label and filter out all features with less than 0.3 correlation.

Step 2 builds a base model with the features selected in Step 1 and adds other features one by one to check if it could bring significant increase in performance. The significance test was done by running a 10 times 10-fold cross validation with 10 degrees of freedom Paired T-Test. After step 1 and step 2, the intermediate feature set is formed and listed in Table 6-4.

Feature	Descriptions
m stat 8	Mean of mouse movements' movement speed
m stat 18	Mean of mouse movements' acceleration
m stat 20	Standard deviation of mouse movements' acceleration
m stat 32	Skewness of ratio of shortest-path distance to travel distance
m stat 33	Mean of ratio of shortest-path distance to travel distance
m stat 35	Standard deviation of ratio of shortest-path distance to travel distance
m AOI 14	Probabilities of mouse transition pattern [1>1>1]
m AOI 19	Probabilities of mouse transition pattern [2>0>1]
m alpha 1	Regularities of travel distance (Kurtosis)
m alpha 6	Regularities of movement speed (Kurtosis)
m alpha 21	Regularities of movement angle (Kurtosis)
m alpha 31	Regularities of ratio of shortest-path distance to travel distance (Kurtosis)

Table 6-4 Intermediate Mouse Feature Set.

Final Features Set selected after Step 3

A single factor one-way ANOVA was performed to determine whether the mouse interaction features perform differently under different level of engagement. Before we run the test, we also checked the required assumptions and found that not all our mouse data fits the assumption of homogeneity of variance. Thus, Table 6-5 shows the significant values of the Levene’s Test of Homogeneity of Variance. If the significant value p is less than the alpha level .05, the null hypothesis of no variance difference is rejected and thus, $p < .05$ indicates that the assumption of homogeneity of variance is not met.

Feature	Descriptions	p value of Levene’s Test
m stat 10	Standard deviation of mouse movements’ movement speed	.000
m stat 18	Mean of mouse movements’ acceleration	.004
m stat 20	Standard deviation of mouse movements’ acceleration	.000
m stat 32	Skewness of ratio of shortest-path distance to travel distance	.777
m stat 33	Mean of ratio of shortest-path distance to travel distance	.740
m stat 35	Standard deviation of ratio of shortest-path distance to travel distance	.378
m AOI 14	Probabilities of mouse transition pattern [1>1>1]	.006
m AOI 19	Probabilities of mouse transition pattern [2>0>1]	.206
m alpha 1	Regularities of travel distance (Kurtosis)	.810
m alpha 6	Regularities of movement speed (Kurtosis)	.531
m alpha 21	Regularities of movement angle (Kurtosis)	.530
m alpha 31	Regularities of ratio of shortest-path distance to travel distance (Kurtosis)	.688

Table 6-5 Significant values of the Levene’s Test of Homogeneity of Variance.

As the results indicate, features m stat 8, m stat 18, m stat 20 and m AOI 14 do not meet the assumption of homogeneity of variance. We therefore use the Kruskal-Wallis test for determining if there are statistically significant differences between the groups and use Games-Howell for the post-hoc test for those features.

Feature	Descriptions	<i>p</i> value of one-way ANOVA
m stat 32	Skewness of ratio of shortest-path distance to travel distance	.234
m stat 33	Mean of ratio of shortest-path distance to travel distance	.231
m stat 35	Standard deviation of ratio of shortest-path distance to travel distance	.191
m AOI 19	Probabilities of mouse transition pattern [2>0>1]	.742
m alpha 1	Regularities of travel distance (Kurtosis)	.003
m alpha 6	Regularities of movement speed (Kurtosis)	.219
m alpha 21	Regularities of movement angle (Kurtosis)	.169
m alpha 31	Regularities of ratio of shortest-path distance to travel distance (Kurtosis)	.155

Table 6-6 Result of doing single factor one-way ANOVA test on the intermediate mouse feature set. Features in green color are having statistically significant difference under different level of engagement.

For m alpha 1 “Regularities of travel distance (Kurtosis)”, there was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,116) = 1.539, p = .003 < .05$). A Tukey post hoc test revealed that the kurtosis of regularities of travel distance was statistically significantly higher in low engagement level ($-0.53 \pm .29$) compared to the medium engagement level ($-0.81 \pm .20, p = .004$) and high engagement level ($-0.83 \pm .20, p = .002$). There was no statistically significant difference between the medium and high groups ($p = .757$).

Feature	Features Left after Step 2	<i>p</i> value of Kruskal Wallis Test
m stat 10	Standard deviation of mouse movements’ movement speed	.253
m stat 18	Mean of mouse movements’ acceleration	.343
m stat 20	Standard deviation of mouse movements’ acceleration	.244
m AOI 14	Probabilities of mouse transition pattern [1>1>1]	.039

Table 6-7 Result of doing Kruskal Wallis test on the intermediate mouse feature set. Features in green color are having statistically significant difference under different level of engagement.

For mAOI14 “Probabilities of mouse transition pattern [1>1>1]”, there was a statistically significant difference between groups as determined by Kruskal-Wallis H test ($\chi^2(2) = 6.514, p = 0.039$). A Games-Howell test revealed that the probabilities of mouse transition pattern staying in the area of information receiving was statistically

significantly lower in low engagement level ($.0003 \pm .0008$, $p = .000$) compared to the medium engagement level ($.0085 \pm .0161$). There was no statistically significant difference between the low and high groups ($p=.148$) and medium and high groups ($p = .211$). Based on our observation, this pattern is mainly caused by the fact that some of the subjects in medium level of engagement tends to use the mouse to guide their reading, while the users with low engagement tends to sit back and read more passively.

Table 6-8 listed the features of the final feature sets of mouse interaction features.

Feature	Description
m _{alpha 1}	Regularities of travel distance (Kurtosis)
m _{AOI 14}	Probabilities of mouse transition pattern [1>1>1]

Table 6-8 Final Mouse Feature Set.

6.3.2 Results

	Dataset	Feature Set	\overline{CCR}	Baseline	ΔCCR	F1
M_{Mouse1}	Language Learning Task	FeatureSet _{without ANOVA}	66.4%	58.8%	7.6%	.612
M_{Mouse2}		FeatureSet _{with ANOVA}	64.7%		5.9%	.598
M_{Mouse3}	Web Searching Task	FeatureSet _{without ANOVA}	61.0%	55.8%	5.2%	.547
M_{Mouse4}		FeatureSet _{with ANOVA}	61.0%		5.2%	.577

Table 6-9 Summary table of the results of the models that are using different of Mouse feature sets and dataset.

In order to detect the level of engagement with mouse interaction data, we used a 3 class support vector machine (SVM) to train models, with RBF (radial basis function) kernel using C (C=1.0) as the default parameter. All prediction models are constructed using the SMO [63] method implemented in the Weka data mining tool [51].

In this subsection, we evaluate models built with the mouse interaction features. As shown in the summary Table 6-9, we focus on comparing the performance of models with different combinations of different datasets (Language Learning task vs Web Searching task) and feature sets (*FeatureSet_{without ANOVA}* vs *FeatureSet_{with ANOVA}*). Here, *FeatureSet_{without ANOVA}* uses the intermediate feature set, listed in Table 6-4, constructed

after the first two steps of feature selection. *FeatureSet* with ANOVA, which is the final feature set, described in Table 6-8, is constructed by using single factor one-way ANOVA and Kruskal-Wallis Test to analyse and select features.

From the results summary in Table 6-9, we see that the mouse feature set is general enough to apply in the Web Searching datasets M_{Mouse3} and M_{Mouse4} , achieving 5.2% of improvement over baseline. Meanwhile, using different mouse feature sets (*FeatureSet* without ANOVA and *FeatureSet* with ANOVA) does not create any significant difference in for the models of Web Searching dataset, but results in a small drop (around 2%) for the Language Learning dataset. Thus, we carry out a 10 times 10-fold cross validation and run a 1 tailed Paired T-Test with 10 degrees of freedom on the models M_{Mouse1} and M_{Mouse2} , and find that the results are not statistically significantly different. The result of using mouse feature set *FeatureSet* without ANOVA and *FeatureSet* with ANOVA with single factor one way ANOVA and Kruskal-Wallis for feature selection do not hurt the model performance and helps to remove the useless features.

6.4 Summary

In this chapter, we propose the method of using mouse cursor-based features to recognize human's engagement level during daily tasks – Language Learning task and Web Searching task. We identify the mouse movement behaviors from mouse cursor locations and also the screen videos captured by the system. We then extract and select the useful features for building user-independent models. The models are evaluated with two different datasets and results show that (1) mouse cursor based features are useful for detecting users' engagement level, (2) the feature selection method helps to remove useless features from the initial feature set without hurting performance of the models, and (3) the selected features are general enough to apply in another type of task and still perform reasonably.

Furthermore, we also investigate the use of mouse behaviors in user intention prediction. Specifically, we extract features from mouse movements signal and train the user-independent models to determine the intended nature of an interaction activity just when it about to occur and predicting the time interval within which it would occur. The results show that the model is able to predict the type of interaction activity and the time interval within which it would occur for an unseen user with reasonable accuracy.

Chapter 7 Extending Multi-Modality Engagement Detection into Real Life

Apart from detecting engagement levels with different modalities (facial, eye and mouse) separately, we would like to know how well the models could perform if we combine and use the modalities together. Considering different combinations of the modalities allows us to know how well the model would perform even if we cannot obtain signals from all modalities. This is especially relevant for real-life usage. For example, suppose we have a user who wants to detect his/her level of engagement but he/she does not have an Eye Tracker. In this case, without the eye gaze position data, how well can the model perform?

We therefore group the type of features in three different groups according to the means of data collection: (1) webcam-based features, (2) Eye Tracker-Captured features, and (3) mouse cursor-based features. Webcam-based features are the features extracted from the recorded videos of the off-the-shelf webcam, which is usually embedded into most consumer-grade monitors and laptops these days. Eye Tracker-Captured features are extracted from the commercial infrared equipment for eye tracking – in our case, we use the Tobii EyeX Controller for collecting the data. Mouse cursor based features are the features extracted from the mouse cursor movement.

7.1 Results of Multi Modalities

Feature	Descriptions
f 3	Standard Deviation of the location of the head - z
f 14	Count of the Existence of AU 5 (Upper Lid Raiser)
f 47	Presence of AU 28 (Lip Suck)
f 49	Percentage of Frame with Low Confidence in Face Detection
e webcam 3	Numbers of Fixation
e webcam 11	Numbers of Saccades
e webcam 12	Percentage of times of Saccades exist in the segment
e webcam 13	Ratio between Count Fixation and Count Pursuit
e webcam 14	Ratio between Count Fixation and Count Saccade
e webcam 15	Ratio between Count Pursuit and Count Saccade

Table 7-1 List of Final Feature Set Obtained from Webcam Signals.

Feature	Descriptions
e AOI 15	Probabilities of eye gaze transition pattern [1>1>2]
e AOI 18	Probabilities of eye gaze transition pattern [2>0>0]
e AOI 23	Probabilities of eye gaze transition pattern [2>1>1]
e AOI 24	Probabilities of eye gaze transition pattern [2>1>2]
e AOI 26	Probabilities of eye gaze transition pattern [2>2>1]
e AOI 27	Probabilities of eye gaze transition pattern [2>2>2]
e AOI 30	Time fixated in decision making AOI
e stat 13	Count of eye blinks

Table 7-2 List of Final Feature Set Obtained from Tobii Signals.

Feature	Description
m alpha 1	Kurtosis of regularities of travel distance
m AOI 14	Probabilities of mouse transition pattern [1>1>1]

Table 7-3 List of Final Feature Sets Obtained from Mouse Cursor Signals.

Table 7-1 is the final webcam-based feature set, which includes the facial features and the webcam-based eye gaze features. Table 7-2 is the final Eye Tracker-Captured feature set and Table 7-3 is the final of mouse cursor based feature set.

In a similar manner as in the previous sections, we train the SVMs to detect the 3 classes user engagement level in 2 tasks. The results of using different combinations of modalities are discussed in the following part.

Results on Training and Testing on Language Learning Tasks Data

Webcam-Based Facial and Eye Features	Eye Tracker-Captured Gaze Features	Mouse Cursor Based Features	\overline{CCR}	Baseline	ΔCCR
✓	X	X	68.9%	58.8%	10.1%
X	✓	X	67.2%		8.4%
X	X	✓	64.7%		5.9%
✓	✓	X	72.3%		13.5%
✓	X	✓	69.7%		10.9%
X	✓	✓	68.9%		10.1%
✓	✓	✓	73.1%		14.3%
✓	✓	✓	73.1%		14.3%

Table 7-4 Summary Table of Using Different Combination of Modalities for Engagement Detection in Language Learning Tasks Dataset.

As shown in the Table 7-4 , the highest performance, producing 14.3% of improvement over baseline, is attained when we use all features extracted from the 3 modalities. It only drops around 1% if we remove the mouse modality, which implies that even though the mouse feature itself could help to produce 5.9% of improvement over baseline, its contribution is redundant with the other 2 modalities.

The lowest performance comes from the model that uses mouse feature set alone which only bring 5.9% of improvement over the baseline. Meanwhile, results show that using webcam-based features alone outperforms using Eye Tracker-Captured feature set, which implies that specialized devices are not necessary for good performance.

Results on Training and Testing on Web Searching Tasks Data

Webcam-Based Facial and Eye Features	Eye Tracker-Captured Gaze Features	Mouse Features	\overline{CCR}	Baseline	ΔCCR
✓	X	X	64.9%	55.8%	9.1%
X	✓	X	64.9%		9.1%
X	X	✓	61.0%		5.2%
✓	✓	X	68.8%		13.0%
✓	X	✓	66.2%		10.4%
X	✓	✓	67.5%		11.7%
✓	✓	✓	74.0%		18.2%

Table 7-5 Summary Table of Using Different Combination of Modalities for Engagement Detection in Web-Searching Tasks Dataset.

Table 7-5 shows the summary of results of using different combinations of modalities in engagement detection on the Web Searching task dataset. It is surprising and encouraging that the performance of the model using all modalities could achieve around 18% improvement over the baseline, which is even higher than the model trained and tested in Language Learning task dataset – which is supposedly cleaner with a more tightly constrained flow and interface, which theoretically should make the prediction easier!

Meanwhile, similar to the results of testing on the Language Learning task dataset, the second-highest accuracy was achieved by the model using both webcam-based features and Eye Tracker-Captured features, achieving a CCR which is 5.2% less than the best model. Unsurprisingly, the poorest performance was obtained from the model that uses mouse features only, which produced 5.2% improvement over baseline. This differs from the Language Learning task data set, in which the mouse features bring improvement into the model when coupled with all three modalities.

7.2 Summary

In this chapter, we evaluated the performance of combining different modalities

(webcam, eye tracker and mouse cursor) for engagement detection. We adopt machine learning techniques to model the captured data and build user-independent models that are able to detect the engagement level of the users. We demonstrate that combining all three modalities could achieve the best performance for engagement detection, that reach 74% of accuracy (which is 18% over the baseline) for detecting engagement in daily tasks such as Web Searching.

Meanwhile, as eye tracker is not as commonly found as webcam, we also investigate the feasibility of measuring engagement without eye tracker. Our proposed models can still reach around 10-11% improvement over the baseline.

Chapter 8 Limitations and Future Work

This thesis investigates techniques for engagement level detection. The experimental results are promising, however, there are still some limitations and potential future work of the related studies.

8.1 Use of Deep-learning Techniques

Our studies use support vector machines as the machine learning method. Given that the deep learning techniques give sweeping performance gains in many recognition problems, it should be of interest to benchmark with these techniques. Specifically, the long short-term memory neural networks (LSTM NN) could be one of the good choice for engagement detection if we have enough data. However, at this point, we are limited by the amount of data, since it is not easy to collect large amounts of data with in HCI applications.

8.2 Process of Features Selection

Features, for example facial features, may be inter-related with each other. This raises the research question about whether the current approach for feature selection is able to handle these inherent correlations between features. Even though we treat the features as independent, our current method uses correlation coefficient as the first criteria of feature selection. This means that features will be selected if they are moderately correlated with labels, which implies that there is a chance of selecting features that are inter-related as they both contribute. As one of our research questions is to identify indicative behaviors, treating features as independent features could help us to identify and understand behaviors in an easier way. However, clustering the features into different groups and then perform group based feature selection may help to improve model performance.

8.3 Calibration of Individual Persons Data

For the purpose of user independence, the work in this thesis, uses raw measurements for most of the features, including features such as head movements, which may be fairly user-dependent. This use of absolute raw values without normalization may result in a decrease in performance. It is easy to see that calibration

or normalization could help minimize individual variations – for example, determining the maximum range of subject’s head movements from the available data and representing the head movement as a calibrated value from 0-1. This could help with model performance, however, more thought is needed to balance this with our requirement for user-independent features.

8.4 Recognition of Engagement Level in Three Dimensions

This thesis focuses on the single label classification problem for engagement studies. We foresee, however, that recognizing and understanding the three dimensions of engagement (affective, behavioral and cognitive) could be helpful. It is possible that even though the overall engagement level may be at the same value, different combinations of the three dimensions of engagement may lead to different kinds of user behaviors.

8.5 User Engagement in Mobile Contexts

Another limitation is that the current model focuses on KVM devices, but nowadays more and more people are choosing to use mobile or tablet. Therefore, it would be interesting to investigate going beyond the current modalities (facial, eye gaze and mouse movement), for example, incorporating tapping or swiping information.

In addition to detecting user engagement of daily computer task, we would like to expand our scope of investigation to mobile based interaction task. In particular, we wish to apply the engagement/affect detection to the photorefractive mobile application for vision screening, both user experience study and user engagement study. One of the possible modality is to use the front camera of the mobile phone to capture user facial expression and analyse how engaged the users are in the application learning process.

8.6 Extension on Video-Sharing Platform

Currently, the platform for videos sharing considers the facial expression for detecting emotion of the viewer. The platform could be extended by integrating users’ engagement detection models and users’ intention prediction model. For example, we could suggest different types of videos if the system finds that the viewer is dis-engaged with the current video, or change the Emotar to one that shares similar feeling with the viewer.

Chapter 9 Conclusion

User Experience as part of the “third wave” of Human-Computer Interaction, is raising the attention of researchers. The necessity of broadening the scope of User Experience (UX) is shown in studies of computer interaction, as opposed to just focusing on usability metrics such as the effectiveness of the application. Users spend time, emotion, attention and effort when they interact with technologies, and a successful application or task should be able to engage users, instead of simply being a “job” that needs to be completed. Thus, detecting user engagement in daily computer interaction is essential.

There has been much previous work that tries to detect user engagement via various means, like facial expression, eye gaze movement and mouse movement, however the work is limited by three main challenges: (1) the constraints caused by using intrusive devices, (2) limitations of specific tasks (like gaming) which may produce user behavior which is different from daily computer usage, (3) and incomprehensive ground truth as collected by straightforward and direct survey questionnaires that capture users’ self-reported numeric level of engagement, but which may not cover the three dimensions of engagement.

We focus on non-intrusive interaction cues, especially facial expression, eye gaze and mouse cursor signals, for recognizing users’ engagement level in daily computer interaction tasks. We conducted experiments and study users’ behaviors with the Language Learning tasks and Web Searching tasks and data are collected via webcam, commercial infrared eye tracker and mouse cursor. Our approach can effectively identify when the users are experiencing different engagement levels in daily tasks.

Overall, our approach is able to achieve a performance improvement of 14.3% above baseline for the dataset (Language Learning task) used for feature selection; and achieve a performance improvement of 18.2% above baseline when we use the same feature set for another task (Web Searching task).

Our contributions of this work can be summarized as:

- We collect a comprehensive dataset containing different levels of User Engagement, containing two separate modes of ground truth annotation;
- We investigate the detection of engagement level based on 2 common daily

human-computer interaction tasks;

- We identify good features that are effective in describing specific facial, eye and mouse movement behaviors when the users are in different levels of engagement;
- We apply machine learning techniques to build user-independent models to recognize the level of engagement in daily task with different modalities in a non-intrusive manner;
- We conduct experiments with human subjects to evaluate the accuracy of our approach in various context;

As an extended study of understanding human mouse behaviors, we investigate the use of mouse behaviors in user intention prediction. Specifically, we extract features from mouse movements signals and train the models for determining the intended nature of an interaction activity just when it about to occur and predicting the time interval within which it would occur.

Our contributions of this work can be summarized as:

- We investigate user intention prediction based on common daily web-based tasks: crowdsourcing annotation task and Web Searching task;
- We identify good features that are effective in describing mouse movements behaviors of these two tasks;
- We propose user-independent models to predict the user intention from two aspects: the type of the next interaction, and an approximate time when this interaction would be triggered.

9.1 Other Relevant Contributions

In addition to the main contributions previously described, the following describes some relevant contribution arising from my thesis project:

9.1.1 Using Interaction Data to build Gaze Model

Most of the eye gaze estimation systems rely on explicit calibration and is inconvenient to the users. As the eye gaze location and the interaction cues is likely to be strong correlated with each other, using the interaction cues and the corresponding eye gaze location generated from daily human computer interaction to train a supervised learning models is possible. The information cues could be found from keyboard and mouse interaction, such as mouse cursor and caret locations. Therefore, we develop a set of robust geometric eye gaze features and corresponding data validation mechanism for identifying good training data that are collected unobtrusively in real-use scenarios. The evaluation on the proposed model shows that it could achieve an average error of 4.06° .

M. X. Huang, **T. C. K. Kwok**, G. Ngai, H. V. Leong, and S. C. F. Chan, “Building a Self-Learning Eye Gaze Model from User Interaction Data,” *Proc. ACM Int. Conf. Multimed. - MM '14*, pp. 1017–1020, 2014.

9.1.2 PACE – Personalized, Auto-Calibrating Eye Tracker

PACE, a Personalized, Auto-Calibrating Eye Tracker, identifies and collect data unobtrusively from daily interaction events on standard computer for eye tracking. With the set of robust geometric gaze features extracted from webcam, a two-layer validation mechanism to identify the quality of the samples from daily interaction data. PACE was founded on a detailed investigation of the gaze and interaction cues relation with the consideration of user habits. It continuously recalibrates and improves when more data is collected. The in-situ study using real-life tasks on a set containing interaction behavior from various interactive applications shows that PACE achieves an average error of 2.56° , which is comparable to the state-of-the-art, without explicit training or calibration.

M. X. Huang, **T. C. K. Kwok**, G. Ngai, S. C. F. Chan, and H. V. Leong, “Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions,” in *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5169–5179. - **Best Paper Award**.

9.1.3 Photorefraction

Vision problems, such as refractive error (e.g. nearsightedness, astigmatism etc.) are common ocular problems, which, if uncorrected, may lead to serious visual impairment. The diagnosis of such defects conventionally requires expensive specialist equipment and trained personnel, which is a barrier in many parts of the developing world. We aim to democratize optometric care by utilizing the computational power inherent in consumer-grade devices and the advances made possible by multimedia computing.

The system employs the photorefractive approach with the graphical calibration method developed by Chan, Edwards and Brown [19]. If a patient has refractive error, the photographic reflex of the eye towards a flash of light, which is essentially the reflection of the light off the retina, manifests itself as a crescent in the photograph. The position of the crescent shows whether a patient has hyperopia or myopia. In myopic cases, the eye is focused in front of the light flash, producing a crescent that appears on the same side as the light source. The opposite is true for hyperopic cases, where the crescent appears on the opposite side of the eye [98].

We therefore present a vision-based, data-driven approach to identifying and measuring refractive errors in human subjects with low-cost, easily available equipment and no specialist training. Our system uses only a standard mobile device and the embedded camera, and is successful at detecting and measuring myopia with less than 1.0 diopter of error. Some manual correction is required, but it takes no special expertise or training (beyond being able to read instructions and operate a mobile device), and only 19% of data requires this additional manual processing.

Table 9-1 shows the concept of photorefraction. A flash source is positioned at a distance e above the camera lens. d is the distance from the lens to the eye being tested. Light enters the myopic eye and is focused in front of the retina. Image AB is formed on the retina and forms an aerial image B'A' at the far point plane of the eye. If the eye is sufficiently myopic, the light returning from this image will enter the camera lens and manifests as a crescent-shaped reflex z on the camera film/sensor. The photograph of the eye shows a crescent on the same side as the light source

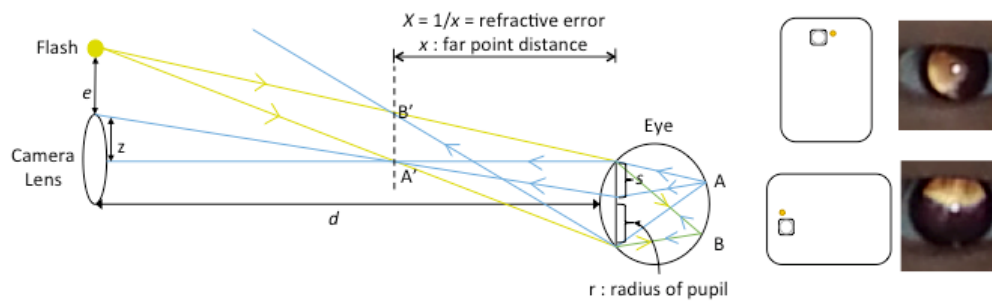


Table 9-1 Photorefraction. Top photo: vertical orientation of mobile device with flash to the left of the eye; bottom photo: horizontal orientation with flash to the top of the eye. (Ray diagram adapted from Chan, Edwards and Brown [19])

T. C. K. Kwok, C. N. Shum, G. Ngai, H. V. Leong, G. A. Tseng, H. Choi, K. Mak, and C.W. Do, “Democratizing Optometric Care: A Vision-Based, Data-Driven Approach to Automatic Refractive Error Measurement for Vision Screening,” in *2015 IEEE International Symposium on Multimedia (ISM)*, 2015, pp. 7–12

References

- [1] Abbasi, A.Z., Ting, D.H., and Hlavacs, H.Engagement in games: Developing an instrument to measure consumer videogame engagement and its validation. *International Journal of Computer Games Technology 2017*, (2017).
- [2] Adibuzzama, M., Jain, N., Steinhafel, N., et al.In situ affect detection in mobile devices: a multimodal approach for advertisement using social network. *Applied Computing Review 13*, 4 (2013), 66–77.
- [3] Aigrain, J., Dubuisson, S., Detyniecki, M., and Chetouani, M.Person-Specific Behavioural Features for Automatic Stress Detection. *International Workshop on Context Based Affect Recognition*, (2015), 1–6.
- [4] Alyuz, N., Okur, E., Oktay, E., et al.Semi-supervised model personalization for improved detection of learner’s emotional engagement. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, (2016), 100–107.
- [5] Amos, B., Ludwiczuk, B., and Satyanarayanan, M.*OpenFace: a General-Purpose Face Recognition Library with Mobile Applications*. 2016.
- [6] Andujar, M. and Gilbert, J.Let’s Learn!: Enhancing User's Engagement Levels through Passive Brain-Computer Interfaces. *CHI’13 Extended Abstracts on Human Factors in ...*, (2013), 703–708.
- [7] Aoki, S. and Uchida, O.A Method for automatically generating the emotional vectors for emoticons using weblog articles. *World Scientific and Engineering Academy and Society Press*, (2011).
- [8] Arapakis, I. and Leiva, L. a.Predicting User Engagement with Direct Displays Using Mouse Cursor Information. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2016), 599–608.
- [9] Arapakis, I. and Valkanas, G.Understanding Within-Content Engagement through Pattern Analysis of Mouse Gestures. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, (2014), 1439–1448.

- [10] Attfield, S., Kazai, G., and Lalmas, M. Towards a Science of User Engagement (Position Paper). *WSDM Workshop on User Modelling for Web Applications*, (2011).
- [11] Baldauf, D., Burgard, E., and Wittmann, M. Time Perception as a Workload Measure in Simulated Car Driving. *Applied Ergonomics* 40, 5 (2009), 929–935.
- [12] Bales, R.F. *Social interaction system: Theory and Measurement*. Transaction Publishers, 2001.
- [13] Bartlett, M.S., Hager, J.C., and Ekman, P. Measuring facial expressions by computer image analysis. *Psychophysiology* 16, (1999), 253–263.
- [14] Bosch, N., D’Mello, S.K., Baker, R.S., et al. Detecting Student Emotions in Computer-Enabled Classrooms. *IJCAI International Joint Conference on Artificial Intelligence 2016-Janua*, (2016), 4125–4129.
- [15] Bote-Lorenzo, M.L. and Gómez-Sánchez, E. Predicting the decrease of engagement indicators in a MOOC. *Seventh International Conference on Learning Analytics and Knowledge*, (2017), 143–147.
- [16] Bouckaert, R.R. Choosing between Two Learning Algorithms Based on Calibrated Tests. *Proceedings of the 20th International Conference on Machine Learning*, (2003), 51–58.
- [17] Bradley, M.M. and Lang, P.J. Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *J. B&w Thu. & Exp. Psvchrar.* 25, 1 (1994), 49–59.
- [18] Brave, S. and Nass, C. Emotion in human-computer interaction. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, (2002), 81–96.
- [19] Chan, O.Y.C., Edwards, M., and Brown, B. Calibration and Validity of an Eccentric Photorefractor. *Ophthal. Physiol. Opt.* 16, 3 (1996), 203–210.
- [20] Chang, C., Amick, B.C., Menendez, C.C., et al. Daily Computer Usage Correlated with Undergraduate Students’ Musculoskeletal Symptoms. *American Journal of Industrial Medicine* 50, 6 (2007), 481–488.

- [21] Chengyao, S. and Zhao, Q. Webpage Saliency. *Computer Vision-ECCV*, 2014, 33–46.
- [22] Chew, S.W., Lucey, P., Lucey, S., Saragih, J., Cohn, J.F., and Sridharan, S. Person-independent facial expression detection using Constrained Local Models. *Face and Gesture 2011*, (2011).
- [23] Chon, J.F. Foundations of Human Computing: Facial Expression and Emotion. *Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06)*, (2006).
- [24] Cohen, I., Sebe, N., Chen, L., Garg, A., and Thomas, S. Facial Expression Recognition from Video Sequences: Temporal and Static Modelling. *Comput Vis Image Understand*, (2003).
- [25] Collewijn, H., Erkelens, C.J., and Steinman, R.M. Binocular Co-ordination of Human Horizontal Saccadic Eye Movements. *The Journal of physiology* 404, (1988), 157–82.
- [26] Crouzet, S.M., Kirchner, H., and Thorpe, S.J. Fast Saccades Toward Faces: Face Detection in just 100 ms. *Journal of Vision* 10, 4 (2010), 1–17.
- [27] Cui, Y., Kangas, J., Holm, J., and Grassel, G. Front-Camera Video Recordings as Emotion Responses to Mobile Photos Shared Within Close-Knit Groups. *CHI'13*, (2013), 981–990.
- [28] D'Mello, S.K., Craig, S.D., Sullins, J., and Graesser, A.C. Predicting Affective States Expressed Through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. *International Journal of Artificial Intelligence in Education* 16, 1 (2006), 3–28.
- [29] Derks, D., Fischer, A.H., and Bos, A.E.R. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, (2007).
- [30] Dinges, D.F., Venkataraman, S., McGlinchey, E.L., and Metaxas, D.N. Monitoring of Facial Stress During Space Flight: Optical Computer Recognition Combining Discriminative and Generative Methods. *Acta Astronautica* 60, 4-7 SPEC. ISS. (2007), 341–350.
- [31] Do2Learn. Emotions Color Wheel.
<http://www.do2learn.com/organizationtools/EmotionsColorWheel/>.

- [32] Duprez, C., Christophe, V., Rime, B., Congard, A., and Antoine, P. Motives for the social sharing of an emotional experience. *Journal of Social and Personal Relationships*, (2014), 1–31.
- [33] Ekman, P. and Friesen, W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.
- [34] Essa, I.A., Darrell, T., and Pentland, A. Tracking Facial Motion. *Proceedings of the IEEE Workshop on Nonrigid and Articulate Motion*, (1994).
- [35] F, L., Y, S., T, O., and Y., S. Gaze Estimation From Eye Appearance: A Head Pose-Free Method via Eye Image Synthesis. *IEEE Trans Image Process.* 24, 11 (2015).
- [36] Feng, L., Takahiro, O., Yusuke, S., and Yoichi, S. Learning Gaze Biases with Head Motion for Head Pose-Free Gaze Estimation. *Image and Vision Computing* 32, 3 (2014), 169–179.
- [37] Franchak, J.M., Kretch, K.S., Soska, K.C., and Adolph, K.E. Head-Mounted Eye-Tracking : A New Method to Describe Infant Looking. *Child Development* 82, 6 (2010), 1738–1750.
- [38] Frank, M., Tofighi, G., Gu, H., and Fruchter, R. Engagement Detection in Meetings. *CoRR abs/1608.0*, (2016).
- [39] Fredricks, J. a, Blumenfeld, P.C., and Paris, a. H. School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research* 74, 1 (2004), 59–109.
- [40] Fu, Y., Leong, H.V., Ngai, G., Huang, M.X., and Chan, S.C.F. Physiological Mouse: Toward an Emotion-Aware Mouse. *Universal Access in the Information Society* 16, 2 (2017), 365–379.
- [41] Georgescu, M. and Zhu, X. Aggregation of Crowdsourced Labels Based on Worker History. *WIMS '14 Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, (2014).
- [42] Google. A google a day. <http://www.agoogleaday.com/>.

- [43] Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., and Lester, J.C. Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. *EDM*, (2013).
- [44] Graham, W., Euan, F., and Stephen, B. Multimodal Affective Feedback: Combining Thermal, Vibrotactile, Audio and Visual Signals. *ICMI 2016 Proceedings of the 18th ACM International Conference on Multimodal Interaction*, (2016), 400–401.
- [45] Granka, L. a., Joachims, T., and Gay, G. Eye-Tracking Analysis of User Behavior in WWW Search. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, (2004), 478 – 479.
- [46] Grinberg, N., Dow, P.A., Adamic, L. a., and Naaman, M. Changes in Engagement Before and After Posting to Facebook. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (2016), 564–574.
- [47] Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. Multi-Pie. *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, (2008).
- [48] Guanqing, L., Jiannong, C., Xuefeng, L., and Xu, H. Cushionware: a Practical Sitting Posture-Based Interaction System. *CHI EA '14 CHI '14 Extended Abstracts on Human Factors in Computing Systems*, (2014), 591 – 594.
- [49] Guo, Q. and Agichtein, E. Ready to Buy or Just Browsing?: Detecting Web Searcher Goals from Interaction Data. *SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, (2010), 130–137.
- [50] H.Meyer., C., Lasker, A.G., and A.Robinson, D. The Upper Limit of Human Smooth Pursuit Velocity. *Vision Research* 25, 4 (1985), 561–563.
- [51] Hall, M., Frank, E., Holmes, G., and Pfahringer, B. The WEKA Data Mining Software: An Update. *SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

- [52] Hansen, D.W. and Ji, Q. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (2010), 478–500.
- [53] Hart, S.G. and Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52, (1988), 139–183.
- [54] Hassenzahl, M. and Tractinsky, N. User Experience - a Research Agenda. *Behaviour & Information Technology* 25, 2 (2006), 91–97.
- [55] Hoffman, D.L., Novak, T.P., and Peralta, M. Building Consumer Trust Online. *Communications of the ACM* 42, 4 (1999), 80–85.
- [56] Huang, M.X., Kwok, T.C.K., Ngai, G., Leong, H.V., and Chan, S.C.F. Building a Self-Learning Eye Gaze Model from User Interaction Data. *Proceedings of the ACM International Conference on Multimedia - MM '14*, (2014), 1017–1020.
- [57] Huang, M.X., Li, J., Ngai, G., and Leong, H.V. StressClick. *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, (2016), 1395–1404.
- [58] Iida, A., Campbell, N., Iga, S., Higuchi, F., and Yasumura, M. A Speech Synthesis System with Emotion for Assisting Communication. *ITRW on Speech and Emotion Newcastle*, (2000).
- [59] Janssen, J.H., Ijsselstein, W.A., and Westerink, J.H.D.M. How affective technologies can influence intimate interactions and improve social connectedness. *International Journal of Human-Computer Studies* 72, (2014), 33–43.
- [60] Jennings, M. Theory and Models for Creating Engaging and Immersive Ecommerce Websites. *SIGCPR '00: Proceedings of the 2000 ACM SIGCPR conference on Computer personnel research*, (2000), 77–85.
- [61] JR, F., KR, S., EB, R., and PC., E. The World of Emotions is Not Two-Dimensional. *Psychological Science* 18, (2007), 1050–1057.
- [62] Kato, Y., Kanda, T., and Ishiguro, H. May I help you?: Design of Human-like Polite Approaching Behavior. *HRI '15 Proceedings of the Tenth Annual*

- ACM/IEEE International Conference on Human-Robot Interaction*, (2015), 35–42.
- [63] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., and Murthy, K.R.K.Improvements to Platt’s SMO Algorithm for SVM Classifier Design. *Neural Computation* 13, 3 (2001), 637–649.
- [64] Kim, S., Valente, F., Filippone, M., and Vinciarelli, A.Predicting Continuous Conflict Perception with Bayesian Gaussian Processes. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING* 5, 2 (2014), 187–200.
- [65] Kloos, H. and Van Orden, G.Voluntary Behavior in Cognitive and Motor Tasks. *Mind and Matter* 8, 1 (2010), 19–43.
- [66] Kotsia, I. and Pitas, I.Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Image Processing* 16, 1 (2007), 172–187.
- [67] Kuppens, P., Tuerlinckx, F., Russell, J.A., and Barrett, L.F.The Relation Between Valence and Arousal in Subjective Experience. *Psychological Bulletin*, 2012.
- [68] Kwok, T.C.K., Huang, M.X., Tam, W.C., and Ngai, G.Emotar: Communicating Feelings through Video Sharing. *IUI ’15 Proceedings of the 20th International Conference on Intelligent User Interfaces*, (2015), 374–378.
- [69] Lagun, D. and Agichtein, E.Infering Searcher Attention by Jointly Modeling User Interactions and Content Saliency. *SIGIR ’15 Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2015), 483–492.
- [70] Lalmas, M., O’Brien, H., and Yom-Tov, E.*Measuring User Engagement*. 2014.
- [71] Laufer, L. and Németh, B.Predicting User Action from Skin Conductance. *Proceedings of the 13th international conference on Intelligent user interfaces - IUI ’08*, (2008), 357.
- [72] Law, E.L.C., Van Schaik, P., and Roto, V.Attitudes Towards User Experience (UX) Measurement. *International Journal of Human Computer Studies* 72, 6 (2014), 526–541.

- [73] Li, J., Ngai, G., Leong, H.V., and Chan, S.C.F. Your Eye Tells How Well You Comprehend. *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, (2016), 503–508.
- [74] Li, J., Ngai, G., Va Leong, H., and Chan, S. Multimodal Human Attention Detection for Reading. *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16*, (2016), 187–192.
- [75] Liu, Y. and Liu, M. An Online Learning Approach to Improving the Quality of Crowd-Sourcing. *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, (2015), 217–230.
- [76] Mathur, A., Lane, N.D., and Kawsar, F. Engagement-Aware Computing: Modelling User Engagemet from Mobile Contexts. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, (2016), 622–633.
- [77] Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., and Cohn, J.F. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Transactions On Affective Computing* 4, 2 (2013).
- [78] McDaniel, B., D’Mello, S., King, B., Chipman, P., Tapp, K., and Grasser, A. Facial Features for Affective State Detection in Learning Environments. *Proceedings of the Annual Meeting of the Cognitive Science Society* 29, (2007).
- [79] McDuff, D., Kaliouby, R., Demirdjian, D., and Picard, R. Predicting Online Media Effectiveness Based on Smile Responses Gathered Over the Internet. *2013 Tenth IEEE International Conference on Automatic Face and Gesture Recognition*, (2013).
- [80] McDuff, D., Kaliouby, R. El, Kodra, E., and Picard, R. Measuring Voter’s Candidate Preference Based on Affective Responses to Election Debates. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, (2013), 369–374.
- [81] Metzger, M.J. Privacy, Trust, and Disclosure: Exploring Barriers to Electronic Commerce. *Journal of Computer-Mediated Communication* 9, 4 (2004).

- [82] Michel, P. and Kaliouby, R. El.Real Time Facial Expression Recognition in Video using Support Vector Machines. *ICMI'03*, (2003), 258–264.
- [83] Mikkelsen, S., Vilstrup, I., Lassen, C.F., Kryger, A.I., Thomsen, J.F., and Andersen, J.H. Validity of Questionnaire Self-Reports on Computer, Mouse and Keyboard Usage during a Four-Week Period. *Occupational and environmental medicine* 64, 8 (2007), 541–547.
- [84] Mok, R.K.P., Chang, R.K.C., and Li, W. Detecting Low-Quality Workers in QoE Crowdstesting: A Worker Behavior-Based Approach. *IEEE Transactions on Multimedia* 19, 3 (2017), 530–543.
- [85] Monkaresi, H., Bosch, N., Calvo, A.R., and D’Mello, S.K. Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing* 8, 1 (2017), 15–28.
- [86] Nakano, Y.I. and Ishii, R. Estimating User’s Engagement from Eye-Gaze Behaviors in Human-Agent Conversations. *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '10*, (2010), 139.
- [87] O’Brien, H.L. and Toms, E.G. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *Journal of the American Society for Information Science and Technology* 59, 6 (2008), 938–955.
- [88] O’Brien, H.L. and Toms, E.G. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69.
- [89] O’Brien, H.L. and Toms, E.G. Examining the generalizability of the User Engagement Scale (UES) in Exploratory Search. *Information Processing and Management* 49, 5 (2013), 1092–1107.
- [90] Partha Pratim, D., A. F. M. Rashidul, H., and Dipankar, D. Detection and Controlling of Drivers’ Visual Focus of Attention. *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, (2017), 301–307.
- [91] Paul, C., Niall, A., and Brent, H. Voting Experts: An Unsupervised Algorithm for Segmenting Sequences. *Intelligent Data Analysis* 11, 6 (2007), 607–625.

- [92] Pavlovic, V.I., Sharma, R., and Huang, T.S. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. *IEEE Transactions On Pattern Analysis And Machine Intelligence* 19, 7 (1997).
- [93] Peng, H., Hu, B., Zheng, F., et al. A method of identifying chronic stress by EEG. *Personal and Ubiquitous Computing* 17, 7 (2013), 1341–1347.
- [94] Ramot, D. Average Duration of a Single Eye Blink. <http://bionumbers.hms.harvard.edu/bionumber.aspx?s=y&id=100706&ver=0>, 2008.
- [95] Read, J., Macfarlane, S., and Casey, C. Endurability, Engagement and Expectations: Measuring Children's Fun. *Interaction Design and Children* 2, January (2002), 1–23.
- [96] Reichle, E.D., Reineberg, a. E., and Schooler, J.W. Eye Movements During Mindless Reading. *Psychological Science* 21, 9 (2010), 1300–1310.
- [97] Rodrigue, M., Son, J., Giesbrecht, B., Turk, M., and Höllerer, T. Spatio-Temporal Detection of Divided Attention in Reading Applications Using EEG and Eye Tracking. *IUI '15 Proceedings of the 20th International Conference on Intelligent User Interfaces*, (2015), 121–125.
- [98] Roorda, A. and Campbell, M.C.W. Slope-based Eccentric Photorefraction: Theoretical Analysis of different Light Source Configurations and Effects of Ocular Aberrations. *J. Opt. Soc. Am. A* 14, 10 (1997), 2547 – 2556.
- [99] Salvucci, D.D. and Goldberg, J.H. Identifying Fixations and Saccades in Eye-Tracking Protocols. *ETRA '00 Proceedings of the 2000 symposium on Eye tracking research & applications*, (2000), 71–78.
- [100] Sánchez, J.A., Elizabeth S., F., and Natalia, V. Challenges for Establishing a Latin American Community in HCI / UX. *CSCW Workshop*, (2014).
- [101] Sebe, N., Lew, M.S., Cohen, I., Sun, Y., Gevers, T., and Huang, T.S. Authentic facial expression analysis. *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference*, (2004), 517–522.
- [102] Senechal, T., Bailly, K., and Prevost, L. *Impact of action unit detection in automatic emotion recognition*. Springer-Verlag London Limited, 2012.

- [103] Sheng, H., Lockwood, N.S., and Dahal, S. Eyes don't lie: Understanding users' first impressions on websites using eye tracking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (2013), 635–641.
- [104] Sheng, V.S., Provost, F., and Ipeirotis, P.G. Get another label? improving data quality and data mining using multiple, noisy labelers. *KDD*, (2008), 614–622.
- [105] Slanzi, G., Balazs, J. a., and Velásquez, J.D. Combining Eye Tracking, Pupil Dilation and EEG Analysis for Predicting Web Users Click Intention. *Information Fusion 35*, (2017), 51–57.
- [106] Soleymani, M., Lichtenauer, J., and Pun, T. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING 3*, 1 (2012).
- [107] Summerfield, P. Getting to Engagement: Media Agencies Start to Realize it's Part Art, Part Science. *Strategy Magazine*, 2006.
- [108] Tan, Y., Kanade, T., and Cohn, J.F. Recognizing Action Units for Facial Expression Analysis. *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, (2000).
- [109] Tomas, K., Jan, M., Pavla, D., and Katerina, S. Searching Time Time Series Series Based Based On On Pattern Pattern Searching Extraction Using Using Dynamic Dynamic Time Time Warping Warping Extraction. *Databases, Texts, Specifications, and Objects (Dateso 2013)*, (2013), 129–138.
- [110] UCL. Blink and you miss it! <http://www.ucl.ac.uk/media/library/blinking>, 2006.
- [111] Virpi, R., Effie, L., Arnold, V., and Jettie, H. *User Experience White Paper. Bringing Clarity to the Concept of User Experience*. 2011.
- [112] Wacharamanotham, C. Making Bare Hand Input More Accurate. *CHI EA '14 CHI '14 Extended Abstracts on Human Factors in Computing Systems*, (2014), 307–310.
- [113] Wang, Y., Shen, T., Yuan, G., Bian, J., and Fu, X. Appearance-Based Gaze Estimation Using Deep Features and Random Forest Regression. *Knowledge-Based Systems 110*, (2016), 293–301.

- [114] Warnock, D. and Lalmas, M. An Exploration of Cursor tracking Data. *arXiv:1502.00317*, (2015).
- [115] Whitehill, J., Serpell, Z., Yi-Ching Lin, Y.-C., Foster, A., and Movellan, J.R. The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.
- [116] Wiebe, E.N., Lamb, A., Hardy, M., and Sharek, D. Measuring Engagement in Video Game-Based Environments: Investigation of the User Engagement Scale. *Computers in Human Behavior* 32, (2014), 123–132.
- [117] Williams, O., Blake, A., and Cipolla, R. Sparse and Semi-Supervised Visual Mapping with the S3GP. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1, (2006), 230–237.
- [118] Xiong, X. and Torre, F.D. la. Supervised Descent Method and Its Applications to Face Alignment. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013).
- [119] Zeng, Z., Tu, J., Pianfetti, B., Liu, M., Zhang, T., and Zhang, Z. Audio-visual Affect Recognition through Multi-stream Fused HMM for HCI. *IEEE Trans. Multimedia* 10, 4 (2005), 570–577.
- [120] Zeng, Z.H., Pantic, M., Roisman, G.I., and Huang, T.S. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions On Pattern Analysis And Machine Intelligence* 31, 1 (2009).
- [121] Zhang, X., Sugano, Y., Fritz, M., Bulling, A., and Planck, M. Appearance-Based Gaze Estimation in the Wild. *Computer Vision and Pattern Recognition (CVPR 15)*, (2015).
- [122] Zheng, N., Paloski, A., and Wang, H. An Efficient User Verification System Using Angle-Based Mouse Movement Biometrics. *ACM Transactions on Information and System Security* 18, 3 (2016), 1–27.
- [123] Live Gaze-Based Authentication and Gaming System. *International Conference on Human Aspects of Information Security, Privacy, and Trust. HAS 2017 10292*, (2017).