UNDERSTANDING HUMAN COMPREHENSION AND ATTENTION IN READING

LI JIAJIA

Ph.D

The Hong Kong Polytechnic University

2017

The Hong Kong Polytechnic University

Department of Computing

Understanding Human Comprehension and Attention in Reading

Li Jiajia

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

January 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)

Li Jiajia (Name of student)

Abstract

Reading is one of the most common computer interaction activities and also one of the most fundamental means of knowledge acquisition. With the development of computing technologies and the growing popularity of e-Learning platforms, understanding human attention and comprehension through reading behaviors has the potential to become an important means to enhance the learning experience and effectiveness.

Eye gaze pattern is known to play an important role in the study of reading behaviors since reading can be considered as a task where visual processing and sensorimotor control takes place in a highly structured visual environment [79]. Many studies have shown that eye movement and eye behavior during reading is closely related to cognitive human mental states, such as comprehension and attention [81][88][89].

There are two main drawbacks in current state-of-the-art research on comprehension and attention detection based on eye gaze patterns. First, many of them use expensive and intrusive devices, like the electrooculography systems, to track the eye movement, or detect the user's mental state as ground truth, through the use of electroencephalography (EEG) devices. Second, numerous methods study how lexical and linguistic variables affect the eye gaze behavior during reading. These methods therefore rely on the availability of linguistic analysis of the reading materials.

Addressing the limitations mentioned above, we conduct experiments with human subjects and do an in-depth study of eye gaze patterns related to the change of comprehension level and attention level during reading. Both Tobii eye tracker and off-the-shelf webcam are used to capture the eye gaze signals based on which the eye gaze features are extracted. By adopting machine learning algorithms, we conduct feature evaluation and compare the classification performance with different kinds of eye gaze features. From the investigation, we have a better understanding of relation between the studied human mental states, i.e. comprehension and attention, and certain eye gaze patterns. We also find that the features extracted based on accurate eye gaze location on the screen captured by Tobii eye tracker contribute more to the comprehension and attention level detection during reading.

In order to recognize human mental states, input signals reflecting human mental states need to be acquired and processed. Under traditional KVM (keyboard-video-

mouse) settings, input signals are mostly tied to keyboard and mouse dynamics. One can deduce some information about human mental states and affects from keyboard [12][111] and mouse [110][123], but the accuracy is not particularly high. Thanks to the popularity of interactive social networking applications, the webcam has become a de facto device. Recent research in video processing and machine learning has demonstrated that human affects can be recognized via webcam video, noticeably via human facial features [127]. Inspired by previous research, we look into other modalities, i.e. facial expressions and mouse dynamics, for attention detection during reading. A two-level facial feature extraction approach is proposed to represent the static and dynamic states of the facial expressions of the subjects. Similarly, the mouse dynamic features are extracted from the captured log mouse events and evaluated for reading attention detection.

To evaluate our method, we apply machine learning techniques to build up user-independent models to recognize human attention and comprehension level on reading tasks. We compare the performances of models built on single modality and multiple modalities. The findings suggest that the multimodal approach outperforms the unimodal approach in our studies. The results also demonstrate that eye gaze pattern and facial expressions show more potential in predicting attention level than the mouse dynamics, which may be caused by the rare usage of mouse as an input device in the reading task.

List of Publications

Jiajia Li, Grace Ngai, Hong Va Leong, Stephen Chan. 2016. Multimodal Human Attention Detection for Reading from Facial Expression, Eye Gaze, and Mouse Dynamics. *Applied Computing Review* 16, 3: 37-49.
http://doi.org/10.1145/3015297.3015301

Jiajia Li, Grace Ngai, Hong Va Leong, Stephen Chan. 2016. Your Eye Tells How Well You Comprehend. *Proceedings of 40th IEEE International Conference on Computer Software and Applications - COMPSAC'16*. 503-508.
https://doi.org/10.1109/COMPSAC.2016.220

Jiajia Li, Grace Ngai, Hong Va Leong, Stephen Chan. 2016. Multimodal Human Attention Detection for Reading. *Proceedings of 31st ACM/SIGAPP Symposium on Applied Computing - SAC'16*. 187-192. https://doi.org/10.1145/2851613.2851681

Jiajia Li, Grace Ngai, Stephen C.F. Chan, Kien A. Hua, Hong Va Leong, Alvin Chan. 2014. From Writing to Painting: A Kinect-Based Cross-Modal Chinese Painting Generation System. *Proceedings of 22nd ACM International Conference on Multimedia - MM'14*. 57-66. https://doi.org/10.1145/2647868.2654911

Michael Xuelin Huang, **Jiajia Li**, Grace Ngai, Hong Va Leong. 2016. StressClick: Sensing Stress from Gaze-Click Patterns. *Proceedings of 24th ACM International Conference on Multimedia - MM'16*. 1395-1404.
http://doi.org/10.1145/2964284.2964318

Michael Xuelin Huang, **Jiajia Li**, Grace Ngai, Hong Va Leong. 2017. ScreenGlint: A Practical Cue to Gaze Estimation on Smartphones. *Proceedings of 34th CHI Conference on Human Factors in Computing Systems - CHI'17*. 2546-2557.
http://doi.org/10.1145/3025453.3025794

Jiajia Li, Michael Xuelin Huang, Grace Ngai, Hong Va Leong. "Detecting Reading Comprehension Level through Eye Movement Pattern Analysis". In submission.

Jiajia Li, Michael Xuelin Huang, Grace Ngai, Hong Va Leong. "A Multimodal, Non-intrusive Approach to Reading Comprehension Detection". In submission.

Acknowledgments

I would like to express my sincere gratitude to all the people that have assisted me to complete this degree. The following acknowledgments are by no means exhaustive, for which I apologize.

I would love to thank Grace Ngai for being the best advisor I could have wished for. She gave me the wonderful experience to be in the CHILab for these years with unwavering support and encouragement.

I would also like to thank the professors in my research group: Hong-Va Leong, Alvin Chan, and especially my co-supervisor Stephen C.F. Chan. They generously shared the knowledge, experience, and inspiring thoughts in our every week discussion. This work would never have been done without their generous contributions.

I am also deeply grateful to Kien A. Hua for being a fantastic supervisor and friend, during the period of his visit in the Hong Kong Polytechnic University.

I have had great pleasure working with members in CHILab: Michael Xuelin Huang, Yuanyuan Wang, Will Tang, Kenneth Lo, Kin Lau, Tiffany Kwok, Eugene Fu, Jun Wang and Hugo Sun. The creativity of all my colleagues has been a constant inspiration throughout my time.

My utmost thanks go to my parents and my husband, who unconditionally support me in all my decisions.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Understanding Reading Behaviors through Eyes

The development of computer technologies has made digital devices and environments, such as mobile phones and tablet computers, an alternative media for presenting text. As one of the most common computer interaction activities and also one of the most fundamental means of knowledge acquisition, reading is a task that frequently happens in people's daily life. In this context, understanding reading behaviors is vital to detect human mental state and enhance the interaction experience.

In the same way as a human teacher might observe his/her students to gauge their reading behaviors, it is easy to see how good understanding of reading behaviors through computers would be helpful for intelligent digital platforms to assist the users. Although there has been much research on reading behavior analysis in the past decades, eye gaze behaviors obtain the most interest in understanding reading behaviors considering that reading is a task where visual processing and sensorimotor control takes place in a highly structured visual environment [79]. In general, eye movements during reading can be categorized into saccades and fixations, which alternately occur during reading [81]. A saccade is a fast movement of the eye, which is usually in a direction parallel to that of the text. A fixation is the maintaining of the visual eye gaze on a single location. The purpose of a saccade is to locate a point of interest on which to focus, while processing of visual information takes place during fixations. The studies of the eye movement patterns shed light on the relation between reading behaviors, reading contexts and the reader's mental states. For example, previous research has demonstrated that readers who are experiencing difficulty in processing

Figure 1-1. The EOG systems (left) [131] and head-mounted eye trackers (right) [132] are intrusive devices widely used to sense the eye movements.

the visual information tend to make more fixations and the fixation duration becomes longer [47][84][82]. Moreover, under these circumstances, readers often make backwards or regressive saccades, to re-read the materials and get a better comprehension of the text [50].

There are many kinds of devices used in the state-of-the-art research on reading behavior analysis through eye movements. Many of them are intrusive devices, which may be uncomfortable to the users. For example, electrooculography (EOG) systems (as shown in the left picture in Figure 1-1) are used for eye movement data recording in many studies [18][7]. The eye can be modeled as a dipole with its positive pole at the cornea and its negative pole at the retina. Assuming a stable corneo-retinal potential difference, the eye is the origin of a steady electric potential field. The electrical signal that can be measured from this field is called the electrooculogram (EOG) [18]. When using the EOG systems, several skin electrodes are placed around the users' eyes and forehead. Head-mounted cameras (as shown in the right picture in Figure 1-1) are also used in video-based eye trackers to track the eye gaze locations. Although the head-mounted eye tracker detects the eye gaze location with high accuracy, they are usually expensive and complicated with cameras installed on the frame, which is not convenient for pervasive applications [30][108].

In contrast, non-intrusive and low-cost devices are attracting more and more attention in the research of human-computer interaction. Remote commodity eye trackers are becoming commonplace in eye movement studies for their high reliability, easy usability and relatively low cost. One of the advantages of using remote eye trackers is that they are reasonably precise and also non-intrusive. The precise eye gaze location information is very important in understanding reading behaviors in different contexts. There are a growing number of researchers using remote eye trackers to analyze the users' reading behaviors. For example, there has been research on eye movement analysis while reading a web page [9] and evaluating the list of ranked results of WWW search engines [32] using the Tobii eye trackers. Compared with using devices which restrain the user's head and neck movements, using the remote eye trackers tends to result in behaviors that are more "normal" [22]. Although these studies explored web reading or searching behaviors by analyzing the eye gaze data captured by eye trackers, their aims are to investigate how the design of the webpage or the feedback provided to the user during the web-reading process affects the reading pace or pattern, instead of understanding human mental states from the eye gaze behaviors.

In addition to understanding reading behaviors through eye movement patterns, there are numerous methods studying how lexical and linguistic variables affect the eye movement patterns during reading [86][47][85]. The basic idea is to find out how the eye movement patterns are affected by the word's lexical properties, such as length and frequency, the word's meaning and the sentence context. These methods usually involve complicated word and semantic pre-processing, which is considerably time consuming. Moreover, they often need to rely on linguistic analysis of the specific materials. However, with the increasing amount of information available in the digital environment, it is likely that more efficient approaches for reading behavior understanding will be developed in the future.

1.1.2 Detecting Affects during Reading

Nowadays, the widespread use of computer technologies and social networks has made reading one of the most common and important activities in human-computer interaction. A report in 2014 said that 21% of American adults reported that they had read an e-book in the past year and this number was still increasing [57]. The shift of reading from printed to digital materials is more obvious in education. An increasing number of individuals, corporations, and institutions are turning to e-Learning as they recognize its effectiveness and convenience. Therefore, the development of intelligent e-Learning platforms which could be aware of the user's mental states is of much interest to both computer science and education. Given how e-Learning platforms involves much reading, it is easy to see how a good understanding of reading behavior and successful detection of human mental states would be helpful for intelligent e-Learning platforms, to provide help to the user, or to adjust the material to improve the learning effectiveness. In other contexts, detecting human affects during reading is also valuable to explore the level of enjoyability and the usefulness of the conveyed information.

The automatic detection of human affect is not a new topic. Since the concept of "Affective Computing" was proposed almost 20 years ago, it has attracted great interest in the domain of computer science. Human affect understanding is one of the leading topics in affective computing since affect is a fundamental component of human expression and communication [130]. The fact that computers could be aware of human affect will significantly facilitate human-computer interaction.

In affective computing, the emotional state of the user can be measured through his/her physiological and behavioral signals. The physiological signals sensed from different parts of human body have been demonstrated to be promising for exploring human affect. Autonomic measurements can be used to objectively detect emotion-related physiological responses of autonomic nervous system (ANS), such as skin conductance responses (SCRs) and heart rate variability (HRV). However, the ANS responses are considered not representative or reliable to reflect some emotional state. There are also inconsistent findings across studies on the mapping between ANS response and emotions [65]. Neurophysiologic measurement based on electrophysiological and neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) can detect a wide range of dynamics of the emotional state by directly accessing the fundamental structure in the brain from which an emotional state emerges [71], and hence, clearly provide a direct and comprehensive means for emotion recognition. Nevertheless, the high cost and immobility of fMRI and MEG prevents them from being used for practical emotion recognition systems in the real life.

In recent years, electroencephalography (EEG) recording devices have been developed to become more cost-effective and mobile with increased practicability and less physical restriction [17], and so they have become more widely used for emotion recognition. However, there are two drawbacks of using EEG for affect detection. First, the devices used to record EEG are quite intrusive, since they need to be pasted or worn on the subject's head during the whole signal recording process [8][125]. Second, the EEG signal has poor spatial resolution and high susceptibility to noise. These make it impractical to use physiological signals for affect detection in daily interaction scenarios.

Comparatively speaking, understanding human affect from the behavioral signals is promising for pervasive applications. In human-computer interaction, various modalities can be involved when human express themselves and communicate with each other. Under these circumstances, it is essential to take into consideration of the contexts of the applications because the modalities and sensors that need to be involved in human affect interpretation vary with different applications. For example, speech text [49] and vocal tones [130] have been demonstrated to be informative for emotion expression. These two channels are available and reliable in contexts in which speaking is an appropriate form of communication.

In other contexts where human-computer interactions are accomplished in relatively silent environments, vision-based technologies are widely used in human behavior analysis and affect detection. Since the webcam has become a common piece of equipment, there has been much development of computer vision technologies that are based on webcam signals. By adopting proper technologies, human affects can be recognized effectively from webcam video [41].

Facial expression is a crucial channel that has been widely investigated for affect detection with vision systems. Previous research proposed many effective methods to encode human facial expressions from the original face images [72][97][120]. Many efforts have been made to study facial expression to successfully infer basic human affects like happiness, sadness, anger, fear, surprise and disgust [27]. These have also been extended to the recognition of fatigue [48], embarrassment [48], and pain [5]. Moreover, high-level mental states like agreeing, disagreeing, interested, thinking, concentrating, unsure, and adult attachment have been investigated as well in previous research [126]. Facial expression recognition, being a powerful technique, also finds its application in understanding the student engagement in a classroom [3]. Cognitive engagement is found to have close relationship with a person's cognitive abilities, including focused attention, memory, and creative thinking in learning [4].

The eyes and their movements, as one of the most salient features of the human face, play an important role in expressing a person's emotions and cognitive processes. The importance of the eye movements to the individual's perception of the visual world is widely acknowledged. For reading in particular, the visual contents are mainly texts and the eye movements can usually be classified into several typical eye gaze patterns, such as saccades and fixations. A fair amount of work has been carried out to analyze the reader's mental state through the eye gaze patterns. For example, it is found that readers who are experiencing difficulty in processing the visual information tend to make more fixations and the fixation duration becomes longer [47][84][82]. Moreover, under these circumstances, readers often make backwards or regressive saccades, to re-read the materials and get a better comprehension of the text [50].

To recognize the eye gaze patterns, it is crucial to obtain the eye gaze locations on the reading interface. Nowadays, the availability of commodity eye trackers has led to much work that uses eye gaze localization to detect human mental states during reading. For example, researchers have studied the eye gaze patterns under human mind wandering [88], boredom and disengagement [24] with the use of eye trackers.

However, these studies did not thoroughly investigate eye gaze behaviors. In [88], the reading behaviors were measured mainly via fixations to detect mindless reading. In [24], the disengagement of the user was predicted by simply detecting when he/she looked away from the screen for an extended period of time. However, the characteristics of some typical eye gaze behaviors that are supposed to play important roles in mental state detection, such as the rate of eye blinks, the patterns of eye fixations and eye saccades, were not studied in this research.

The trend to adopt gaze-aware systems in daily human-computer interaction and social interaction is growing with the availability of the equipment [38]. Compared to eye trackers, off-the-shelf webcams are cheaper and more versatile, which give them more potential for pervasive applications. A great number of efforts have been put on vision-based eye detection and eye tracking in the past few years. The proposed technologies efficiently deal with the problem of identifying the eyes with large variability of appearance and dynamics [124][115][39]. However, effective eye gaze estimation that uses only a single camera and a single light source is still challenging. It has been demonstrated that the use of single camera and single light usually constrains gaze estimation in head variant scenarios [34][99].

In digital reading environments, readers usually need to use the input and output devices to accomplish the reading tasks. The signals captured by various input devices often bears information that reflects the reader's mental states. Under traditional KVM (keyboard-video-mouse) settings, input signals are mostly tied to keyboard and mouse dynamics. One can deduce some information about human affect from the keyboard [12][111] and the mouse [110][123], but the accuracy is not particularly high.

Since there are so many potential modalities involved in human-computer interaction, choosing the useful information is an important issue. Some previous work relied on a single modality to detect certain affects. For example, facial expression is widely studied to infer basic human emotion [27] and high-level mental states [48][3][4]. Unimodal affect detection achieves good performance when the modality has a strong correlation with the target affects. Nevertheless, existing research [94][73] suggest that it is beneficial to integrate multimodal information to recognize human affects automatically. Since different modalities convey different information to reflect human affects, the advantage of taking account of multiple aspects of human behaviors is obvious. Researchers have showed that learning from more than one modality achieved better classification than that with individual modality [98][36]. To date,

much work has been done in the study of fusing multimodal information. For instance, research has been done to detect basic emotions by integrating visual cues, including facial expression, body gesture and head movement etc., [37][35][76]. However, the research on multimodal human affect detection and behavior understanding during reading is still in the early stage.

1.2 Study Overview

As one of the most common activities in human-computer interaction, reading not only plays a role in information transformation, but also provides cues to understand the readers' mental states. With the development of computer science and the widespread use of digital media, the research on reading behavior understanding and mental state detection is attracting more and more interest. Since reading is a complex cognitive process, the reader's comprehension level and attention level are critical mental states raised in the reading tasks. This thesis proposes effective methods of predicting the reader's comprehension level and attention level in the common reading tasks. To gain comprehensive understanding of the relation between the target mental states and the human behaviors, we conduct systematic human behavior analysis based on multiple modalities collected using non-intrusive devices.

1.2.1 Detecting Reading Comprehension from Eye Gaze

Although there has been a certain degree of success reported in existing research on reading behavior understanding and mental state detection, there are still significant challenges, such as the aforementioned constraints of using intrusive devices, conducting content-based eye movement analysis, and building user-dependent model. This thesis attempts to address these issues by introducing user-independent models for reading comprehension level detection through eye movement analysis. We obtain in-situ data, by inviting human subjects to carry out experiments in reading a variety of English articles while recording their eye movements with a non-intrusive remote eye tracker, which does not require the reader to bear or wear special devices or sensors. The difficulty level of the articles is varied so as to induce different levels of comprehension with the users.

The recorded eye gaze data is firstly denoised and subsampled for preprocessing. The preprocessed data is thoroughly analyzed to identify three typical eye gaze

behaviors, which are eye blink, eye fixation and eye saccade. We construct features that will be used to describe these behaviors within a given time period of reading. In addition to the behavior-based features, we also adopt gaze-based features to capture the general characteristics of the unfiltered eye gaze positions, including the kurtosis and skewness of the eye gaze data. Finally, we incorporate some features that are meant to capture the context of the reading task through combining basic information on the read article and the overall task.

Our goal of this study is to produce a set of resilient user-independent models, which are universal to different users and able to accommodate new unseen users, to predict the reading comprehension level. We analyze the performance of our method in different real-use situations with the extracted features, and we explore different ways of building the user-independent model, in particular running prediction models that consider different lengths of available eye gaze segments for recognition. This evaluation gives us a sense of the confidence levels we can expect of our prediction with respect to the length of available data segment. We also investigate the indicative eye gaze features on different datasets. These eye gaze features represent the eye movement patterns during reading. The set of features that contribute to the comprehension level detection may change in different contexts. To explore this problem, we apply machine learning techniques to identify a subset of useful features capable of assisting in the determination of human comprehension level for each model mentioned above. Our findings reveal the consistency of the contribution of the eye gaze features for models built with different segment granularity.

1.2.2 Multimodal Reading Attention Detection

Apart from the level of comprehension, the level of attention is another critical cognitive mental state involved in the reading task. It is common to observe the attention drift away from the task at hand to off-task thoughts because of the task itself or the environment [52][53]. In the learning context, detecting the reader's attention level during reading can help to compensate for the negative effects on the performance of the reading task and potentially provide necessary feedback to the reader in time.

Our study on reading comprehension detection and much existing research [13][28] shows the important role played by eye gaze pattern in reflecting human cognitive mental states. Although in some cases, considering eye gaze pattern alone can lead to promising performance for affect detection, a large number of previous studies on

affective computing demonstrate the advantage of fusing the information expressed by different modalities over considering single modality. In the digital reading environments, readers usually need to use various input and output devices to accomplish the reading tasks. The behavioral signals generated by the input devices often reflect the reader's mental states. We, therefore, propose a multimodal approach for attention level detection in reading with ubiquitous hardware available in most computer systems, namely, webcam and mouse.

We invite human subjects to carry out experiments in reading English articles while being subjected to different kinds of distraction to induce them into different levels of attention. During the experiment, a webcam installed on the monitor is used to record the subjects' frontal face view non-intrusively, and a C++ program run at the background to capture and log mouse events. We investigate the data obtained from each modality for feature extraction and evaluation before building a user-independent model for the attention level detection.

Based on the video clips recorded by the webcam, we present a two-level facial feature extraction approach in our work: frame-level and segment-level. 66 facial landmarks are tracked to extract 26 frame-level facial features. After performing frame-level facial feature extraction, we extract three kinds of segment-level facial features based on the frame-level facial feature vectors reflecting different statistical behaviors. Each segment is composed of a good number of frames. The first set of segment-level features derived from the 26 frame-level facial features is calculated as the mean value of the features of the embedded frames. The second set of segment-level features is computed based on moving windows representing the overall change of each frame-level feature in a given time period. The third set of segment-level features calculates the change of the whole face in the video clip with respect to the neutral face, which is defined as the face in the first frame of the video clip.

We extract eye gaze features from the webcam videos by eye gaze tracking and eye gaze behavior recognition. In this study, we analyze three kinds of eye gaze behaviors for reading attention detection, including eye blink, eye fixation and eye saccade. 6 landmarks associated with the contour of each eye are identified. We establish the eye geometry, namely, the eye openness, the relative horizontal position and vertical position of the eye gaze based on the landmark positions. Eye openness is employed in the detection of the eye blink, whereas temporal changes in the horizontal and vertical positions of the eye gazes are adopted in the detection of eye fixation and saccadic

movements. After the three different eye gaze behaviors have been identified from the sequence of eye gaze positions, we construct 9 statistical features that will be used to describe these three behaviors and predict the attention level for each video clip.

Mouse dynamics have been shown to provide indicative information for affect detection in various research [110][123]. In this study, we attempt to relate mouse dynamics with human reading attention level, by analyzing typical mouse dynamics, including mouse click, mouse movement and mouse scrolling. Similar to facial expression recognition, we process raw mouse events to establish mouse dynamics over time. We then extract features representing mouse dynamics for each segment to align with the segment in the video clip. This enables signal fusion among the different modalities, namely, mouse signals and webcam signals.

Our work extracts an initial set of 80 facial features, 9 eye gaze features and 7 mouse features, which is too many to be effective for practical real-time recognition, especially for facial features. After extracting the set of potentially useful features, feature selection is conducted to remove non-indicative features, which is known to improve classification performance in pattern recognition and machine learning applications. We adopt the wrapper method for feature selection which is reported to outperform filter method by considering the relationship between different features and selecting one feature subset that is the best for the chosen classifier [106]. We adopt the best first search approach for the efficiency, based on the Linear Support Vector Machine (SVM) for classification. This filtering step is very efficient in reducing the set of potential facial features from 80 down to 11. After initial feature selection in trimming down the feature set to a manageable size, we can explore different feature combinations via an exhaustive search for the optical feature set to build up our attention level recognition model. We end up with 7 top facial features, 5 top eye gaze features and 3 top mouse dynamics as the best combination for the attention level detection.

We evaluate our multimodal attention detection approach by building user-independent models based on the combined dataset of all subjects. We compare the classification performance of our multimodal approach with the performance produced using only a single modality. The results illustrate that the multimodal models perform better than the single modality ones, achieving higher correct classification rate (CCR) and F-measures. Moreover, we explore the contribution of each individual modality by conducting three more experiments based on (*a*) combined facial and eye gaze features, (*b*) combined facial and mouse features, and (*c*) combined eye gaze and mouse features.

We find that they achieve a performance between that achieved by single component modalities and that achieved by the full set of modalities.

1.3 Thesis Aims and Outline

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

- To investigate the detection of comprehension level based on a commonly occurring task, i.e. reading, by investigating the eye gaze behavior with non-intrusive devices.
- To identify indicative features that are effective in describing specific eye gaze behaviors and build user-independent models to recognize the level of comprehension on various lengths of available eye gaze data for a human in reading tasks.
- To propose a multimodal approach to the detection of attention level during reading through human facial features, eye gaze features and mouse dynamic with off-the-shelf devices.
- To represent the facial expressions, eye gaze behaviors and typical mouse dynamics properly, and explore the useful features of each modality for the attention level detection.
- To compare the performance of the unimodal and multimodal user-independent models for the attention level detection for a deeper analysis of the different modalities.

The reminders of this thesis will cover the following material:

Chapter 2 provides the literature reviews that describe the background information for the work introduced in this thesis. More specifically, it reviews the research efforts on eye gaze behavior analysis, gaze estimation, and human affect and mental state detection for reading.

Chapter 3 explores comprehension level detection for reading based on eye gaze behavior analysis. The eye gaze features are extracted and evaluated to build user-independent models for reading comprehension level detection on various lengths of available eye gaze data.

Chapter 4 proposes a multimodal approach to attention level detection in reading based on facial expressions, eye gaze behaviors, and mouse dynamics. The performance of unimodal classification and multimodal classification is also discussed.

Chapter 5 introduces other contributions we have made that are related to or beyond the scope of this thesis.

Chapter 6 concludes the whole thesis and plans for the future research.

Chapter 2

Literature Review

This chapter begins with a review of relevant topics of eye gaze behaviors, including vision-based eye gaze estimation, fixation identification, and gaze movement studies in reading. This is followed by the review of affect and mental state detection in reading. The purpose of this chapter is to provide an understanding of the prior research in the fields of gaze analysis and multimodal affect and mental state detection, especially in reading.

2.1 Eye Gaze Behaviors in Reading

Reading is one of the most common computer interaction activities and also one of the most fundamental means of knowledge acquisition. As a complex cognitive task, reading involves attention level, comprehension ability, visual interest, oculomotor processing constraints. From a mechanical view point, reading can be considered as a task where visual processing and sensorimotor control takes place in a highly structured visual environment [79] which is a symbolic and abstract source of information. In the past decades, a great number of studies have been done on understanding reading behaviors and related human mental states based on the eye gaze patterns. To successfully identify the eyes and measure the eye movement, many efforts have also been made to propose solid eye-tracking technologies in different contexts.

2.1.1 Eye Movements in Reading

Human behaviors are often better reflected by studying human-oriented signals. In reading tasks, the eye is the essential sensory organ involved, besides the brain. In particular, the eye is known to play an important role in reading by researchers, in addition to the more obvious electroencephalogram (EEG) signal, oriented from the human brain. Towards an eye tracker, it is capable of detecting the location of the eye gaze on the screen, which in turn can be processed into eye movement information.

Many research conclusions have been drawn on eye movements in reading in the domain of psychology and computer science. In general, eye movements during reading can be categorized into saccades and fixations, which alternately occur during reading [81]. A saccade is a fast movement of the eye, which is usually in a direction parallel to that of the text. It is from left to right for western languages or right to left in some others. It is from top to bottom for Chinese in some cases. A fixation is the maintenance of the visual eye gaze on a single location. The purpose of a saccade is to locate a point

of interest on which to focus, while processing of visual information takes place during fixations. Humans typically alternate saccades and fixations in daily life.

The eye movements when reading silently differ from those when reading aloud. The research we introduce in this chapter is for silent reading. When reading English, the length of a saccade can be as short as a single word or can span across multiple sentences of paragraphs. The average saccade length is 7-9 letter space for readers of English and other alphabetic writing system [100]. The number of letters traversed by saccades is relatively invariant when the same text is read at different distance, even though the letter spaces subtend different visual angles [67]. Although most of the saccades in reading English are made from left to right, about 10-15% of the saccades are regressions, which are right-to-left eye movements along the line of the reading contents. The length of the regressions is important information to tell the reader's reading behaviors and mental states. Specifically speaking, the short regressive saccades may be necessary for preceding the reading efficiently after too long saccades. The longer regressions usually occur when the readers have difficulty understanding the text. In such cases, good readers are found to be able to relocate their eyes to the part of text causing the difficulty, whereas poor readers engage in more backtracking through the text [68]. Moreover, the readers tend to have shorter saccades and more regressions when the text gets more difficult [83].

During fixations, the eye stays relatively steady for a period ranging from 60 to 500ms [59]. Although there is a large variation in the duration of individual fixations, fixations tend to focus on long content words rather than short function words [40]. Previous research shows that the frequency and length of the word can also affect the duration of the fixation on the word, with the gaze duration on longer or low frequency words being lengthier than that on shorter and high frequency words [81]. Readers who are experiencing difficulty in processing the visual information tend to make more fixations and the fixation duration becomes longer [47][84][82]. Moreover, under these circumstances, readers often make backwards or regressive saccades, to re-read the materials and get a better comprehension of the text [50].

There has been much research on reading behavior and associated eye gaze behaviors [81][59][47][84][82]. Studies have shown that eye movement and eye behavior during reading is closely related to human comprehension and attention [81][88][89]. Some efforts were make on detecting human mental state during reading by eye tracking. Reichle et al. [88] used eye tracker to monitor the gaze location of four

subjects to detect mindless reading and investigate the relation between the reading behavior and comprehension. The reading behaviors are measured mainly via fixations. Although they demonstrated the feasibility of detecting mindless reading from some fixation features, they had not investigated the performance of their method in real-use situations. Rodrigue et al. [89] studied the attention level during reading using EEG and eye tracking. They worked on a small number of only three subjects and performed a 10×10-fold-cross-validation with four selected eye gaze features for classification. They achieved an average classification accuracy lower than 70% in a well-controlled setting with the use of eye gaze features, and their method is not evaluated on unseen new users.

In addition to eye movements, eye blinks have also been studied in conjunction with human cognition. There are three main types of eye blink: reflex blinks, voluntary blinks and endogenous blinks [101]. Reflex blinks are caused by foreign bodies invading the eye, and voluntary blinks result from a conscious decision. Endogenous blinks are usually triggered by some aspects of information processing, and it has been shown that endogenous blinks occurring during reading and speaking reflect changes of attention and changes in thought processes [102]. Prior research has made use of the eye blinks as an indicator for fatigue detection. Divajak et al. [25] used eye dynamics and blinks to estimate human fatigue in computer use. They reported that primary eye fatigue indicators include the frequency and duration of blinks as well as the speed of eye closure. Dynamic Bayesian network have also been used to relate fatigue with eye movement patterns such as fixation occurrence and fixation saccade ratio, as well as facial expressions and head movements [48].

2.1.2 Gaze Estimation and Eye Tracking

The release of modern commercial eye trackers facilities precise eye gaze estimation and eye tracking in many research domains. Among the many kinds of eye trackers, optical eye trackers are favored for being non-intrusive and inexpensive. In optical eye trackers, the light, typically infrared, is reflected from the eye, and sensed by a video camera or some other specially designed optical sensors. Despite the convenience of using eye trackers, gaze estimation and eye tracking by using the webcam alone attracts much interest and achieves great success in the past few years.

Gaze estimation methods can mainly be categorized into feature-based and appearance-based methods. Feature-based methods infer the gaze point by extracting

local features from the eye regions, such as the location of eye corners and the eye contours. Appearance-based methods (e.g. [60][61]) detect and track eyes directly based on the image of the eye regions without explicitly extracting the eye features. Current gaze estimation methods are mostly feature-based methods. The two types of existing features-based methods are model-based (geometric) (e.g. [43][119]) and the interpolation-based (regression-based) (e.g. [16][66]).

Compared to the eye tracking systems based on stereo vision, light source, or multiple cameras, using single webcam to estimate gaze and track eyes is work with great challenges. Many factors including the variations of eye-camera distance, head pose and glass occlusion should be taken into consideration when building the gaze model. In this case, most of the studies on gaze estimation rely on data-driven methods to learn the mapping from gaze features into gaze point, and these methods generally rely on massive training data. Huang et al. [46] prepared a large dataset called TabletGaze for gaze learning, which contains over 100 thousand images. Zhang et al. [128] constructed a convolution neural network to recognize the gaze angle from the MPIIGaze dataset, which consists of over 200 thousand images. Xu et al. [122] presented TurkerGaze, which collects large-scale gaze data through crowdsourcing. Krafka et al [55] developed a convolutional neutral network (CNN) to model gaze from the face and eye regions of 1.5 million frames. Similarly, Zhang et al. [129] presented a CNN to learn gaze model using only the full face images.

Instead of using real images, Wood et al. [118] leveraged graphical rendering techniques to generate a million realistic eye images for gaze angle estimation. However, results in [55] show that gaze model calibrated by user-specific data can markedly outperforms the user-independent counterpart, meaning that it is still hard to overcome individual differences by simply learning from large-scale data. Lu et al. [62] synthesized user-specific training samples for unseen head poses from images under a certain reference head positions. However, the computer-generated data may not fully coincide with the real user-specific data.

An alternative is to collect user-specific data implicitly through interactions. Sugano et al. [103] applied the saliency map of video frames to estimate the gaze points according to the eye appearances. Wang et al. [113] proposed to minimize the overall difference between the estimated fixations and the saliency map. However, the consistency between image saliency and real gaze locations can be affected by the semantics and complexity of visual stimuli.

The use of mouse and keyboard as cues to gaze has also been investigated. Sugano et al. [104] used the mouse-click locations as ground truth to update the gaze model. Similarly, Papoutsaki et al. [74] presented a browser-based eye tracker that learns from mouse-clicks and mouse movements. To ensure consistency between gazes and interactions, Huang et al. [42] identified the temporal and spatial alignments between keypresses, mouse-clicks, and the gaze signals. These methods achieve impressive performance, but they are all based on the desktop platform. They also require sufficient well-aligned data from user interactions, which can be slow to acquire, which therefore limits their applicability.

2.1.3 Identifying Fixation from Eye-Tracking Data

Identifying different eye movement behaviors from the eye-tracking data is critical in gaze-based research and applications. This process mainly involves identifying fixation and saccade. Fixation identification is considered as a convenient method of minimizing the complexity of the eye-tracking data while retaining its most essential characteristics for the purposes of understanding cognitive and visual processing behavior. Salvucci et al. [90] classified the existing fixation identification algorithms can be into five categories with respect to their spatial and temporal characteristics. According to the spatial characteristics, there are three primary types of algorithms: velocity-based, dispersion-based and area-based.

Velocity-based algorithms emphasize the velocity information in the eye-tracking data, taking advantage of the fact that fixation points have low velocities and saccade points have high velocities. It is notable that with a constant sampling rate, velocities are simply distances between sampled points and thus we can ignore the temporal component implicit in velocities. Velocity-threshold fixation identification (I-VT) is a velocity-based method that the simplest to understand and implement. I-VT begins by calculating point-to-point velocities for each point in the eye-tracking data. Each velocity is computed as the distance between the current point and the next (or previous) point. I-VT then classifies each point as a fixation or saccade point based on a simple velocity threshold: if the point's velocity is below threshold, it becomes a fixation point otherwise it becomes a saccade point.

Dispersion-based algorithms emphasize the dispersion (i.e., spread distance) of fixation points, under the assumption that fixation points generally occur near one another. Dispersion-threshold identification (I-DT) is one of the most widely used

dispersion-based method. It utilizes the fact that fixation points tend to cluster closely together because of their low velocity. I-DT identifies fixations as groups of consecutive points within a particular dispersion, or maximum separation [117]. Dispersion-based identification techniques often incorporate a minimum duration threshold of 100-200ms to help alleviate equipment variability. The I-DT algorithm uses a moving window that spans consecutive data points checking for potential fixations. The moving window begins at the start of the eye-tracking data and initially spans a minimum number of points, determined by the given duration threshold and sampling frequency. I-DT then checks the dispersion of the points in the window by summing the differences between the points' maximum and minimum x and y values. If the dispersion is below the dispersion threshold, the window represents a fixation. In this case, the window is expanded (to the right) until the window's dispersion is above threshold. The final window is registered as a fixation at the centroid of the window points with the given onset time and duration. This process continues with window moving to the right until the end of the eye-tracking data is reached. It is notable that I-DT algorithm requires two parameters: the dispersion threshold and duration threshold. Sometimes, it is necessary to do experimental analysis and consider the task processing demands to determine the parameters.

Area-based algorithms identify points within given areas of interest (AOIs) that represent relevant visual targets. In contrast to the velocity-based and dispersion-based methods, area-of-interest fixation identification (I-AOI) identifies only fixations that occur within specified target areas. The target areas are rectangular regions of interest that represent units of information in the visual field. These target areas, generally used in later analyses like tracing, keep identified fixations close to relevant targets. I-AOI also utilizes a duration threshold to help distinguish fixations in target areas from passing saccades in those areas.

If considering the temporal characteristics of whether the algorithm uses duration information, and whether the algorithm is locally adaptive. We can find that I-VT doesn’t have these temporal characteristics. In contrast, I-DT have both the temporal characteristics and I-AOI is only duration sensitive. The comparison of the three kinds of algorithms shows that velocity-based and dispersion-based algorithms both fare well and provide approximately equivalent performance. However, area-based algorithms are too restrictive and can generate deceptive results that bias later analyses. Second,

the use of temporal information can greatly facilitate fixation identification of eye-tracking data.

2.2 Affect Detection for Reading

Recent advances in miniature hardware have accelerated human-computer interaction research, in enabling the computer to interact better with human. Affective computing research [19][77] had gained tremendous momentum in recent years, demanding computers to understand human affects or emotions and to react accordingly in enhancing user experience. In order to recognize human affects, input signals reflecting human affects need to be acquired and processed. Under traditional KVM (keyboard-video-mouse) settings, input signals are mostly tied to keyboard and mouse dynamics. One can deduce some information about human affect from the keyboard [12][111] and the mouse [110][123], but the accuracy is not particularly high.

Webcam has become a de facto device thanks to the popularity of interactive social networking applications. A human can oftentimes deduce the emotion of a person sitting in front of a webcam to a certain degree of accuracy. Recent research in video processing and machine learning has demonstrated that human affects and mental states can be recognized via webcam video, noticeably via human facial features [127] and eye gaze behaviors [44]. Though there has been work on mind detection based on facial features and body gestures, research on mental state detection in reading is still limited in the aspects of feature recognition. There is also much work on reading behavior and the associated eye gaze behaviors [47][59][81]. Studies have shown that eye movement and eye behavior during reading is closely related to human comprehension and attention [81][88][89]. However, there are three main drawbacks in current state-of-the-art research works. First, many of them used intrusive devices, like the electrooculography systems, to track the eye movement, or detect the user's mental state as ground truth, through the use of electroencephalography (EEG) devices. Second, numerous methods studied how lexical and linguistic variables affect the eye gaze behavior during reading instead of performing a thorough analysis on the eye gaze pattern for a more ubiquitous and efficient affect or mental state detection. They need to rely on linguistic analysis of the materials being read by the human. Third, some other work designed user-dependent models for the affect or mental state detection which might not be able to accommodate unseen new users in practical applications,

since it is often not practical to ask a new user to strain up the model before actually using it. We believe that reading tasks form a major category of computer usage for many users, especially for laymen and students, to warrant more systematic investigation.

In human computer interaction research, one would often exploit the expressive power resulted from multimodal interaction [69], in which the intention of a user is jointly specified by a plurality of input interaction modalities or signals representing the user. It could be effective in combining and fusing input signals acquired from the keyboard, the mouse and the webcam. In this thesis, we investigate the detection of human attention level when users are carrying out reading tasks based on a multimodal approach with ubiquitous hardware, namely, the webcam and the mouse, without relying on sophisticated devices such as head-mount devices, electrocardiogram devices or heart-beat belts for additional modalities. The webcam is capable of returning a stream of video frames, which is analyzed for eye gaze behavior recognition, face recognition and then temporal change in facial expression. The mouse is capturing its movement and clicking events, indirectly modeling the user activities of moving down a page for reading. For simplicity, we do not consider keyboard dynamics, since users in general do not utilize the keyboard in reading tasks.

The eye is found to play an important role in reading, which is also proven in our experiments. Human cognition detection in reading has become an important research topic since reading is not only a remarkable human skill but also a good sample case to study the working of internal processes of the human mind and the external stimuli on the generation of complex human actions. However, human can get distracted when reading [1], for instance, by Instant Messaging [29]. It is therefore important to recognize the human attention level when formulating feedback for interactive applications to enhance user experience. Human reading cognition detection can contribute in applications such as e-Learning by predicting the readers' mental state through their external behavior and brain activity during the reading process.

The major source of inputs that can closely reflect human reading cognition rests with video streams, often captured via the webcam. Facial expression analysis based on webcam video stream has been applied to analyze cognitive states, psychological states, social behaviors, and social signals [23]. Most recent research on facial expression analysis has been focused on basic emotions, or prototypic emotions, including happiness, sadness, surprise, anger, disgust and fear [26]. These have also

been extended to the recognition of fatigue [48], embarrassment [48], and pain [5]. Cognitive states, like agreeing, disagreeing, interested, thinking, concentrating, unsure, and adult attachment have been investigated [126]. Human mental states can be recognized effectively from webcam video [41]. Facial expression recognition, being a powerful technique, also finds its application in understanding the student engagement in a classroom [3]. Cognitive engagement is found to have close relationship with a person's cognitive abilities, including focused attention, memory, and creative thinking in learning [4].

Human behaviors are often better reflected by human-oriented signals. In reading tasks, the eye is the essential sensory organ involved, besides the brain. In particular, the eye is known to play an important role in reading by researchers, in addition to the obvious electrocardiogram signal, oriented from the human brain. In general, eye movements during reading can be categorized into saccades and fixations, which alternately occur during reading [81]. A saccade is a fast movement of the eye, which is usually in a direction parallel to that of the text. A fixation is the maintaining of the visual eye gaze on a single location. The purpose of a saccade is to locate a point of interest on which to focus, while processing of visual information takes place during fixations. Previous work has found that fixations tend to focus on long content words rather than short function words [40]. The frequency and length of the word can also affect the duration of the fixation on the word, with the gaze duration on longer or low frequency words being lengthier than that on shorter and high frequency words. In addition to eye movements, eye blinks have also been studied in conjunction with human cognition. Prior research has made use of the eye blinks as an indicator for fatigue detection. Divajak et al. [25] used eye dynamics and blinks to estimate human fatigue in computer use. They reported that primary eye fatigue indicators include the frequency and duration of blinks as well as the speed of eye closure. Techniques have been developed to accurately capture eye gaze behaviors from webcam videos [43] rather than relying on the use of proprietary external devices, such as Tobii [114]. It is also possible to derive human mental states from eye gaze behaviors, such as stress level [44].

Despite the simplicity of the mouse in tracking movement and clicking events, it has been found to deliver interesting signals indicating user anxiety [123], or for stress detection [111]. It is in general useful for e-Learning environments [110]. Like mouse dynamics, keystroke dynamics has been studied to correlate human behavior [12].

Keystroke dynamics is particular useful in the analysis of writing tasks which rely primarily on keyboard activities. Reading tasks are more challenging, since the keyboard is often not well-utilized, and the mouse is only used to a limited extent.

The area of multimodal interaction research was pioneered by the seminal "Put-That-There" system [14], augmenting video for location recognition and audio for command recognition. Multimodal interfaces process two or more combined user input modes, for instance, speech and gesture, in a coordinated manner with multimedia system output, aiming to recognize naturally occurring forms of human language and behavior [69]. Human-smart environment can be built based on combined modalities of deictic gestures, symbolic gestures and voice [20].

There have been much research works on understanding the relationship between human cognitive states such as attention and comprehension and human behaviors during reading [78][87]. Efforts have been made into the detection of human mental state from different modalities in the non-instructive manners. Huang et al. [41] inferred interest and boredom from facial expression. Li et al. [58] detected reading attention from mouse dynamics and facial expression. Although facial expression is an important channel for mental state detection, the privacy issue concerned with capturing the human face still poses certain limitation on its applicability in the real world. The mouse is a very common tool to be used as a movement tracking and event selection device. Researchers have found that it is able to encode interesting signals indicating user anxiety [123] or stress [111]. The mouse log can in general provide useful information for e-Learning environments [107]. Huang et al. [42] exploited the alignment between mouse-click and eye movement for gaze learning. Similar pattern describing the coordination between mouse and gaze has also been investigated for mental stress detection [44]. These studies were basically done in specific contexts where the mouse is used frequently as a key input device. However, using the mouse log in the context of reading tasks would be challenging, since the mouse is oftentimes used to quite a limited extent, primarily in the scrolling bar to move around in the article.

Chapter 3

Reading Comprehension Detection from Eye Gaze Behaviors

Affective computing has become an important area in human-computer interaction research. Techniques have been developed to enable computers to understand human affects or emotions, in order to predict human intention more precisely and provide better service to users.

In this chapter, we investigate the detection of the level of reading comprehension as a useful form of human affect, which could be useful in intelligent e-Learning applications. Specifically, we focus on the eye gaze behaviors, in the form of eye gaze signal captured by a commodity eye tracker (Tobii eye tracker). We invite human subjects to carry out experiments in reading articles of different difficulties to induce different levels of comprehension. Machine learning techniques are applied to identify useful features to recognize when the readers are experiencing difficulties in understanding their reading material, leading to a lower level of comprehension.

A user-independent model that is able to identify different levels of user comprehension is built. We find that our approach is able to achieve a performance improvement of over 30% above baseline, translating to more than 50% reduction in detection error. It is found to be quite robust in catering for new unseen users. Finally, we explore different ways of building the user-independent model, in particular running prediction models with respect to different length of available eye gaze segments for recognition. We further investigate the performance demonstrated by the various models.

3.1 Introduction

Reading is one of the most common computer interaction activities and also one of the most fundamental means of knowledge acquisition. In the same way as a human teacher might observe his/her students to gauge their level of understanding, it is easy to see how a good understanding of reading behavior and successful detection of reading difficulty would be helpful for intelligent e-Learning platforms, to provide help to the user, or to adjust the material to improve the learning effectiveness.

Recent advances in hardware and sensors have enabled new modes of sensing signals from users and made these sensing technologies more accessible to the general public. Kinect and Leap Motion are among these common commodity devices that are able to capture human body gestures and finger gestures for game playing and for the development of interesting human-driven applications. Likewise, commodity eye

tracking sensors have become more readily accessible in recent years. There are games and augmented reality applications built based on those eye trackers. It is not inconceivable to imagine that more and more consumer computing devices of the future may be equipped with sensors of this kind.

There has been much research on reading behavior and the associated eye gaze behaviors [81][59][47][84][82]. Studies have shown that eye movement and eye behavior during reading is closely related to human comprehension and attention [81][88][89]. However, there are three main drawbacks in current state-of-the-art research. First, many of them used intrusive devices to track the eye movement like the electrooculography systems, or detect the user's mental state through the use of electroencephalography (EEG) devices. Second, numerous methods studied how lexical and linguistic variables affected the eye gaze behavior during reading instead of doing a thorough analysis on the eye gaze pattern for a more ubiquitous and efficient affect detection. They need to rely on linguistic analysis of the materials being read by the human. Third, many other works designed user-dependent model for the affect detection which might not be able to accommodate new unseen users in practical applications, since it is often not practical to ask a new user to train up the model before actually using it.

In this chapter, we attempt to address the three limitations mentioned above. We make use of a commodity eye tracker to track the eye movement, without making much intrusion into the reader, in bearing or wearing special devices or sensors. We make use of commonly available English articles without analyzing their content in details to build up the model. Finally, we build user-independent models to cope with new unseen users. In order to build up the model, we invite human subjects to carry out experiments in reading a variety of articles while recording their eye movements with a commodity eye tracker. The difficulty level of the articles is varied so as to induce different levels of understanding or comprehension with the users. The captured eye movement data is analyzed to identify specific eye gaze behaviors, and eye movement features are extracted to describe these behaviors. We then apply machine learning techniques to identify a subset of useful features that are capable of assisting in the determination of human comprehension level. Our goal is to produce a set of resilient user-independent models, which are universal to different users and able to accommodate new unseen users.

Our work demonstrates the feasibility of determining a useful and interesting human affect, namely, reading comprehension level. It could find various applications in the real world. For instance, an affectively-aware learning system could automatically ramp up the difficulty of the reading materials if it senses that the reader is not being adequately challenged enough. Similarly, an affectively-aware e-reader could suggest books for its reader based on the understanding of his/her comprehension levels of similar works based on past reading history and pattern. An e- Learning platform could also change the presentation paradigm of the materials, in much the same way as an attentive teacher adapts to perceived student attentiveness in the classroom by adjusting the teaching method and content delivery.

The rest of this chapter is organized as follows. In Section 3.2, we describe our method of comprehension level detection based on the identification of eye gaze behaviors and the extraction of useful features to describe these behaviors. Section 3.3 describes the experimentation with human subjects carrying out reading tasks, and the data collection process. We then evaluate the effectiveness and accuracy of our models in Section 3.4 under many different situations. Finally, we conclude this chapter briefly in Section 3.5.

3.2 Eye Gaze Behavior Recognition and Feature Extraction

In this section, we analyze the eye gaze behaviors captured by a commercial eye tracker in order to detect the level of comprehension of a user when reading an article. In the subsequent subsections, we will describe the actual feature extraction mechanisms based on the stream of eye movement signals captured by the eye tracker. We also present the mechanism to select the set of useful features and finally the means by which we classify the level of human comprehension when reading an article. The overall processing mechanism is depicted in Figure 3-1.

3.2.1 Identifying Eye Gaze Behaviors

The eye gaze data used in this work is captured by the Tobii X1 Light Eye Tracker, which is marketed for “consumer” use. The eye tracker represents the position of each eye as a timestamped sequence in terms of (x, y) screen coordinates, which are normalized to [0, 1] if the eye gaze detection is effective. Occasionally, the eye tracker

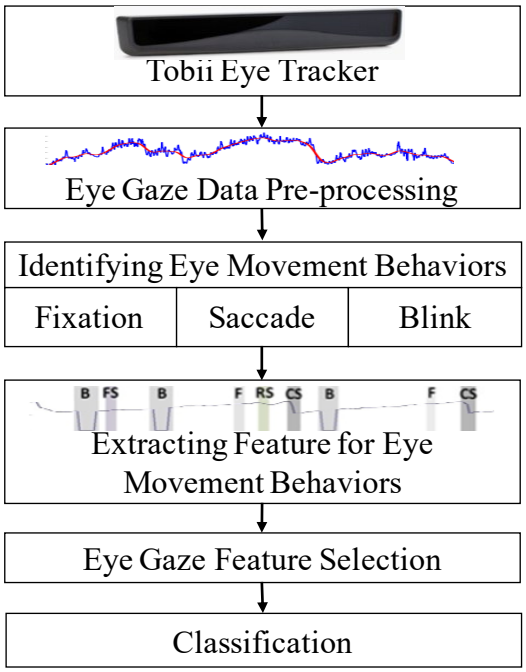


Figure 3-1. System components.

may lose track of the eye, when the person is moving the head rather rapidly sideway or back and fro, or when the person turns his/her head around, or also for eye blinks. The output value for any unrecognized eye position is given a special value of (-1, -1).

There are two challenges in working with the raw eye gaze data. First, the eye tracker occasionally fails to detect one eye, when the eyesight of the reader moves out of the effective recording area. The eye tracker reports a normal pair of screen coordinates for the detected eye, while the position of the other eye is reported as (-1,-1) due to noise. It sometimes fails to detect both eyes, usually when the reader moves his/her head too violently and the positions of both eyes would be reported as (-1,-1). While we may still be able to extrapolate the position of a missing eye, there is a risk that the failure to detect may be due to some other unforeseeable situations. Further, we notice that these erroneous positions take up less than 5% of the data instances. We thus decide to discard them in our study, since we already obtained sufficient amount of data.

Second, the sampling rate of the Tobii X1 light eye tracker is not a constant [107]. This has been reported in previous work and in our pilot study, we observed that the sampling rate varies between 20 to 30Hz. To facilitate the data analysis and pattern recognition processes in the next stage, the eye gaze data is down-sampled to 15 Hz using linear interpolation. A median filter, which has been shown to be effective in preserving the characteristics of the original signal without introducing signal artifacts

[18], is then applied to the down-sampled eye gaze signal to further remove noisy artifacts. The window size of the median filter is set to be 120 ms, which is small enough to retain short pulses indicating eye gaze movements.

After preprocessing, the resultant denoised, subsampled eye gaze data can be represented as a sequence E of *eye position vectors*:

$$E_{\mathrm{i}} = [e_{l_{x_i}}, e_{l_{y_i}}, e_{r_{x_i}}, e_{r_{y_i}}]$$

where $e_{l_{x_i}}, e_{l_{y_i}}$ are the normalized x and y coordinates of the gaze location of the left eye of the i^{th} item in the sequence, as captured by the eye tracker, and $e_{r_{x_i}}, e_{r_{y_i}}$ are the corresponding coordinates for the right eye. For most human, both left and right eyes will be moving together. We therefore would like to simplify the representation by computing the mean value of the coordinates of the left and right eyes. The eye gaze feature vectors can then be simplified into a sequence EG of *eye gaze vectors*:

$$EG_i = [EG_{h_i}, EG_{v_i}]$$

where EG_{h_i} is the horizontal component of the eye gaze location of the i^{th} item in the sequence, defined as the average of $e_{l_{x_i}}$ and $e_{r_{x_i}}$ for the same item, and EG_{v_i} is the corresponding vertical component, defined as the average of $e_{l_{y_i}}$ and $e_{r_{y_i}}$.

3.2.1.1 Detecting eye blinks

Given the eye gaze sequence, EG, eye blinks can be easily detected by identifying the moments in which EG_h and EG_v are both equal to -1. The duration of the blink is defined as the length of the sequence of consecutive eye blink points. In previous research work, an eye blink is defined as the eyelid closure for a duration of 50 to 500 ms [93]. We follow the convention to discard eye closures which are shorter than 50 ms and longer than 500 ms as in relevant research. An inspection shows that eye closures that are shorter than 50 ms are usually due to noise from the eye tracker which occasionally fails to track the position of the eye, and those longer than 500 ms are due to failures from the eye tracker, or momentary human reflection or a taking a break to rest via the closure of both eyes.

3.2.1.2 Detecting eye fixations

Eye fixations are defined to be periods in which the gaze remains stationary on a specific location. However, the inherent error present in the eye tracker makes detecting

fixations from the eye gaze signal more than simply looking for periods during which the eye coordinates do not change.

To determine the extent of the inherent sensor error, a pilot experiment was carried out with 3 subjects (age: 22-28 years, M = 24.7, SD = 2.5). Subjects were asked to sit at a normal distance away from the screen and fix their gaze on a single word displayed in the center of the screen for 10 seconds. The recorded eye gaze signal gives us the error of the eye tracker, which was measured to be around 1% of the potential range of the horizontal component of the eye gaze signal, or $EG_{h_{range}}$. This is then used as the fixation amplitude threshold τ_{FA}. We adopt a method which is similar with I-VT [90] to identify fixations. Since our processed eye gaze data has constant sampling rate, the calculation of point-to-point velocities for each point of the eye gaze data is simply the calculation of distance between the sampled points. Defining D_i as the Euclidean distance between successive gaze locations EG_i and EG_{i+1}, a vector F can then be constructed from EG, where:

$$F_i = \begin{cases} 1, & D_i < \tau_{FA} \\ 0, & D_i \geq \tau_{FA} \end{cases}$$

This gives us a binary vector in which elements with a value of 1 correspond to the moments when the eye gaze could be considered to be stationary. Since it has been found that fixations are rarely less than 100 ms and usually in the range of 200 to 400 ms [90], we label fixations as continuous sequences that last for longer than 100 ms but shorter than 500 ms. Too long a potential fixation may indicate abnormal situation or sometimes erroneous situation. We would like to filter out those relatively uncommon scenarios to clean up the data stream, by removing the outliers that could have affected the statistical measures that we compute as the potential features.

3.2.1.3 Detecting eye saccades

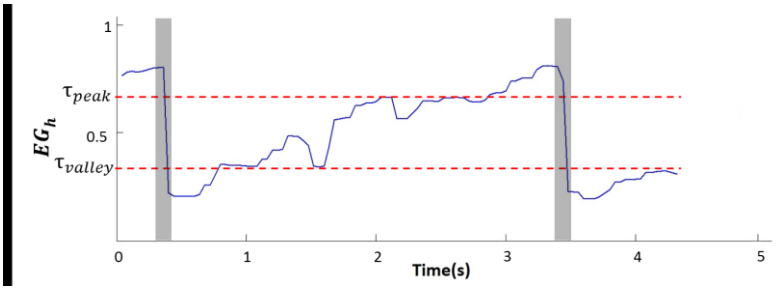


Figure 3-2. Changes in horizontal component of eye gaze.

Once the eye blinks and fixations have been identified, the sequences in between the fixations are considered for saccadic eye movements.

Given the nature of the task, we classify the eye saccades into three categories: line-change saccades, forward saccades and regressive saccades. Figure 3-2 shows an example. Line-change saccades involve large, fast eye movements that coincide with moments when the reader finishes reading one line and the eye jumps to the beginning of the next line. Forward saccades follow the direction of the text (left to right in our case of English text) during normal reading. Regressive saccades, on the other hand, go against the flow of the text when the eye position moves back to re-read previously-read content.

Previous work using intrusive electrooculography (EOG) [18] has found that the horizontal component is sufficient for identifying saccadic behaviors when reading English. Our saccade detection algorithm similarly uses the horizontal component EG_h of the eye gaze position to identify the different saccadic types.

We use a two-level saccade detection algorithm that first identifies line-change saccades. The definition of line-change saccades and typical reading behavior suggests that these saccades can be identified by looking for eye movements that span over a threshold corresponding to the end of a line and the beginning of the next line of text.

Figure 3-2 illustrates the horizontal component of the eye gaze signal EG_h as a function of time. It can be seen that the signal experiences periodic fall-off "cliffs" when the eye gaze switches from the far right of the page (closer to 1) to the left edge (closer to 0). It gives us periodic "peaks" and "valleys" which correspond to line-change saccades.

In order to perform analysis on the signal for line-change saccades, we define two thresholds: τ_{peak}, corresponding to the maximum value of the eye gaze signal before the fall-off "cliff", and τ_{valley}, corresponding to the minimum value of the eye gaze signal right after the fall-off "cliff". Line-change saccades are then defined as eye gaze movements that start at positions greater than (to the right of) τ_{peak} and end with a position smaller than (to the left of) τ_{valley}.

The two thresholds, τ_{peak} and τ_{valley}, are determined empirically through an experiment with 4 subjects (age: 21-30 years, M = 25.3, SD = 3.5). The subjects are asked to read English articles presented on a LCD monitor with a resolution of 1680 x 1050. The gaze data from both eyes is then visualized onto images of the screen display,

which allows us to observe the trajectory of the eye movements. We would like to reduce the likelihood that we mistakenly label regressive saccades as line-change saccades by carefully selecting the thresholds. The best performance of the algorithm (with 100% recall and 100% precision) is achieved when we select $\tau_{peak} = 0.6$ and $\tau_{valley} = 0.3$ and these values would therefore be adopted in our experiments.

The second level of the saccade detection algorithm distinguishes between forward and regressive saccades. Previous research [100] has shown that saccadic eye movements during reading usually span a distance of about 7 to 9 letters, equivalent to about 2 degrees of visual angle, with duration between 10 to 100 ms.

We define the *saccade amplitude* S_{amp} as:

$$S_{amp} = \sum_{EG \in S} |EG_{h_i} - EG_{h_{i-1}}|$$

where S is a saccade, EG are the eye gaze points composing the saccade S, EG_{h_i} and $EG_{h_{i-1}}$ are the horizontal components of temporally successive gaze points.

Given the saccade amplitude, and knowing that normal reading activities generate saccades that span approximately 2 degrees of the visual angle, we can use the arc length formula to calculate the required amplitude that would be expected if this saccade resulted from the subject's reading activity. We therefore define the *saccade amplitude threshold* τ_{SA} as:

$$\tau_{SA} = \frac{\pi d EG_{h_{range}}}{90 W_{article}}$$

where d is the distance from the eyes to the screen, $EG_{h_{range}}$ is the range of EG_h, $W_{article}$ is the width of the article as displayed on the screen.

According to the hardware setup in our experiment, $W_{article} = 31.5\ cm$, and on average, $d = 60\ cm$ and $EG_{h_{range}} = 0.68$. This leads to the setting of $\tau_{SA} = 0.045$. Since a full line of text has on average 73 characters in our setup, a saccade that spans 7 to 9 characters would give us saccades ranging between 0.065 to 0.084. Combining this with the calculated τ_{SA} gives us the experimental parameter settings of $\tau_{SA_{min}} = 0.045,\ \tau_{SA_{max}} = 0.084$.

The use of the amplitude threshold allows us to detect a potential saccade, but not its direction. Given a candidate saccade S with n eye gaze points, we identify the first and last eye gaze points EG_1 and EG_n. The horizontal components allow us to recognize

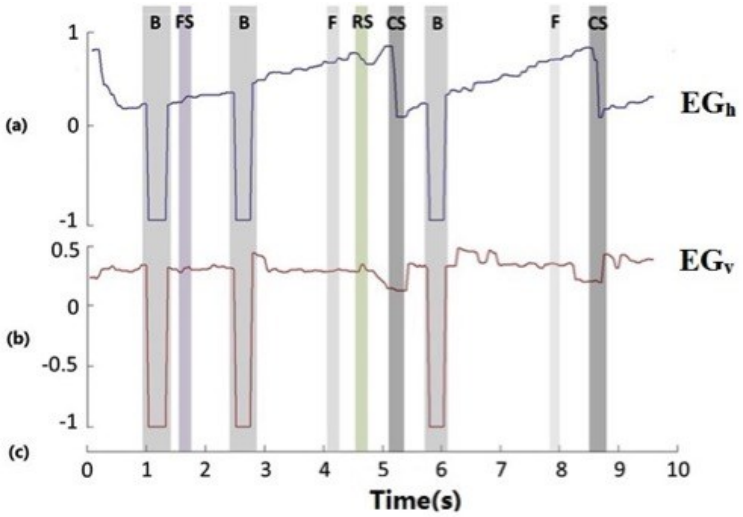


Figure 3-3. Identifying eye movement patterns.

the direction of saccade. Thus, the saccade can be classified as forward or regressive saccade:

$$S = \begin{cases} forward & if\ \ \tau_{SA_{\min}} \le s_{amp} \le \tau_{SA_{\max}} \text{ and } EG_{h_1} < EG_{h_n} \\ regressive & if\ \ \tau_{SA_{\min}} \le s_{amp} \le \tau_{SA_{\max}} \text{ and } EG_{h_1} > EG_{h_n} \\ non-saccadic & otherwise \end{cases}$$

Once again, we would like to clean up the saccadic stream for outliers. Saccadic segments which are shorter than 20 ms and longer than 200 ms are discarded, assuming an average angular saccade speed of 20 degrees/second [95].

Figure 3-3 illustrates an example of identified eye gaze behaviors given the eye gaze positions in both x (horizontal) and y (vertical) dimensions. The eye blinks, fixations, forward, regressive and line-change saccades are illustrated. Given the horizontal component EG_h and vertical components EG_v of the eye gaze positions, we consider changes in both dimensions as a whole. We use the notation $\boldsymbol{F}$ to indicate a fixation, $\boldsymbol{B}$ to indicate an eye blink, $\boldsymbol{FS}$ to mean a forward saccade, $\boldsymbol{RS}$ a regressive saccade, and $\boldsymbol{CS}$ for a line-change saccade. A blink is easily identified for a period of non-eye-detection. A line-change saccade is reflected by backward changes in both x and y dimensions in a sequence of eye gaze points, whereas a regressive saccade is reflected by backward changes in x dimension.

3.2.2 Features Describing Eye Gaze Behaviors

Once the different eye gaze behaviors have been identified from the eye gaze points, we construct the features that will be used to describe these behaviors.

Table 3-1. Features describing saccadic eye behaviors.

Feature	Meaning	Formulation
$s_{1_{FS}}$, $s_{1_{RS}}$, $s_{1_{CS}}$ $s_{2_{FS}}$, $s_{2_{RS}}$, $s_{2_{CS}}$	Distance covered by saccades	Average (s_1) and standard deviation (s_2) of amplitude (screen distance) covered by forward, regressive and line-change saccades
$s_{3_{FS}}$, $s_{3_{RS}}$, $s_{3_{CS}}$ $s_{4_{FS}}$, $s_{4_{RS}}$, $s_{4_{CS}}$	Duration of saccades	Average (s_3) and standard deviation (s_4) of duration of forward, regressive and line-change saccades
$s_{5_{FS}}$, $s_{5_{RS}}$, $s_{5_{CS}}$ $s_{6_{FS}}$, $s_{6_{RS}}$, $s_{6_{CS}}$	Speed of saccades	Average (s_5) and standard deviation (s_6) of speed of forward, regressive and line-change saccades
$s_{7_{FS}}$, $s_{7_{RS}}$, $s_{7_{CS}}$	Rate of saccades	Number of forward, regressive and line-change saccades in window W_{EG}

The objective of our work is to automatically detect when the user is having difficulty with a text, with a reduced level of reading comprehension. Given this ultimate objective, we consider the eye gaze behaviors within a given time period of reading.

For each saccade type, we define 6 different useful features measuring the saccades more precisely. In particular, we measure the mean and standard deviation of the metrics of interest for each saccade type, namely, the amplitude (defined in terms of the screen distance covered in the saccade), the duration and the speed. Given three types of saccades, three interesting metrics and two statistical measures, this gives rise to 18 potentially useful features. Finally, we are also interested in the global manifestation of each type of saccadic behavior throughout the period of the article reading task. We measure the number of saccades of each type over the window W_{EG} that spans over the duration of each individual article reading task. This gives us a final list of 21 features. Table 3-1 depicts those 21 features useful in describing the saccadic eye behavior adopted in building our comprehension level recognition model.

We extract features to describe fixations in a similar manner. These include the rate of fixation normalized by the window W_{EG}, i.e. the time spent in reading the article. We also calculate the mean and standard deviation of the duration of the fixations, as

Table 3-2. Features describing fixations.

Feature	Meaning	Formulation
f_1, f_2	Duration of fixations	Average (f_1) and standard deviation (f_2) of duration of fixations
f_3, f_4	Interval between fixations	Average (f_3) and standard deviation (f_4) of elapsed time between successive fixations
f_5	Rate of fixations	Number of fixations in window W_{EG}

Table 3-3. Features describing eye blinks.

Feature	Meaning	Formulation
b_1, b_2	Duration of eye blinks	Average (b_1) and standard deviation (b_2) of the duration of eye blinks
b_3, b_4	Interval between eye blinks	Average (b_3) and standard deviation (b_4) of elapsed time between successive eye blinks
b_5	Rate of eye blinks	Number of eye blinks in window W_{EG}

Table 3-4. Features describing eye movements.

Feature	Meaning	Formulation
e_1, e_2	Variation of eye movements	Kurtosis (e_1) and skewness (e_2) of horizontal component of eye gaze locations captured by eye tracker

Table 3-5. Contextual features.

Feature	Meaning	Formulation
c_1	Reading speed	Number of lines in article segment, divided by time
c_2	Repetition rate	Number of line-change saccades divided by actual number of lines

well as the interval between successive fixations. Table 3-2 details the 5 fixation-related features used in our work.

We compute 5 eye blink features. They are extracted by calculating the rate of blink normalized by the window W_{EG}, the mean and standard deviation of the duration of the eye blinks, and the interval between successive blinks. Table 3-3 shows the list of eye blink features.

In addition to the behavior-based features, we also adopt gaze-based features to capture the general characteristics of the unfiltered eye gaze positions. These include the kurtosis and skewness of the horizontal component of the EG_h signal, as shown in Table 3-4.

Finally, we incorporate some features that are meant to capture the context of the reading task through combining basic information on the read article and the overall task. These are the per-line reading speed and the repetition rate of reading, as shown in Table 3-5.

3.3 Experimentation and Data Collection

The evaluation for our model for detecting the level of reading comprehension involves reading tasks in a real-world setup. The experiment subjects read the article in a full-screen mode, as depicted in Figure 3-4. Different pre-selected articles of different levels of difficulty are used to induce varying levels of comprehension on the subjects. Note that even though the level of difficulty is roughly related to the level of comprehension, it does vary across different persons. A hard article could be hard to comprehend for someone, but perhaps just medium for another. An easy article may also take some

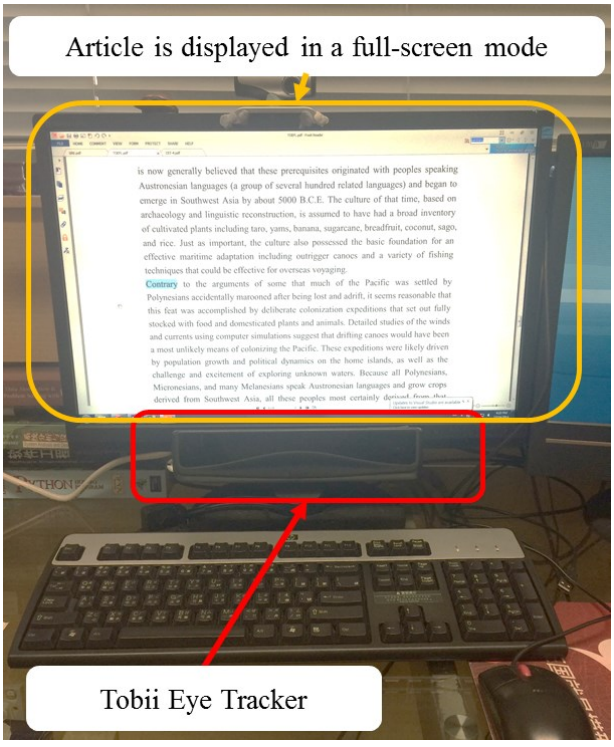


Figure 3-4. Experimental Setup.

weak readers much effort to understand. So we rely on experiment subjects to report their level of comprehension when reading a specific article. The eye gaze tracking logs for individual subjects are recorded in real-time during the experiment, and pre- and post-surveys are used for further data collection about the subject.

3.3.1 Participants and Experiment Setup

We recruit 10 subjects (age 20-33 years, M = 24.6, SD = 4.2) for this experiment. All of the experiment subjects are undergraduate or graduate students, four of them are female. They are all non-native English speakers, and their "mother tongue" is either Mandarin Chinese or Cantonese Chinese. A pre-experiment survey revealed that they are all comfortable with using the computer, and are able to read and write in English, though there is some variation in the level of comfort and their grasp of the language.

The experiment is performed within a standard laboratory environment, as shown in Figure 3-4. The experimental setup consists of a 22-inch flat LCD screen with resolution of 1680 × 1050 pixels for displaying the English article and a commercial Tobii X1 eye tracker mounted under the monitor facing the subject for eye gaze tracking. Paging is done either via the keyboard or the mouse, depending on the user's preference. The subjects are seated about 60 cm away in front of the monitor and are free to move their head or body throughout the experiment, though they are requested to avoid violent body movements.

3.3.2 Experiment Design

In order to induce different levels of reading comprehension in this experiment for all the subjects, we make use of English articles from standardized sources that are widely used worldwide to evaluate English reading comprehension ability. Specifically, we pick articles from the reading comprehension material pools of the GRE (Graduate Record Examination), TOEFL (Test of English as a Foreign Language), and CET-4 (College English Test Band 4[1]). TOEFL is widely used for undergraduate and graduate admissions of non-native speakers to English-speaking colleges and universities. GRE is required by many graduate schools in the US and Canada as an admission criterion. CET-4 is used for calibrating the level of English ability for university students in China, being also a common admission criterion to graduate schools. Two articles were chosen

[1] CET-4 is an English test used to evaluate the English ability of Chinese undergraduates in China. Most participants are sophomores or juniors. All the reading materials used in the reading comprehension of this test are chosen from English publications.

from each of the three tests, and the length of each article was constrained to be around 500 words. To make sure that the subjects were focused on the reading task and attempted to understand the article as much as possible, a post-reading guarding procedure was imposed. Before the reading task began, the subjects were told that they should give a detailed explanation of each paragraph of the article to the experiment instructor after finishing reading each article. Moreover, to guide the subjects to do a precise report of the level of comprehension for each article and to minimize the potential inconsistency among subjects, we showed them the self-report guidelines with our required levels of reading comprehension before the experiment. The guidelines provide them with criteria for their judgement with which to determine the level. After reading all the articles, the subjects were presented a survey to write down some feedback about the setting of the experiment and provide further suggestion.

Considering the background and English level of our subjects, we expect them to have a reasonable understanding of the TOEFL articles, while the GRE would be considered to be difficult, and the CET-4 should pose no difficulty to them. Post-experiment surveys confirmed that the subjects indeed did find that the articles were of different levels of difficulty and they did experience varying levels of comprehension. In general, our experimental subjects also found the setup of the experiment to be comfortable and non-intrusive.

The subjects were informed in advance about the eye gaze tracker and that their gaze movements were being recorded during the experiment. To minimize the impact of the ordering effects commonly occurring in HCI experiments, the order of the articles was randomized so each subject would be presented with the articles at different levels of difficulties following one of the different permutations. The subjects were not constrained for reading time, but they were asked to read the article thoroughly so as to fully comprehend it, as much as possible. Immediately after reading each article, the subjects were required to label their perceived level of comprehension as “low”, “medium” or “high” while their memory was still fresh.

3.3.3 The Dataset

After the experiment, the collected eye gaze data is visually inspected to ensure that it is usable. In a few cases, large degree of body movements and askew sitting postures from the subject results in eye tracker failure for prolonged periods of time. Reading activity segments that exhibit this kind of phenomena are removed from the dataset.

The final dataset contains 41 instances, corresponding to 41 article-reading activities. The total length of the dataset is 228 minutes. According to the subjects' self-report labeling, of the 41 instances, 9 (22.0%) instances were labeled as "high level of comprehension", 17 (41.4%) as "medium level of comprehension", and 15 (36.6%) as "low level of comprehension". Even though the "medium" class is slightly dominating the other two, we still consider this to be a reasonably well-mixed dataset since the deviation from the completely even distribution (33.3%) is not too far.

The baseline of the dataset is 41.4% since the bottom line for random guessing is to output the label of the largest class to achieve the "best" result. Various user-independent models for comprehension level detection are built based on this dataset and evaluated in the following sections.

3.4 Feature Selection and Model Evaluation

In this section, we evaluate our reading comprehension detection system by building user-independent models based on the dataset, as well as various partitioning of the dataset based on different temporal segmentations on the reading data. In classification research, a user-independent model is usually not as accurate as a user-dependent model, especially for classifying phenomena that are heavily dependent upon individual characteristics of the subject. However, their appeal is that they are more applicable in practice, being universally applicable to all users. Once trained, a user-independent model can be used for any new user, whereas its user-dependent counterpart, in contrast, would require new data to be collected and the model to be re-trained for each individual new subject. This places a heavier burden upon the deployability of such systems, especially in real-usage scenarios. Therefore, in this section, we build up different user-independent models based on different experimental settings and evaluate our approach for recognition performance.

3.4.1 Evaluation on Reading the Whole Article

3.4.1.1 Feature evaluation and selection

Based on the process described in Section 3.2.2, 35 eye gaze features are extracted from the reading activity data. These include 21 saccade features, 5 fixation features, 5 eye blink features, 2 eye movement features, and 2 contextual features. The duration of the window W_{EG} is set to be the duration of reading the whole article.

Our initial set of features contains 35 features, which are likely too many to be effective for practical real-time recognition. Therefore, after extracting the set of potentially useful features, feature selection is conducted to remove non-indicative features to improve classification performance.

We adopt the wrapper method for the feature selection, which is reported to outperform the filter method [106]. The wrapper method considers the selection of a set of features by comparing different feature combinations to identify a subset that is best for the chosen classifier. Unlike the filter method, the wrapper method is able to take into consideration interactions between different features rather than ranking each of the features on their own. We make use of the Linear Support Vector Machine (SVM) for classification and its error rate to indicate the performance of a feature subset. We adopt the best first searching approach which searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility, with a comparison window of 5 consecutive non-improving search nodes to determine the set of potential features. This strikes a balance between computational efficiency and effectiveness of features selected.

We analyze how the feature evaluator ranks all the features by conducting a 10-fold-cross-validation evaluation. The number of training sets that each of the features is selected in the cross-validation for classification is used to measure the degree of importance of the features.

Table 3-6 shows the top 10 most indicative features, together with the number of partitions (in 10-fold cross-validation) in which they are found to be of potential contribution in recognition. Unsurprisingly, the most indicative features are the reading speed and the rate of the forward saccades. This makes sense, as people who are having difficulties in understanding the reading material will tend to slow down and have longer fixation [81]. They also tend to make more regressive saccades and repeatedly read sections of the text more often, which is also evidenced by the fact that the repetition rate is selected as one of the most indicative features, and regressive saccades feature heavily dominates among the top 10 most useful features as well.

We note that some of the top 10 ranked features are actually indicative only for a small number of training partitions instead of being useful across the board and they may not always work synergistically together. Some of them may still be useful when combined with other features.

Table 3-6. Top 10 indicative features.

Feature	Description	Number of indicative training sets (10-fold cross-validation)
c_1	Per-line reading speed	10
$s_{7_{FS}}$	Rate of forward saccades	8
e_2	Skewness of eye movements	4
$s_{3_{RS}}$	Average duration of regressive saccades	3
e_1	Kurtosis of eye movements	3
c_2	Repetition rate	2
f_5	Rate of fixations	2
$s_{3_{FS}}$	Average duration of forward saccades	2
$s_{6_{RS}}$	Standard deviation of duration of regressive saccades	2
f_4	Standard deviation of elapsed time between fixations	2

Table 3-7. Final set of selected features.

Feature	Description
c_1	Reading speed
$s_{7_{FS}}$	Rate of forward saccades
e_2	Skewness of eye movements
c_2	Repetition rate
f_5	Rate of fixations
$s_{3_{FS}}$	Average duration of forward saccades

To select the best feature subset, we run the feature selection algorithm with the same feature evaluator to yield a subset which would be the most effective. Having trimmed the set of potentially useful features from 35 to 10, we explore the different combinations of subsets of features drawn from the list of 10 potentially useful features. Again, we make use of SVM for classification and adopt 10-fold cross-validation. Finally, there are 6 important features that together give us the best performance. This list of features selected in building the user-independent model is illustrated in Table 3-7.

Table 3-8. Confusion matrix for comprehension detection.

Comprehension level / Ground truth	High	Medium	Low
High	7	2	0
Medium	3	11	3
Low	0	3	12

Table 3-9. Leave-one-subject-out comprehension detection.

Performance / Comprehension level	Precision	Recall	F-measure
High	0.70	0.78	0.74
Medium	0.69	0.65	0.67
Low	0.80	0.80	0.80
Overall	0.73	0.73	0.73

Table 3-10. Normalized confusion matrix.

Comprehension level / Ground truth	High	Medium	Low
High	0.78	0.22	0
Medium	0.18	0.64	0.18
Low	0	0.20	0.80

3.4.1.2 User-independent model

We would like to evaluate the performance of the user-independent model built. The gold standard in evaluating the effectiveness of user-independent models is the leave-one-subject-out cross-validation test. From the set of $n = 10$ subjects, we train the user-independent model using the data from $n - 1 = 9$ subjects and test the model on the left-out subject. We repeat the experiment n times, each time leaving out a different subject, and the average performance for the 10 experiments is reported.

Table 3-8 presents the confusion matrix for the leave-one-subject-out comprehension level detection experiment on the 10 subjects and 41 instances. Table 3-9 illustrates the classification performance of the classifier, averaged over 10 subjects.

Table 3-11. All-subject-included comprehension detection.

Performance / Comprehension level	Precision	Recall	F-measure
High	0.69	1.00	0.82
Medium	0.82	0.53	0.64
Low	0.76	0.87	0.81
Overall	0.77	0.76	0.74

Table 3-12. Normalized confusion matrix.

Comprehension level / Ground truth	High	Medium	Low
High	1.00	0	0
Medium	0.23	0.53	0.24
Low	0	0.13	0.87

Table 3-10 provides further details with the confusion matrix normalized by the ground truth, giving us a better picture of the relative performance.

From Table 3-8, we can compute the correct classification rate (CCR) as defined to be the proportion of the correctly classified instances over all the instances. This CCR is found to be 73.2%, which is significantly higher than the baseline of 41.4%**,** with an improvement of 31.8%, achieving 54.2% error reduction. Table 3-9 indicates that we are able to perform well with an overall F-measure of 0.73. The performance of each individual class indicates that we are able to achieve a high precision as well as a high recall, without having to sacrifice one metrics for the other. In fact, we can achieve a precision and a recall up to 0.8 and even the worst recall stands at 0.65, far much better than the baseline performance of 41.4%. We are able to recognize very well low level of comprehension for probably difficult articles, and comparatively not that well for medium ones. This is perhaps the low level of comprehension is often associated with lengthy reading with many regressive saccades, making it easier to detect. For the medium one, it is somewhere in between, making it more prone to being misclassified. This becomes more evidenced when we normalize Table 3-8 to produce Table 3-10, which shows that most of the errors come from misclassifying relatively similar comprehension classes, i.e. conflating *low* and *medium* and *medium* with *high*. There

are no errors in which the classifier erroneously classifies an instance into an extreme class, i.e. confusing *low* with *high* and vice versa.

According to our study, most of the eye gaze behaviors and their corresponding level of reading comprehension are intuitive and follow common sense. For example, compared with "high" level comprehension, when the comprehension level is "low", the subjects tend to read more slowly, have more fixations, and have more backward saccades. These behaviors suggest that the subjects spend more time processing the information during reading and often need to read repeatedly to have a better understanding of the reading contents. These behaviors can be reflected from the features of c_1 (Reading speed), f_5 (Rate of fixations), and c_2 (Repetition rate). We also find the value of $s_{3_{FS}}$ (Average duration of forward saccades) becomes smaller and $s_{7_{FS}}$ (Rate of forward saccades) increases when the comprehension level is "low". These phenomena are consistent with the changes of the eye fixation behaviors.

It is notable that among the three groups of eye gaze features, eyep blink features neither appear among the top ranked features nor are they selected for the classification. This is despite the fact that blink frequency and duration have been widely reported to be useful for human affect or behavior detection, such as visual engagement measurement [70], fatigue detection [25], activity recognition [18], etc. This would suggest that there is no obvious relationship between eye blinks and the human comprehension of reading material, at least as perceived by the user. There has been research demonstrating that blink rate is unreliable as a measure of the difficulty of the reading task when the difficulty is varied by introducing glare conditions or auditory distractions [11][10]. Our work seems to further confirm this conclusion by demonstrating that eye blinks are also not particularly helpful when the difficulty of the reading task is varied by the reading materials.

To have a better understanding of the influence of the used features to the performance of our method, we further evaluate our method in two other cases. First, we use all the 35 features presented in Table 3-1 to Table 3-5 in Section 3.2.2 to build a user-independent model for the reading comprehension detection. By running a leave-one-subject-out cross-validation test, a CCR of 56.1% is achieved. This performance is higher than the baseline (41.4%) by 14.7%, but lower than the performance of the user-independent model built on the selected features (as presented in Table 3-7) by 17.1%. There may be two reasons for this result. First, some of the 35 features are not helpful

for the classification, and even do harm to it because of the ambiguous patterns across the data. For example, we find the value of the feature b_5 (Rate of eye blinks) is not consistent with the level of comprehension at all across the data. Instances labeled as "high" level and "low" level comprehension both have b_5 with big value. According to our observation, the big b_5 value can be cause by both the eye fatigue during reading and the "low" level reading comprehension. As a result, the different physical conditions of the subjects make b_5 a useless feature for the comprehension level detection in reading. Second, 35 features are too many to use in our study considering our small dataset, which contains only 41 instances. It is easy to cause overfitting problem in the machine learning process.

Second, instead of selecting features based on the whole dataset as we did in previous studies, we evaluate our method by doing feature selection only on the training set and testing on the test set. We do a leave-one-instance-out cross-validation to evaluate our method. In each round of the evaluation, one of the 41 instances is left out for testing and the rest of the 40 instances are used for training. We adopt the wrapper method for the feature selection on the training set by conducting a 10-fold-cross-validation evaluation. The selected set of features are used to build the model and tested on the left-out instance. We finally get an average CCR of 68.3%, which is higher than the baseline by 26.9%. Although the performance slightly drops by 4.9% from that of the user-independent model built on the features in Table 3-7, it shows the robustness of our model built with a harsher machine learning process and in a more real-use situation. By analyzing the selected features, we find c_1 (Reading speed) and f_5 (Rate of fixations) are the most frequently selected features, which is not surprising.

According to the performance of our method with different evaluation approaches, we further improve the importance of doing feature selection for the comprehension level detection in reading.

3.4.1.3 Absence of new unseen users

In our evaluation, we assume the setting of leave-one-subject-out for recognition performance to cater for new unseen users. However, it is also common in real-use for the model to be used by one or more dedicated users. One would expect that the accuracy in those scenarios to be higher. In this evaluation, we keep all subjects in the 10-fold cross-validation and compare the performance with the leave-one-subject-out

Table 3-13. CCR improvement for existing users.

Evaluation method	CCR	F-measure
Leave-one-subject-out	73.2%	0.73
All-subjects-included	75.6%	0.74

setting. We repeat the experiment in Section 3.4.1.1 by including all subjects together. The performance is depicted in Table 3-11 and Table 3-12.

We can observe from the two tables that there is a bit of improvement across the board for all the metrics. Table 3-13 summarizes the key performance metrics between the two sets of experiments and two types of models. Despite the small improvement observed, it actually is far more encouraging. It does demonstrate that our approach is very robust, delivering good performance even for new unseen users based on training data from a small number of subjects ($n - 1 = 9$). Thus, our research work represents a good initial attempt to reading comprehension detection based on the eye movement patterns captured by a commodity eye tracker.

3.4.2 Evaluation on Incremental Length of Segments

Our previous evaluation demonstrates the effectiveness of our method in recognizing user comprehension level from the gaze behaviors when they read an article. However, in real-use situations, it is more interesting and useful that a prediction can be made well before the user finishes reading the entire article. This is useful in providing real-time feedback to an HCI or e-Learning application in order to tailor its presentation or interaction to the user comprehension level. In such cases, only the beginning segment of the reading eye gaze data is available. We therefore conduct a performance evaluation based on various incremental lengths of available eye gaze data. This evaluation can shed lights on the confidence levels of our prediction with respect to the length of available data segment.

Each of the 41 instances of reading a whole article is of different lengths, some subjects take as little as 2 minutes in reading a possibly easy article, but some subjects take as much as 14 minutes in reading a difficult GRE article. In this experiment, we will progressively perform recognition on the comprehension level based on various lengths of eye gaze data. Figure 3-5 shows the length information of our instances. The x-axis shows the range of the length, and the y-axis indicates the number of instances that satisfies the corresponding length condition. For example, all of the 41 instances

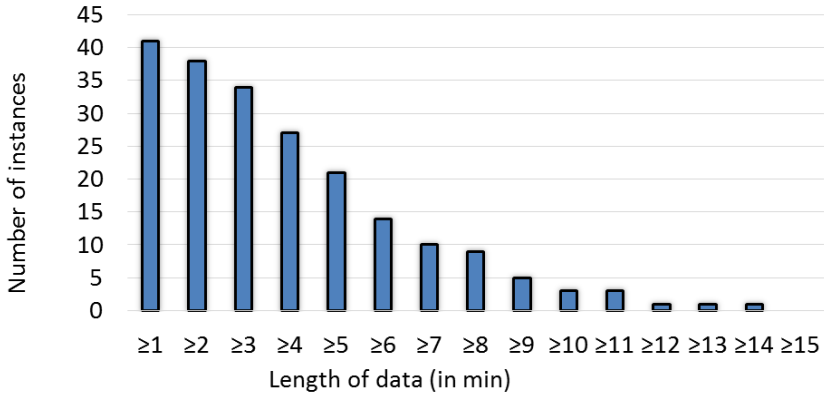


Figure 3-5. Distribution of the length of our dataset.

are longer than 1 minute and only 1 instance is longer than 14 minutes. It can be seen that around half of our instances are over 5 minutes, with a median length of 5.22 minutes.

Since our features reflect the temporal attributes of eye gaze behaviors, we foresee that these features may be sensitive to the length of eye gaze data available. The implication of the features on the performance can vary as their granularity. For instances, the fixation duration in a half minute segment may indicate the time spent on each word, or the reading speed. However, the fixation duration of reading the entire article can provide a cue to the total time of being distracted. The number of line-change saccades in short segments would also be small.

In this study, we perform the leave-one-subject-out evaluation as usual. We first build up a user-independent model for each left-out subject. Instead of performing recognition on the instances of the left-out subject, we perform recognition based on prefix segments, i.e. segments starting from the beginning of the reading tasks, of various lengths drawn from the instances of the left-out subject. Thus, all instances will be evaluated based on prefix segments of the first minute. Almost all instances will be evaluated based on prefix segments of the first two minutes and so on. There will be fewer and fewer instances to be evaluated when the segment length grows. The total number of individual testing prefix segments conforms to the frequency depicted in Figure 3-5. Though the testing segments are of varying lengths, each user-independent model built is based on the full-length instances reading the whole article. We thus call this evaluation the "entire article" model approach (**Model**$_{\textbf{article}}$), and its performance is depicted as the green dashed line in Figure 3-6.

Recently, Pasqual et al. [75] proposed a template-matching-based approach in predicting the endpoints of mouse movements based on classifiers with incremental granularity. In their real-time recognition application for the mouse movement endpoints, it is necessary to make a prediction based on partially collected data, without knowing when the data sequence will become complete. They adopted a template based approach with fixed-length training mouse trajectories for matching and achieved reasonable performance. Inspired by this study, we hypothesize that constructing a set of classifiers with the same granularity as the testing segments may be conducive to the recognition of the short instances, since both training and testing instances will exert similar impacts on the features, especially features that are more well-defined for longer instances, such as line-change saccades. We thus explicitly build user-independent models based on prefix segments from training instances of the same length as the testing segments in order to perform the evaluation. As such, our user-independent models built for segments of different lengths resemble templates. We call this evaluation the "segment" model approach (**Model**$_{\text{segment}}$). Its performance is illustrated as the red solid line in Figure 3-6.

We therefore compare two types of models: (1) a model learnt from the data of the entire article, **Model**$_{\text{article}}$ and (2) a model learnt from the data prefix with a corresponding length as the testing segment, **Model**$_{\text{segment}}$. We will perform full cross-validation on all the instances in our dataset for both models. The only difference is that **Model**$_{\text{segment}}$ is trained on the prefix segments according to the length of the testing data in each epoch.

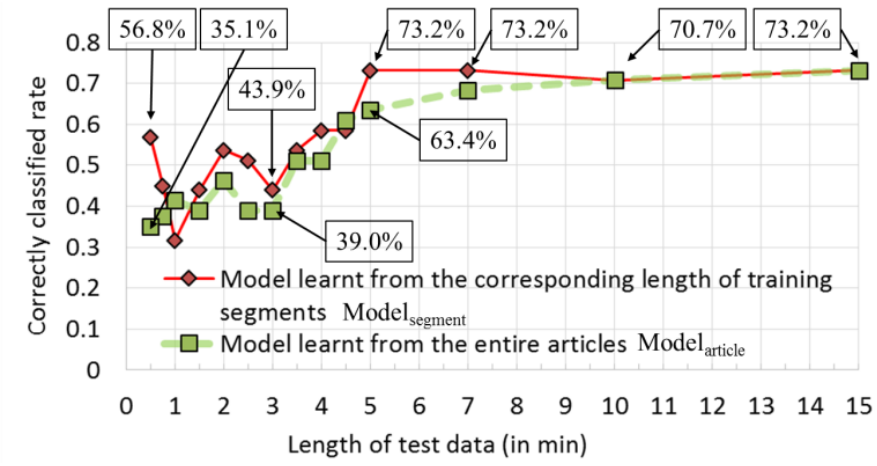


Figure 3-6. Correctly classified rate vs. length of testing prefix segments.

Figure 3-6 shows the performances of **Model**$_{\textbf{article}}$ and **Model**$_{\textbf{segment}}$ on the incremental length of the testing prefix segments, i.e. the beginning stage of reading each article. As predictions in the early stage of reading are of particular interest to us, we evaluate with a finer granularity on the length of the prefix segments in intervals of every half minute. Since the number of testing segments decreases when the length increases, the granularity of the length for testing is reduced beyond 5 minutes. Both models adopt the identical feature set as presented in Table 3-7, and the CCR is calculated from the leave-one-subject-out cross-validation for user-independent models. Key performance CCR metrics are given in numerical values for more accurate comparison. From the figure, we can observe that the performances of both models fluctuate for the short testing segments and gradually flat out given sufficient amount of data.

Although fluctuations exist with short length data (1~3 minutes), we can generally conclude that the performance of **Model**$_{\textbf{article}}$ rises as the length of testing data increases, whereas **Model**$_{\textbf{segment}}$ suffers from a higher variability. It is encouraging that given the testing segments with length longer than 5 minutes, both **Model**$_{\textbf{article}}$ and **Model**$_{\textbf{segment}}$ perform rather stably and they fully converge when testing segments become longer than 10 minutes. In addition, they can deliver an accuracy of over 63.4% accuracy given sufficient length of testing segments (at least 5 minutes). Consider the performance when more data is available, i.e. for testing segments of length 5 minutes or longer. For **Model**$_{\textbf{article}}$, decreasing the testing segment length from the entire article (73.2%) to only 5 minutes (63.4%) results in a 9.8% drop. However, such a decrease in data availability has only a modest impact on **Model**$_{\textbf{segment}}$, just a slight decrease by 2.5% for a relatively poor performance for 10-minute segments. On the other hand, consider the performance when not much data is available, i.e. segments of length less than 5 minutes. Reducing the length from 5 minutes to 3 minutes results in significant performance drops by 29.3% for **Model**$_{\textbf{segment}}$ (from 73.2% to 43.9%) and by 24.4% for **Model**$_{\textbf{article}}$ (from 63.4% to 39.0%). This shows that our method can yield a relatively confident prediction as long as the available reading data is more than 5 minutes. However, the performance is not as high, as well as fluctuates with inadequate eye gaze data. We may need to rely on other learning mechanism for a better prediction for those situations.

Inspecting from the data, when the testing segment is shorter than 5 minutes, it is interesting to note that that **Model**$_{\textbf{segment}}$ outperforms **Model**$_{\textbf{article}}$ for majority cases

(except for "1min"). More importantly, **Model**$_{\textbf{segment}}$ makes a marked improvement for the short length segments in general. Notably the use of **Model**$_{\textbf{segment}}$ (56.8%) can produce a much higher accuracy for "0.5min" over **Model**$_{\textbf{article}}$ (35.1%). This corroborates our hypothesis that building the classifiers with the corresponding length of the testing data is rather conducive to the recognition. However, the decrease of **Model**$_{\textbf{segment}}$ for "1min" also reveals a risk that it is not stable enough. This may be because that the feature set adopted in **Model**$_{\textbf{segment}}$ is suitable for the entire length of article, but not for the short segments. Further discussion on the proper feature set for short segments will be provided in the next section, when we perform more extensive evaluation along another dimension.

In summary, with sufficient user eye gaze data (≥5 minutes) our method can consistently achieve satisfactory accuracy (around 70%). However, for applications that need to make predictions in the early stage of reading (≤1 minute), the standard classifiers built for entire articles would need to be supplemented with additional classifiers specifically trained on shorter prefix segments following a similar approach as **Model**$_{\textbf{segment}}$.

3.4.3 Evaluation on Short Segments

Our results in Figure 3-6 highlight the challenge of making prediction in the very early stage of reading. It will be versatile if the reader's comprehension level can be predicted within a short period, e.g. 1~2 minutes. It can enable the continuous prediction of reading comprehension, which facilitates the instantaneous feedback to the readers in applications, such as e-Learning. We need to tackle with the difficulties. This section is dedicated to further study on the relevant issues of the comprehension recognition from and for short segments.

Given the finding that the model learnt from the corresponding length of segments (**Model**$_{\textbf{segment}}$) can outperform the model learnt from the entire segment (**Model**$_{\textbf{article}}$) in the previous section, we focus on the investigation of learning of **Model**$_{\textbf{segment}}$ in this section. As shown in Figure 3-6, there is a performance fluctuation in the early stage (≤2 minutes) of prediction with short testing segments. It might well be due to the inappropriateness of the feature set selected in building the models. This section further evaluates the proper feature sets for short segments of length 0.5, 1, and 2 minutes. The design of these lengths of segment is in line with those in the previous section. More importantly, it is in accordance with the reading behaviors. A close scrutiny of our data

Table 3-14. Data distribution of the derived datasets.

Class / Dataset	High	Medium	Low	Total
GazeData$_{Article}$	9 (22.0%)	15 (41.4%)	17 (36.6%)	41 (100%)
GazeData$_{2min}$	10 (10.4%)	36 (37.5%)	50 (52.1%)	96 (100%)
GazeData$_{1min}$	24 (11.5%)	78 (37.5%)	106 (51.0%)	208(100%)
GazeData$_{0.5min}$	52 (12.0%)	162 (37.5%)	218 (50.5%)	432(100%)

reveals that 0.5 minutes is a lower bound to ensure the subject to finish reading one line and it also allows our gaze features to capture useful information. On the other hand, the shortest time for the subjects to finish reading one article is between 1 and 2 minutes. As a result, we focus the study on the segment of length within 2 minutes.

We first segment the original dataset (named as **GazeData$_{Article}$**) according to the lengths of interest, namely, 2 minutes, 1 minute, and 0.5 minutes. The derived datasets are defined as **GazeData$_{2min}$**, **GazeData$_{1min}$** and **GazeData$_{0.5min}$**. Table 3-14 shows the number and percentage (in the parentheses) of instance in each class in the datasets. Compared with the original dataset, the number of instances in "Low" level of comprehension increases significantly in all three derived datasets, whereas that for "High" level of comprehension only increases at a much slower pace. Thus "Low" becomes the largest class, and "High" becomes an even smaller class. This can be explained by the fact that subjects generally spend more time on articles that are difficult for them to understand. Dividing the data for each article into segments of equal length would create a greater increase of segments for long instances due to "Low" level of comprehension. From Table 3-14 we can see that the baseline of the derived datasets (> 50%, due to "Low") is much higher than that of **GazeData$_{Article}$** (41.4%, due to "Medium").

Table 3-15. Leave-one-subject-out comprehension detection based on derived datasets.

Dataset / CCR	GazeData Article	GazeData 2min	GazeData 1min	GazeData 0.5min
Baseline	41.4%	52.1%	51.0%	50.5%
FeatureSet$_{Article}$	73.2%	68.8%	67.3%	62.5%
FeatureSet$_{Optimal}$	73.2%	75.0%	73.6%	74.1%

Table 3-16. Data distribution of the random sampled datasets.

Class / Dataset	High	Medium	Low	Total
GazeData$_{Article}$	9 (22.0%)	15 (41.4%)	17 (36.6%)	41 (100%)
GazeData$_{2min}$	10 (21.7%)	19 (41.3%)	17 (37.0%)	46 (100%)
GazeData$_{1min}$	24 (21.6%)	46 (41.4%)	41 (37.0%)	111 (100%)
GazeData$_{0.5min}$	52 (21.4%)	101 (41.6%)	90 (37.0%)	243 (100%)

We evaluate our method of reading comprehension detection based on each of the derived datasets of all the subjects' data. To investigate whether the optimal feature set **FeatureSet$_{Article}$** due to the dataset **GazeData$_{Article}$** still works on the derived datasets, we run the leave-one-subject-out cross-validation with **FeatureSet$_{Article}$** based on each dataset. For comparison, we also find out the optimal set of features for each of the derived datasets. Specifically, from the reading activity data of each dataset, we extract the 35 eye gaze features and adopt wrapper method to evaluate the features through 10-fold cross-validation with SVM as before. The optimal set of features (defined as **FeatureSet$_{Optimal}$**) is selected to build the user-independent model for the comprehension detection. The performance obtained from both methods is depicted in Table 3-15. It can be seen that the performance drops for the derived datasets based on the optimal feature set **FeatureSet$_{Article}$** for the original full article dataset. The shorter the segments, the more the performance drop. As the segments become shorter, the feature sets would be affected to a certain degree, and there would be a higher degree of mismatch between the features in **FeatureSet$_{Article}$** with those needed by the derived datasets, and this degree of mismatch would magnify with shorter segments. However, when we are free to adopt the best feature set for each derived datasets, we discover that the performance with **FeatureSet$_{Optimal}$** is much better than that adopting **FeatureSet$_{Article}$**, which is not surprising. However, it is interesting to note that adopting the best feature sets for the individual derived datasets outperforms adopting the best feature set for the original dataset.

One might suspect that this is due to both increase in the number of training instances, as well as the change in data distribution. To make a fair comparison among different datasets, it is reasonable to keep the data distribution and the baseline the same across all the datasets. Therefore, we perform a random sampling on the datasets and maintain the class distribution as that of **GazeData$_{Article}$**, so as to produce data

Table 3-17. Performance of classification of each group of models with the best feature set.

Model	$\mathbf{GazeData_{2min}}$	$\mathbf{GazeData_{1min}}$	$\mathbf{GazeData_{0.5min}}$
1	84.8%	66.7%	67.9%
2	73.9%	82.0%	76.5%
3	84.8%	82.9%	80.7%
4	73.9%	80.2%	78.2%
5	87.0%	73.9%	80.7%
6	91.3%	80.2%	79.8%
7	89.1%	85.6%	80.7%
8	84.8%	78.4%	84.4%
9	80.4%	85.6%	77.0%
10	87.0%	77.5%	74.9%
Average	**83.7**%	**79.3**%	**78.1**%
Std	5.6%	5.4%	4.2%

distribution of "Low", "Medium", and "High" roughly equal to 36.6%, 41.4%, and 22.0% as presented in Section 3.3.3. More specifically, since there is relatively little data labeled as "High" level of comprehension, we preserve all the "High" data and perform a random sampling on the "Medium" and "Low" data of each subject. In doing so, we guarantee that the derived datasets contain data from each subject and each class. Table 3-16 presents the data distribution after sampling. The baseline of each generated datasets is almost the same with that of $\mathbf{GazeData_{Article}}$, so is the data distribution.

In Section 3.4.1, we conclude that there are 6 important features, i.e. $\mathbf{FeatureSet_{Article}}$ (see Table 3-7), adopting which can produce the best-performing model for the original dataset $\mathbf{GazeData_{Article}}$. However, we understand that the optimal feature set can vary as the length of training segments, as indicated in Table 3-15. In other words, the 6 features in $\mathbf{FeatureSet_{Article}}$ may not be suitable to depict the indicative eye gaze behaviors for the short length segments. We thus perform a study similar to that in Table 3-15 for the random sampled datasets, preserving the data distribution and baseline, but with less data.

To investigate the effective feature set for comprehension detection from the short segments with different lengths, we conduct feature selection analysis as presented in Section 3.4.1. We extract the 35 eye gaze features from each segment and build the

user-independent models. Wrapper method with SVM is used for feature ranking through 10-fold cross-validation. We use the CCR of leave-one-subject-out cross-validation as the evaluation metric. The feature set for each model that gives the highest performance in the cross-validation is selected. Since the dataset are produced through random sampling, we repeat the random sampling and feature selection on each dataset for 10 times to evaluate the average or expected performance.

Table 3-17 presents the CCRs on **GazeData$_{2min}$**, **GazeData$_{1min}$** and **GazeData$_{0.5min}$** based on the corresponding optimal feature sets. We notice that the average performance on the sampled datasets outperforms that of the derived datasets in Table 3-15. The recognition performance on **GazeData$_{2min}$** achieves the highest accuracy, attaining 80%

Table 3-18. Selected features from the derived datasets.

Feature	Description	2min	1min	0.5min
c_1	Per-line reading speed	8	10	10
$s_{3_{FS}}$	Average duration of forward saccades	4	10	3
c_2	Repetition rate	5	8	0
$s_{4_{FS}}$	Standard deviation of duration of forward saccades	6	4	0
$s_{6_{RS}}$	Standard deviation of speed of regressive saccades	3	3	0
$s_{7_{FS}}$	Rate of forward saccades	0	8	8
$s_{3_{RS}}$	Average duration of regressive saccades	4	0	5
e_1	Kurtosis of eye movements	5	0	4
f_2	Standard deviation of duration of fixations	3	0	0
$s_{4_{RS}}$	Standard deviation of duration of regressive saccades	0	6	0
$s_{7_{RS}}$	Rate of regressive saccades	0	5	0
$s_{5_{RS}}$	Average speed of regressive saccades	0	4	0
$s_{2_{FS}}$	Standard deviation of amplitude covered by forward saccades	0	0	6
f_5	Rate of fixations	0	0	5
e_2	Skewness of eye movements	0	0	4
$s_{1_{FS}}$	Average amplitude covered by forward saccades	0	0	3
$s_{5_{FS}}$	Average speed of forward saccades	0	0	3

in the 10 iterations. We believe that the eye gaze features presented in this study are useful for comprehension detection with eye gaze data at different granularity and is quite robust across models. Although the use of the best feature set would lead to a high CCR, it is inapplicable for real-use scenarios, since it is infeasible to perform feature selection for unseen datasets. However, these results give us a sense of the best performance that we can achieve based on segments with particular lengths.

We proceed to find out one set of features that can represent the important features for the dataset so that it can be adopted for real-use scenario. We call this the *indicative feature set*. Making the observation that a more important feature is more likely to be included in the optimal feature set, we consider the features that have been frequently selected (≥3 times over 10 iterations as in Table 3-17) as indicative and construct the feature set with these indicative features for recognition with different lengths of data. Table 3-18 shows the indicative feature sets for **GazeData$_{2min}$**, **GazeData$_{1min}$**, and **GazeData$_{0.5min}$**, i.e. **FeatureSet$_{2min}$**, **FeatureSet$_{1min}$**, and **FeatureSet$_{0.5min}$**. The number of iterations for each dataset that a feature is selected is presented. Different shading is used to highlight feature selected by different number of datasets.

Inspecting from the selected features in Table 3-18, Per-line reading speed (c_1) and Average duration of forward saccades ($s_{3_{FS}}$) contribute to all three datasets. Since these two features are closely related to the reading speed, it implies that the reading speed seems to be indicative of comprehension level across different granularity of segments. We also see that there are some features shared by two datasets, including the features describing saccades ($s_{4_{FS}}$, $s_{7_{FS}}$,$s_{3_{RS}}$, $s_{6_{RS}}$), repetition rate (c_2) and variation of eye movements (e_1). It is interesting that features related to repetitive reading behaviors are more favorable for the datasets with longer lengths (**GazeData$_{2min}$** and **GazeData$_{1min}$**), but not in **GazeData$_{0.5min}$**. This is probably due to the fact that those repetitive features become more stable with longer segments and would be more indicative to the recognition. The optimal feature set for **GazeData$_{0.5min}$** consists of a number of features never selected in other datasets. This indicates that such gaze features may not be reliable to capture indicative patterns in short segments. This further corroborates our hypothesis that models with different segment granularity should be associated with their corresponding feature sets in order to attain good performance.

Upon selecting the best feature sets for models with different segment granularity, we would like to study whether those feature sets would be able to produce comparable

Table 3-19. Cross-model evaluation of indicative feature sets.

Feature set \ Dataset	GazeData Article	GazeData 2min	GazeData 1min	GazeData 0.5min
FeatureSet$_{Article}$	**73.2%**	70.2%	70.2%	69.4%
FeatureSet$_{2min}$	70.7%	**80.0%**	75.4%	74.6%
FeatureSet$_{1min}$	68.3%	74.8%	**75.9%**	74.7%
FeatureSet$_{0.5min}$	68.3%	76.3%	75.4%	**77.8%**

performance for the models that they were not initially intended for. Table 3-19 shows the performance of the evaluation across different feature sets and data sets. From the table, we can see that each indicative feature set achieves the best performance on the corresponding data set. This is reflected as the diagonal and that it dominates the rows and the columns. Compared with the results in Table 3-17, which indicates a possibly upper bound on performance, due to the use of the best feature sets for individual iterations, using the indicative feature sets as a representative across all 10 iterations only causes a slight drop of the overall performance. It is encouraging that these identified indicative feature sets for the individual derived datasets can lead to a performance quite close to the upper bound, and these indicative feature sets become practical choice for building applications under real-use scenarios.

3.5 Summary

In this chapter, we propose the method of using eye movements as a modality to recognize scenarios in which humans have difficulty with reading material. Our method uses only consumer-grade devices, namely, a commercial optical eye tracker, together with some very basic information gathered from the article and the overall reading task. We identify eye movement behaviors from the stream of eye gaze locations captured by the eye tracker. We extract features to describe these behaviors. We then adopt machine learning techniques to model the captured data and build user-independent models that are capable of recognizing the comprehension level for new unseen users.

We conduct our experiments via reading tasks, in which the subjects are induced to different levels of comprehension by exposure to articles of varying difficulties. We explore the comprehension detection based on eye gaze data with different lengths of eye gaze segment available for recognition. Through feature selection methods, we

identify the most indicative features for the models to achieve reasonable results. We also look into the features to figure out the relation between human comprehension level and eye gaze behaviors under different contexts.

Chapter 4

A Multimodal Approach to Attention Level Detection in Reading

The last chapter presents a study on comprehension detection based on gaze behaviors extracted from an infrared-based eye tracker. However, such specialized equipment for eye tracking with a high resolution is not likely to be available for the majority people. Besides, there are different interaction signals, such as mouse dynamics and facial expressions, which can be indicative of human mental states and can be captured non-intrusively. We therefore investigate a multimodal approach to attention detection in reading based on off-the-shelf devices in this chapter.

Specifically, we investigate human attention level detection in reading by using ubiquitous hardware available in most computer systems, namely, webcam and mouse. Information from multiple input modalities, including facial expressions, eye gaze movements and mouse dynamics, is fused together for feature extraction and effective human attention detection. We invite human subjects to carry out experiments in reading articles when being imposed upon different kinds of distraction to induce them into different levels of attention. Machine learning techniques are applied to identify useful features to recognize human attention level by building up user-independent models. We also analyze the similarity of different modalities by investigating their contributions to the attention level detection. Our results indicate performance improvement with multimodal inputs from webcam and mouse over that of a single device. We believe that our work has revealed an interesting affective computing direction with potential applications in e-Learning.

4.1 Introduction

Recent advances in miniature hardware have accelerated human-computer interaction research, in enabling the computer to interact better with human. Affective computing research [19][77] had gained tremendous momentum in recent years, demanding computers to understand human affects or emotions and to react accordingly in enhancing user experience. In order to recognize human affects, input signals reflecting human affects need to be acquired and processed. Under traditional KVM (keyboard-video-mouse) settings, input signals are mostly tied to keyboard and mouse dynamics. One can deduce some information about human affect from the keyboard [12][111] and the mouse [110][123], but the accuracy is not particularly high.

Webcam has become a de facto device thanks to the popularity of interactive social networking applications. A human can oftentimes deduce the emotion of a person

sitting in front of a webcam to a certain degree of accuracy. Recent research in video processing and machine learning has demonstrated that human affects and mental states can be recognized via webcam video, noticeably via human facial features [127] and eye gaze behaviors [44]. Though there has been work on mind detection based on facial features and body gestures, research on cognition detection in reading is still limited in the aspects of feature recognition. There is also much work on reading behavior and the associated eye gaze behaviors [47][59][81]. Studies have shown that eye movement and eye behavior during reading is closely related to human comprehension and attention [81][88][89].

In human-computer interaction research, one would often exploit the expressive power resulted from multimodal interaction [69], in which the intention of a user is jointly specified by a plurality of input interaction modalities or signals representing the user. It could be effective in combining and fusing input signals acquired from the keyboard, the mouse and the webcam. In this chapter, we investigate into the detection of human attention level when users are carrying out reading tasks based on a multimodal approach with ubiquitous hardware, namely, the webcam and the mouse, without relying on sophisticated devices such as head-mount devices, electrocardiogram devices or heartbeat belts for additional modalities. The webcam is capable of returning a stream of video frames, which is analyzed for eye gaze behavior recognition, face recognition and then temporal change in facial expression. The mouse is capturing its movement and clicking events, indirectly modeling the user activities of moving down a page for reading. For simplicity, we do not consider keyboard dynamics, since users in general do not utilize the keyboard in reading tasks.

We invite human subjects to carry out experiments in reading English articles, while recording the multimodal interaction data. Changes in human subject attention level are induced via the imposing of various levels of distraction during reading. We apply machine learning techniques to identify useful features that assist in the determination of human attention level. Unlike in some other recent work relying on user-dependent models, we decide to build up the resilient user-independent model, which is more universal to different users, including unseen new users. Our results indicate that by combining the webcam and mouse inputs, there is a significant improvement in attention recognition over the use of a single modality alone. Our work demonstrates the feasibility of determining an interesting human affect, namely, attention level. It could find various applications in e-Learning. For instance, animation and sound effects

could be useful to attract teacher attention when a student starts to lose attention when learning. Change in materials presentation paradigm would be helpful, in a similar way as a teacher adapting to changes in perceived student attentiveness inside the classroom. Human physiological signals [31] could also be integrated into the framework with respect to human stress level during e-Learning.

The rest of this chapter is organized as follows. Chapter 4.2 describe our recognition framework based on webcam video processing and mouse dynamics analysis, as well as the associated machine learning techniques. Chapter 4.3 explains the experimental setups and the experimentation with human subjects carrying out reading tasks. We then evaluate the effectiveness and accuracy of our method in section 4.4. Finally, we conclude this work briefly in Chapter 4.5.

4.2 Multimodal Architecture

In this chapter, we employ multimodal interaction recognition approach to detect the attention level of a user when reading an article. There are three input modalities in our study: facial features captured and returned by a webcam in the form of a video clip,

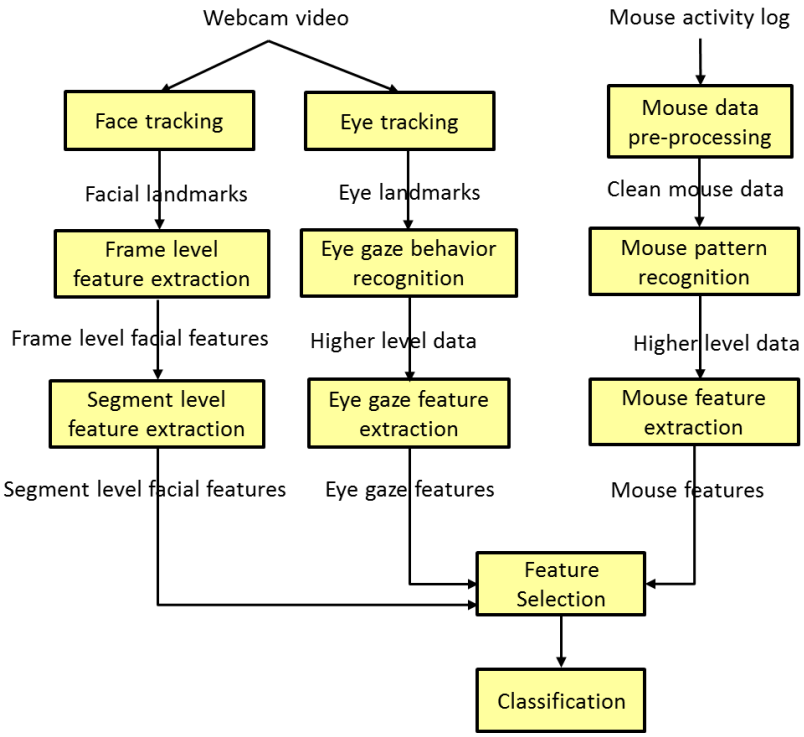


Figure 4-1. Multimodal recognition framework.

eye gaze behavior extracted from the webcam video clip, and the mouse dynamics captured by a mouse logger program. In the subsequent subsections, we will describe the actual feature extraction mechanisms for the three modalities, followed by the way to select the set of useful features. The overall mechanism is depicted in Figure 4-1.

4.2.1 Facial Features

A two-level facial feature extraction approach is adopted in our work: frame-level and segment-level, as depicted in Figure 4-1. We perform feature extraction in each frame of a video clip and generate a set of frame-level facial feature vectors. We divide each video clip for an experimental subject carrying out a task into smaller units called segments. Each segment is composed of a good number of frames. Based on the frame-level facial feature vectors, segment-level feature extraction consolidates and generates a single segment-level feature vector to represent the whole segment. Before we perform frame-level feature extraction, we must first be able to recognize and track the human face in the video. Instead of performing face recognition from scratch for individual video frames, we adopt the face tracking approach. Once a face is recognized in a frame, we assume delta movement of the face in the subsequent frames. This can be achieved by computing for the facial landmarks and then their displacement across frames. Only in the event when the face loses track due to excessive movement (often due to large degree of head rotation) then the face needs to be recognized from scratch.

To perform frame-level facial feature extraction, we apply Constrained Local Models (CLM) [91] to track 66 facial landmarks from the video clips. This model is trained on the CMU Multi-PIE Face database [33], which contains over 750,000 images from 337 people. However, it fails to track some of the mouth movements, such as mouth corner depression. Thus, the Supervised Descent Method [121] is adopted to

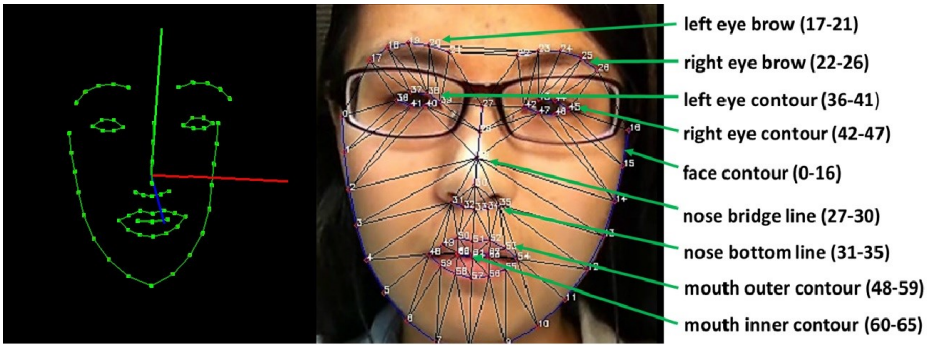


Figure 4-2. Facial landmark tracking via CLM.

validate and optimize the 2D landmark locations. During CLM optimization, the 2D and 3D landmarks and other global and local parameters are adjusted iteratively until the face fitting regression model converges. Removing the rigid transformation from the acquired 3D shape compensates for the influence of out-of-plane rotation and produces the aligned 3D landmarks. Figure 4-2 indicates our usage of CLM to track the 66 facial landmarks.

We follow a standard approach to extract facial features, referred to as Action Units (AUs) [109]. We calculate the normalized distances and angles between the corresponding facial landmarks, which represent the direction and intensity of the facial movements, by extending AUs with only discrete intensity levels.

Table 4-1 summarizes the descriptions and measurements of the 20 facial features (f_1 to f_{20}) that we calculate from the 66 aligned 3D facial landmarks. Observing that the head orientation and position also play an important role in facial expression representation, we augment our feature list with 6 more head-oriented features (f_{21} to f_{26}). The first three features measure head orientation with respect to *x*, *y* and *z* axes in the webcam coordinate system. The remaining three features measure head position, with the face center position represented in the 2D image coordinate and the size of the face, revealing the distance between the face and the screen.

From our pilot study, we discover that variations in both head movement and lighting condition (e.g., heterogeneous illumination and camera exposure) have posed significant challenges for the appearance-based features, especially with elderly people with natural wrinkles. As a result, we move away from texture- and color-based features to geometry-based features which are more resilient to variation to movement and illumination. This has significantly enhanced the robustness of our model in real-usage situations in the presence of uncontrollable environmental variations. The use of geometric facial features has effectively mitigated the noise arising from the textural and appearance channels.

After performing frame-level facial feature extraction, we extract three kinds of segment-level facial features based on the frame-level facial feature vectors reflecting different statistical behaviors. The first behavior that we are interested in is the *average frame* inside the segment. The second behavior is the *variation of frames* contained within the segment over a moving window. The third behavior is the *variation* with respect *to an anchor frame*.

We hope that this three-way representation of the frame statistical variations suffices in providing us with a good sense of the macro-behavior of the user, while is simple

Table 4-1. Facial features extracted from video.

Feature	Meaning	Formulation
$f_{1,2,3,4}$	Inner and outer brow movement	Distance between eye brow corner and the corresponding eye corners (left & right)
$f_{5,6}$	Eye brow movement	Distance between eye center and the corresponding brow center
$f_{7,8}$	Eye lid movement	Sum of distance between corresponding landmarks on the upper and lower lid
f_{9}	Upper lip movement	Distance between landmark 33 of nose bottom and landmark 51 of mouth outer contour
$f_{10,11}$	Lip corner puller	Distance between mouth corner and the corresponding eye outer center
f_{12}	Eye brow gatherer	Distance between inner eye brow corners
f_{13}	Lower lip depressor	Distance between landmark 8 of face contour and landmark 57 of mouth outer contour
f_{14}	Lip pucker	Perimeter of the mouth outer contour
f_{15}	Lip stretcher	Distance between the mouth corners
f_{16}	Lip thickness variation	Sum of distance between corresponding points on the outer and inner mouth contours
f_{17}	Lip tightener	Sum of distance between corresponding points on upper and lower mouth outer contour
f_{18}	Lip parted	Sum of distance between corresponding points on upper and lower mouth inner contour
f_{19}	Lip depressor	Angle between mouth corners and lip upper center
f_{20}	Cheek raiser	Angle between nose wing and nose center
$f_{21,22,23}$	Head orientation	Head orientation in 3D coordinate
$f_{24,25,26}$	Head position	Face center position in 2D image coordinate and face size

enough without introducing too many features to begin with. In our experiments, we select segments of length of 1 minute each.

The first set of segment-level features derived from the 26 frame-level facial features in Table 4-1 is calculated as the mean value of the features. For each feature f_i, we compute for each segment containing S frames the average feature values of all S frames inside the segment. For notational convenience, we denote this set of segment-level features as f_i_mean, where f_i ($i \in [1,26]$) is the corresponding frame-level facial feature. Altogether, there are 26 features in this set.

The second set of segment-level features is computed based on moving windows of size W (we select $W = 15$ based on the frame rate of 15 in our experiment). The frames in the segment are divided into units of W frames each. For each feature f_i, the difference in feature values between the first frame in the window and the last frame in the window is computed. Then we compute the mean and standard deviation of the set of S/W feature value differences for each feature over the segment. We denote this set of segment-level features as $f_i\_window\_mean$ and $f_i\_window\_std$ for the corresponding frame-level feature f_i. There are a total of 52 features in this set.

The third set of segment-level features is computed based on a special anchor frame. In particular, we adopt the face in the first frame of the video as the *neutral face* and consider changes in face in other frames with respect to this neutral face (i.e., *delta face*). In this third set of features, we consider the face as a whole instead of individual features. As a result, we compute one single value for the face in each frame with respect to the anchor frame (first frame) for the neutral face. We treat those 26-element feature vectors for each frame as a unit, and compute the Euclidean distance between the feature vector of frame F_j and that of first frame F_1. This will give us $S - 1$ Euclidean distances for a segment of size *S*. Finally, we compute the mean and standard deviation of those $S - 1$ distances to result in only 2 *global* features. These two features are denoted as $face\_mean$ and $face\_std$.

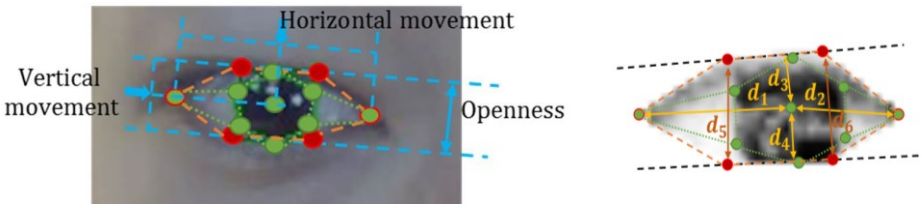
4.2.2 Eye Gaze Features

As illustrated in Figure 4-1, we extract eye gaze features from the webcam videos by eye gaze tracking and eye gaze behavior recognition. In this section, we analyze three kinds of eye gaze behaviors for reading attention detection, including eye blinks, eye fixations and eye saccades. Before this can be done more precisely, we need to

estimate the position of the pupil center of each eye, as well as extracting some other useful eye landmarks.

As presented in the last subsection, the face CLM consists of 66 facial landmarks. Out of them, we identify 6 landmarks associated with the contour for each eye. This is depicted in Figure 4-3a, inclusive of 4 around the eye in red circles and 2 at the corners of the eye in green circles with red border. In order to accurately describe the eye gaze behaviors, it is crucial to properly locate the pupil center, which often cannot be detected from the appearance information of the eye region in unconstrained situations, reflected by the facial landmarks. Furthermore, the low resolution in the video, as well as light reflections on glasses and cornea usually makes the region of the pupil and its periphery almost unobservable. To address these issues, instead of attempting to identify the pupil from individual frames, we apply the CLM based on the eye [43] to track the key pupil center and 8 other eye landmarks with good salient features on the iris contour and eye lid corners across frames, making use of the temporal consistency property. This is depicted in Figure 4-3a in the form of green circles. Note that the 2 landmarks at the eye lid corners (green circles with red border) both serve among the 66 facial landmarks (facial features in Section 4.2.1) as well as among the 9 eye landmarks (eye gaze features in Section 4.2.2).

Based on the 6 landmarks identified from the face and 9 landmarks from the eye CLMs (a total of 13 landmarks), we can compute the 6 key eye landmark distances, d_1 to d_6, in Figure 4-3b accordingly. From these landmark distances for each eye, we would like to establish the eye geometry, namely, the eye openness, the relative horizontal position and vertical position of the eye gaze. Eye openness is employed in the detection of the eye blinks, whereas temporal changes in the horizontal and vertical positions of the eye gazes are adopted in the detection of eye fixations and saccadic movements.



(a) facial and eye landmarks (b) key eye landmark distances

Figure 4-3. Eye landmarks and features.

We first recognize eye blinks according to the value of $d_5 + d_6$ of each eye, which represents the eye openness as shown in Figure 4-3a. As in previous studies, an eye blink is defined as eyelid closure for a duration of 50 to 500 ms [93]. Given the eye openness of each frame in a video segment, eyelid closure events can be easily detected by identifying the moments when the eye openness value of each eye goes down to 0. The sequences of eyelid closure events with duration shorter than 50 ms or longer than 500 ms are discarded as noise, which may be caused by the occasional tracking failure of the eye CLM or the turning away of the subject's head. The remaining eyelid closure event sequences are considered as eye blinks. The duration of the eye blink is the length of corresponding eyelid closure sequence.

Upon identifying eye blinks, we need to classify the remaining eye gaze behaviors into eye fixations and saccades, namely, whether the eye gaze is focused on a word for mental processing, or moving for reading. To distinguish fixations and saccades, we analyze the horizontal and vertical movements of both eyes. For each eye, we compute the relative eye gaze position within the eye, independent on the actual coordinates of the eye in the frame. These relative horizontal and vertical eye gaze positions for an eye are computed as $\frac{d_1}{d_1+d_2}$ and $\frac{d_3}{d_3+d_4}$ in each frame. As illustrated in Figure 4-3a, the movements of the eyes over a temporal period can be analyzed from the eye gaze position sequence. Considering that for most human, both left and right eyes move together, we thus simplify the representation of eye gaze position by computing the mean value of the eye gaze positions of the left and right eyes.

The eye gaze position sequence can then be represented as $EG = < EG_1, \dots, EG_k >$ of k eye gaze points:

$$EG_i = \left[EG_{h_i}, EG_{v_i}\right]$$

where EG_{h_i} is the horizontal component of the eye gaze position of the i^{th} item in the sequence, defined as the average of the horizontal positions of the two eyes, and EG_{v_i} is the corresponding vertical component, as the average of the vertical positions. The movement of the eye gaze is measured as the Euclidean distance between the corresponding eye gaze points in the eye gaze sequence EG.

Eye fixations are defined to be periods in which the eye gaze remains stationary on a specific location. However, due to the inherent error of the eye CLM model and head movement, detecting fixations from the eye gaze signal EG becomes more than simply

looking for periods during which the eye gaze positions do not change. To determine the extent of noises on fixation detection, a pilot study was carried out to analyze the samples of gaze fixation on a single word. The eye gaze position sequences were calculated and the eye gaze movements between successive frames were analyzed to estimate the potential impact of noises. Let us define $mean_{MOVE}$and std_{MOVE} as the mean and standard deviation of the eye gaze movements detected by the eye CLM model between successive frames for the periods of eye fixation. To filter the noise exerted on the eye gaze signal, we define τ_{FA} as the *fixation amplitude threshold*, where

$$\tau_{FA} = mean_{MOVE} + 3 * std_{MOVE}$$

Define D_i as the eye gaze movement between successive eye gaze points EG_i and EG_{i+1}, a vector F can then be constructed from EG, where

$$F_i = \begin{cases} 1, & D_i < \tau_{FA} \\ 0, & D_i \geq \tau_{FA} \end{cases}$$

This gives us a binary vector in which elements with a value of 1 correspond to the moments when the eye gaze could be considered to be stationary. Since it has been found that fixations are rarely less than 100 ms and usually in the range of 200 to 400 ms [90], we label fixations as continuous stationary sequences that last for longer than 100 ms but shorter than 500 ms. Once the eye blinks and eye fixations have been identified, the sequences in between the fixations with duration shorter than 200 ms are

Table 4-2. Eye gaze features adopted.

Feature	**Meaning**	**Formulation**
e_1	Blink rate	Number of eye blinks per minute
$e_{2,3}$	Blink duration	Mean (e_2) and standard deviation (e_3) of the eye blink durations
e_4	Fixation rate	Number of fixations per minute
$e_{5,6}$	Fixation duration	Mean (e_5) and standard deviation (e_6) of the fixation durations
e_7	Saccade rate	Number of saccades per minute
$e_{8,9}$	Saccade duration	Mean (e_8) and standard deviation (e_9) of the saccade durations

considered as saccadic eye gaze movements as defined in [96]. The duration of a fixation and a saccade is the length of the corresponding eye gaze sequence.

After the three different eye gaze behaviors have been identified from the sequence of eye gaze positions, we construct the 9 statistical features that will be used to describe these three behaviors as shown in Table 4-2.

According to our observation of the eye gaze behaviors, the eye fixation is very indicative of the human attention level. It is notable that a reader tends to have long fixations while paying high attention to reading. This implies the reader makes efforts to process the information from the reading materials. In contrast, short fixations happen when the reader's attention level is low. The fixation rate is also important. Readers at low attention level usually read repeatedly until they fully understand the reading materials, which results in a high fixation rate. Besides eye fixations, eye blinks and saccades may also contribute to our research problem. Previous studies [25] have shown that eye blinks are correlated with human cognition, such as fatigue. Eye saccades can reflect the reading speed, which is closely related to the attention level.

4.2.3 Mouse Dynamics Features

Mouse dynamics have been shown to provide indicative information for affect detection in various research works [110][123]. In this section, we attempt to relate mouse dynamics with human reading attention level, by analyzing typical mouse dynamics, including mouse click, mouse movement and mouse scrolling. Similar to facial expression recognition, we process raw mouse events to establish mouse dynamics over time. We then extract features representing mouse dynamics for each segment to align with the segment in the video clip. This enables signal fusion among the different modalities, namely, mouse signals and webcam signals.

As depicted in Figure 4-1, we pre-process the mouse activity log to clean extreme data values that may be due to noise. We then extract mouse patterns and then compute the actual features reflecting the mouse dynamics at the segment granularity. For instance, we compute the total distance traveled by the mouse by summing up the individual Euclidean distances traveled throughout the segment for each pair of sampled mouse coordinates. Similarly, each pair of mouse coordinates indicates a mouse moving direction and the change in mouse movement direction is computed as the absolute difference in angle between the directions indicated by two consecutive pairs of mouse coordinates. Mouse scrolling features are computed based on the log of

Table 4-3. Mouse features adopted.

Feature	Meaning	Formulation
m_1	Mouse click	Number of mouse clicks
m_2	Mouse distance	Distance traveled by the mouse in pixels over the screen
m_3	Mouse direction	Amount of change in direction encountered by the mouse in angle
$m_{4,5}$	Mouse scroll count	Number of scrolls and number of changes in scroll direction (up and down)
m_6	Mouse scroll step size	Number of discrete steps per scroll
m_7	Mouse scroll speed	Average speed of mouse scrolls (step size over time period of scroll)

scrolling events, each of which occurs when the wheel is scrolled one discrete step. Consecutive scrolling events occurring within 1 second are considered to belong to the same scroll when the scroll step size is computed. The set of features extracted for mouse dynamics is depicted in Table 4-3, which can be categorized into three types: mouse click (m_1), mouse movement ($m_{2,3}$), and mouse scrolling ($m_{4,5,6,7}$), generated from the three mechanical components of the mouse (button, trackball and scroll wheel).

We notice that mouse direction is an important feature in demonstrating the "roughness" of the user. A conscious user would normally move the mouse in relatively straight lines without many changes in directions. Rapid directional changes often indicate confusion or restlessness. An increase in number of scrolling steps indicates relatively fast article reading, implying generally a higher level of attention.

4.2.4 Feature Selection and Classification

After extracting the set of potentially useful features, feature selection needs to be conducted to remove non-indicative features and to improve classification performance in pattern recognition and machine learning applications. In our work, we have extracted an initial set of 80 facial features, 9 eye gaze features and 7 mouse features, too many to be effective for practical real-time recognition, especially for facial features.

We adopt the wrapper method for feature selection which is reported to outperform filter method by considering the relationship between different features and selecting

Table 4-4. Potential facial features for consideration.

Ranking	Feature
1	f_{14}_mean
3	$f_{5}\_window\_mean$
6	f_{10}_mean
8	$f_{7}\_window\_std$
13	$face\_std$
17	$f_{7}\_window\_mean$
20	$f_{1}\_window\_std$
25	$face\_mean$
26	$f_{3}\_window\_mean$
32	f_{13}_mean
35	$f_{24}\_window\_std$

one feature subset that is best for the chosen classifier [106]. We adopt the best first searching approach for its efficiency, based on the Linear Support Vector Machine (SVM) for classification. This filtering step is very efficient in reducing the set of potential facial features from 80 down to 11. In other words, many of the original 80 features would not contribute much to the recognition task, manifested by the fact that recognition performance is not affected upon their removal. The list of potentially useful facial features is depicted in Table 4-4.

In Table 4-4, the ranking indicates the relative importance of the single feature contributing to recognition. It is simply computed as the percentage of training sets that the feature is selected for classification. Note that features in pair form are often of similar values and would often contribute similarly towards recognition, so that one of them would suffice and the better one would be selected, e.g., left eyebrow movement (f_5 ranked 3^{rd}) edges out right eyebrow movement (f_6 ranked 5^{th}). The second-ranked feature f_{16} on lip thickness also highly correlates with the first ranked feature f_{14} on lip pucker, so that the use of f_{16} suffices. It also subsumes other top-ranked lip features such as f_9 and f_{11}, and eventually f_{10}. Some features may rank high when used alone, but not compatible with other features in a way that putting them together may actually lower the accuracy. That is why a simple regression-like algorithm in eliminating weak features may not always work, and backtracking is needed in the heuristic feature selection approach. There are only 9 features for eye gaze and 7 features for mouse dynamics, making initial feature selection unnecessary.

Table 4-5. Final set of features adopted.

Facial feature	Attribute	Eye gaze	Attribute	Mouse dynamics	Attribute
f_{14}_mean	lip pucker	e_5	fixation duration	m_6	scroll step size
$f_5\_window\_mean$	eye brow movement	e_4	fixation rate	m_3	mouse direction
$f_7\_window\_std$	eye lid movement	e_8	saccade duration	m_2	mouse distance
$face\_std$	whole face	e_1	blink rate		
$f_7\_window\_mean$	eye lid movement	e_2	blink duration		
$f_1\_window\_std$	inner eyebrow movement				
$f_{24}\_window\_std$	head position				

After initial feature selection in trimming down the set to a manageable size, we can explore different feature subset combinations via an exhaustive search for the most impressive feature set to build up our attention level recognition model. We end up with 7 top facial features, 5 top eye gaze features and 3 top mouse dynamics as the best combination, as depicted in Table 4-5.

4.3 Experiments for Data Collection

We invite experimental subjects to conduct experiments to validate our multimodal approach to attention detection for reading tasks in a real-world setup. The subject is reading an article in full screen, using the mouse to navigate through the article. This is depicted in Figure 4-4.

In order to induce different attention levels for experimental subjects when reading, different types of vocal stimuli are applied to distract the subjects. Subjects would need to self-report their level of attention to serve as ground truth for classification. The subjects' facial expressions and mouse dynamics are both recorded in real time during

the experiment. The subjects are also required to do a pre-experiment survey and post-experiment survey for information collection and labeling.

4.3.1 Participants and Experiment Setup

We have recruited 6 subjects aged between 22 and 30, averaging 25.5. Two are undergraduates and four are graduates, whereas four are female and two are male. According to the pre-experiment survey, all subjects are skilled in using computer and capable of reading in English though their English ability varies. All are non-native speakers; the native language of two subjects is Mandarin while that of the other four subjects is Cantonese. This dictates the choice of the distracting vocal stimuli used in the experiments. Although they share the common written Chinese characters, the two dialects differ enough that speaker in one dialect without proper training or sufficient immersion would have much difficulty in understanding the other.

The experiment is carried out in a common office environment in the CHI Lab. As shown in Figure 4-4, the standard setup of the experiment consists of a 22-in flat LCD screen with a resolution of 1680 × 1050 pixels for displaying the articles to read, a common webcam fixed on the top of the display to record the subjects' face and upper body, and a common wired mouse. All the devices are non-intrusive to the subjects. The light in the room is adjusted to be suitable for reading and is maintained stable

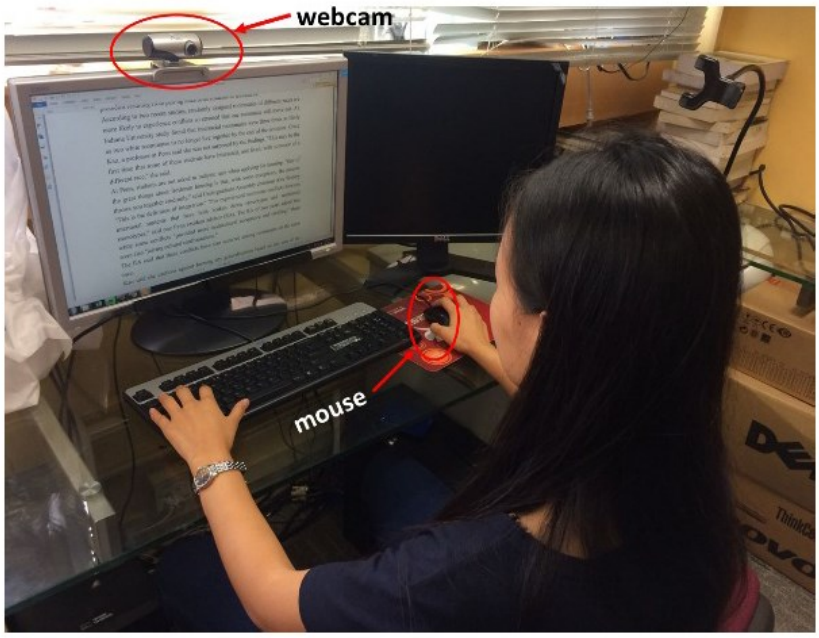


Figure 4-4. Experimental Setup.

throughout the experiment. The subjects are seated about 60 cm away in front of the display.

Data collection programs run on the computer displaying the article for reading. Both the content shown on the screen and the webcam vision are recorded by a free-trial version of the software Camtasia, capturing the two video streams at a frame rate of 15 per second onto the hard disk. We develop a C++ program to capture and log mouse events to determine mouse dynamics, including mouse click, mouse scrolling and mouse movement, together with their timestamps. Mouse click and scroll events are logged when they occur, and mouse coordinate is sampled at a rate of approximately 15 per second for mouse movement. The program is run concurrently and the timestamped information is stored in the hard disk for temporal alignment.

4.3.2 Experiment Design

In our experiment, each subject is required to read three different English articles chosen from TOFEL (Test of English as a Foreign Language) reading comprehension materials. We decide to select articles from TOFEL because the topic, length and difficulty of the articles are proper for our non-native speaker subjects in this reading experiment. The time spent on reading is not constrained. In this study, a reading session refers to the particular experiment in which one subject reads one TOEFL article. To make sure that the subjects really read the articles with a reasonable amount of efforts instead of just killing time, they are required to write a short summary of at least 50 words after finishing reading the article in each session.

In the first set of sessions, the subjects read in a quiet environment without anything to distract them. To induce different levels of attention on the subjects, we choose two kinds of vocal stimuli to distract the subjects on purpose during reading in the second and third set of sessions. One of the vocal stimuli is heavy metal music which carries a “high information-load” and supposedly to be able to impair performance significantly in reading comprehension task [6]. The other vocal stimuli are sound recording of famous funny talk shows that the reader would very likely be interested in. Considering the different native languages of our subjects, we choose Mandarin and Cantonese talk shows for the subjects based on their native language. By doing this we make sure that all the subjects can understand the contents of the talk shows easily even in the background, so as to distract them.

At the end of each session, the subjects label their level of attention throughout the reading tasks with "*low*", "*medium*" or "*high*" on a per minute basis. Prior to the labeling, the subjects were shown a document of guidelines of doing self-reports of the level of reading attention, which provides them with criteria to make a precise decision of the self-reports and avoids the potential inconsistency among subjects. To help the subjects remember the reading process and their mental state so that to make a reliable labeling of the level of attention, they are displayed with video clips of the screen and their face recorded during the reading task minute by minute and they label immediately after watching each minute. It has been demonstrated recently that watching video clips and giving a label for the entire video is a more impressive approach for labeling than giving continuous labels while watching video clips [116].

4.3.3 The Dataset

We have to perform pre-processing to the video clips since there are occasional instances with subjects showing only a partial face, caused by inappropriate sitting position of the subject. As our facial expression model depends on the key landmarks throughout the face, a partial face without the mouth would not be useful. We thus remove those occasional corrupted video data. The amount of such bad data only contributes to less than 10% of the total data. Finally, we are able to collect data of a length of 147 minutes for all the six subjects (about 25 minutes per subject).

We next establish the ground truth and baseline from the dataset for evaluation purpose. According to the attention level labeled by our subjects, 35.4% of the data is labeled as "*high*", 34.7% of the data is labeled as "*medium*" and 29.9 % of the data is labeled as "*low*". This is a set of very well-mixed data, since the three classes are roughly equally represented without much skewness.

The baseline of the dataset is 35.4%, since the bottom line for random guessing in classification is to output the label of the largest class for a "best" result. This baseline of a dataset is widely used to evaluate the classification performance of an algorithm. In this chapter, we build up user-independent models for attention detection based on this dataset.

4.4 Results and Analysis

In this section, we evaluate our multimodal attention detection approach by building user-independent models based on the combined dataset of all subjects. In classification research, a user-independent model is usually not as accurate as a user-dependent model, but is more applicable in practice. The former can be built easily but the latter has to be built for each individual subject and the amount of data needed for training the classifier will be much larger. User-independent models can be applied to new users, but user-dependent models cannot. We build up user-independent models and evaluate our approach in this section.

We compare the classification performance of our multimodal approach with the performance produced by using only a single modality. The results illustrate that the multimodal models perform better than the single modality ones, achieving higher correct classification rate (CCR) and F-measures

4.4.1 Attention Detection with Facial Features

Our first evaluation is concentrated on the use of facial features extracted from the webcam video to recognize attention level in reading tasks. There are a total of 80 extracted facial features across three categories, trimmed down to 11 via the wrapper approach upon adopting the Linear Support Vector Machine (SVM) with 10-fold cross-validation to classify the dataset.

From the set of 11 potential facial features, we attempt different subset combinations and select 7 producing the best performance, as shown in Table 4-5. We can see that features describing the change of frame-level feature vectors are mostly chosen. This indicates that the magnitude of the change of facial expression of specific areas on the face varies with the level of attention of the subjects. Within all the selected features, the change of eyebrow position and eyebrow movement are particularly important when compared with other features.

Since we are building user-independent models, the gold standard in evaluating the effectiveness is the leave-one-subject-out cross-validation test. From the set of n subjects, we train the user-independent model with dataset from $n-1$ subjects and test the model on the left-out subject. We repeat the experiment n times by leaving out a different subject and the average performance is reported. The confusion matrix

Table 4-6. Normalized confusion matrix for facial feature model.

Ground truth \ Classified as	Low	Medium	High
Low	0.73	0.21	0.06
Medium	0.19	0.69	0.12
High	0.07	0.29	0.64

Table 4-7. Classification performance for facial feature model.

Attention Level \ Performance	Precision	Recall	F-measure
Low	0.75	0.73	0.74
Medium	0.60	0.69	0.64
High	0.76	0.64	0.69

normalized by the ground truth and the performance matrix for classification are shown in Table 4-6 and Table 4-7.

From Table 4-6 and Table 4-7, it can be observed that the average CCR for the three classes is 68.7%, and this is significantly higher than the baseline of 35.4% with an improvement of 33.3% (doubling the accuracy). It can also be seen that most of the errors come from misclassifying as the neighboring attention level class, i.e., *low* ↔ *medium* and *medium* ↔ *high*. Only very few errors are due to misclassification of extreme classes between *low* ↔ *high*. Similarly, we are able to achieve a high precision as well as a high recall, without having to sacrifice one metrics for the other. The resultant F-measure is also as high as 0.7, close to the CCR.

4.4.2 Attention Detection with Eye Gaze Features

We believe that the 9 eye gaze features will not contribute equally to the attention level classification. To explore the most indicative set of eye gaze features, we compare the classification performance with different combination of eye gaze features and find out 5 useful eye gaze features, as shown in Table 4-5. Those 5 eye gaze features are e_1, the rate of eye blinks, e_2 , the average blink duration, e_4, the rate of eye fixations, e_5, the average fixation duration, and e_8, the average saccade duration. It is worth noticing that features representing all three kinds of eye gaze behaviors analyzed in this work are

Table 4-8. Normalized confusion matrix for eye gaze feature model.

Classified as / Ground truth	Low	Medium	High
Low	0.77	0.19	0.04
Medium	0.39	0.53	0.08
High	0.27	0.30	0.43

Table 4-9. Classification performance for eye gaze feature model.

Performance / Attention Level	Precision	Recall	F-measure
Low	0.56	0.77	0.65
Medium	0.54	0.53	0.53
High	0.76	0.43	0.55

selected in the subset. It indicates that there is a strong correlation between the eye gaze behaviors and the level of attention in our reading task. Moreover, the top two ranked features are both eye fixation features, which validates our findings that eye fixation is critical to the attention level detection in our work. According to Table 4-5, none of the eye gaze features representing the standard deviation of the eye gaze behaviors (e_3, e_6 and e_9) is selected. It perhaps implies that the eye gaze behavior patterns are quite stable with a certain attention level. We build a user-independent model based on eye gaze features alone and perform the leave-one-subject-out cross-validation test. The confusion matrix normalized by the ground truth and the performance matrix for classification are shown in Table 4-8 and Table 4-9.

As shown in Table 4-8 and Table 4-9, the average CCR for the three classes is 58.5%, which is higher than the baseline by 23.1% with only 5 features. Similar to the facial feature model, the CCR of the *low* class is better than that of the *medium* class, while the *high* class is still the one with biggest misclassifying errors. Although the errors still mainly come from misclassifying between *low* ↔ *medium* and *medium* ↔ *high* as in the facial feature model, we note that the error to misclassify *high* as *low* becomes bigger than the facial feature model. It means the eye gaze behaviors analyzed in this study do not correspond that well with the level of attention as with facial expressions. This may sound intuitive, since the facial expression carries inherently

richer information than the eye gaze alone. Nevertheless, the eye gaze features still contribute a lot to the attention level classification, despite its relatively small amount of features and landmarks required. Finally, the average recall and precision for the three classes are 0.57 and 0.62 respectively, whereas the average F-measure is 0.57, consistent with the CCR and somewhat lower than those performances based on facial features.

4.4.3 Attention Detection with Mouse Dynamics

There are only 7 mouse dynamics features but not all of them contribute well to the classification process. We therefore explore different subsets of feature combinations for mouse dynamics and we land on 3 useful mouse features for classification as shown in Table 4-5. Those useful features are m_6, the amount of scrolling steps, m_3 and m_2, the amount of changes in mouse direction and total distance that the mouse travels. In our experiment, we observe that the mouse click events are not indicative at all. This is because most subjects only use the mouse scrolling button to navigate up and down the article, instead of clicking on the scroll-bar in the application window in this reading task. The distance traveled and direction changed for the mouse come up as important features contributing to the classification of the attention level. The mouse click events would be more useful when writing tasks are studied, so would keyboard dynamics be.

Table 4-10. Normalized confusion matrix for mouse dynamics model.

Classified as / Ground truth	Low	Medium	High
Low	0.44	0.31	0.25
Medium	0.29	0.43	0.28
High	0.23	0.29	0.48

Table 4-11. Classification performance for mouse dynamics model.

Performance / Attention Level	Precision	Recall	F-measure
Low	0.48	0.44	0.46
Medium	0.43	0.43	0.43
High	0.44	0.48	0.46

In any case, the selected features demonstrate that the mouse trajectory is indicative for attention level classification. The normalized confusion matrix on classification and its accuracy based on mouse dynamic features is depicted in Table 4-10 and Table 4-11.

According to Table 4-10 and Table 4-11, the average CCR for the three classes is around 44.9%, which is not as good as the performance of the facial feature model and the eye gaze feature model. When compared with the baseline of 35.4%, there is still an improvement of 9.5%, even with as few as 3 mouse features. Although the improvement is not as impressive when compared with those of facial features, the result is already acceptable with just 3 features. We believe that the lack of useful information about the mouse dynamics during the reading task drags the classification performance to a certain extent. It can also be observed that there are more classification errors across extreme classes, i.e., *low* ↔ *high*. This is perhaps due to the fact that mouse dynamics do not correspond that well with the attention level as with facial features and eye gaze features. Nevertheless, the recall and precision metrics and the F-measures for the three classes remain stable at about 0.45, similar to the CCR.

4.4.4 Attention Detection with Multimodalities

We have already observed good recognition with the unimodal models based on facial features and acceptable recognition based on eye gaze behaviors and mouse dynamics in our study. We now adopt the multimodal model by combining the features of all the modalities. There are a total of 15 features in this multimodal recognition study as shown in Table 4-5. We build user-independent models based on SVM and apply 10-fold cross-validation in the evaluation. As before, we employ the challenging leave-one-subject-out cross-validation experiment over the *n* subjects. The confusion matrix normalized by the ground truth and the performance matrix for classification are shown in Table 4-12 and Table 4-13.

From the two tables, the average CCR for the three classes is found to be 75.5%, an improvement of 40.1% over the baseline, with the accuracy of one class going up to 81%. Although the classification performance based on mouse dynamics is much lower than that one based on facial features or eye gaze features, the overall performance has been improved compared with individual performance, when the three modalities are combined. The classification errors across neighboring classes and especially the extreme classes of *low* ↔ *high* have all been reduced when compared with the use of

Table 4-12. Normalized confusion matrix for multimodal model.

Ground truth \ Classified as	Low	Medium	High
Low	0.81	0.13	0.06
Medium	0.12	0.76	0.12
High	0.11	0.20	0.68

Table 4-13. Classification performance for multimodal model.

Attention Level \ Performance	Precision	Recall	F-measure
Low	0.79	0.81	0.80
Medium	0.71	0.76	0.74
High	0.77	0.68	0.72

features of single modality. It is also worth noticing that the performance of the *medium* class improves dramatically compared with the eye gaze feature model and the mouse dynamics model. When we look at the recall and precision metrics, they show a similar pattern as that of CCR with a comparable F-measure. In summary, we believe that our selected features of different modalities contribute to the attention level detection in reading in a synergic way.

4.4.5 Contributions by Individual Modalities

We can attain different performance based on features generated from each individual input modality. From Section 4.4.1 to 4.4.3, it is easily seen that facial features produce the best performance, followed by eye gaze features and finally mouse features. However, in terms of computational cost, the reverse is true. This is the rationale behind the choice of a proper multimodality feature set to yield a good enough recognition rate. In this section, we proceed to analyze more deeply the individual contribution by each modality and see which combinations would produce a better integrative performance.

We conduct three more experiments based on (*a*) combined facial and eye gaze features, (*b*) combined facial and mouse features, and (*c*) combined eye gaze and mouse features. From there, we would be able to identify the contribution by individual modality more precisely. The corresponding confusion matrices for the three combinations are summarized in Table 4-14. It can be seen that they exhibit

Table 4-14. Normalized confusion matrix for multimodal models.

Facial + eye gaze	**Low**	**Medium**	**High**
Low	0.79	0.15	0.06
Medium	0.25	0.73	0.02
High	0.11	0.20	0.68
Facial + mouse	**Low**	**Medium**	**High**
Low	0.83	0.12	0.06
Medium	0.22	0.69	0.10
High	0.16	0.18	0.66
Eye gaze + mouse	**Low**	**Medium**	**High**
Low	0.83	0.15	0.02
Medium	0.37	0.49	0.14
High	0.18	0.27	0.55

intermediate performance with respect to those for single component modalities and the one for the full set of modalities, as compared with those in the previous tables. The precision/recall metrics show a similar pattern as in the previous experiments and are thus omitted.

For comparison, we report the CCR for these combinations, alongside those of the individual feature sets. We also compute the improvement in CCR performance for each combination. This improvement indirectly measures the "synergic" effect between the two feature sets. It is conceivable that a higher synergic effect is more preferred. The results are depicted in Table 4-15.

We can observe from Table 4-15 that facial features integrate well with mouse features to produce a best improvement of 16% in terms of CCR performance, whereas the other two combinations only produce about 10% improvement. This observation yields a slightly different conclusion based on absolute performance alone, which suggests that the model based on facial features combined with eye gaze features performs the best at 73.5% against 72.8% for facial features combined with mouse features. Nevertheless, this higher performance is attained at the expense of adopting the higher cost eye gaze feature set than the lower cost mouse feature set.

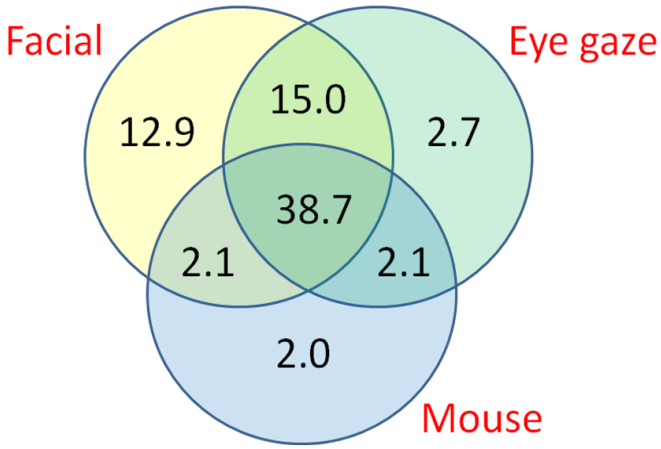


Figure 4-5. Improvement breakdown against models.

Let us make a simplifying assumption that all feature sets are somewhat synergic to one another, in order for us to take a glance on the contributions by the individual modalities. In other words, we assume that the models would not have a negative impact on one another when combined. The synergic effect is much higher than the interference effect. We can then attempt to break down for the individual contributions based on a simple additive model as shown in Figure 4-5. This provides us with a glance on the individual contribution to the overall performance. The more performance that can be "explained" by the overlapping part of two models, the more "similar" are the two sets of features and the higher possibility that the two models are making similar classification. As a result, there would be less additional improvement incurred in the multimodal model. Finally, it can be seen that any of the three models alone would produce an accuracy of close to 40%, which accounts for more than half of the attainable performance for the three models. Actually, this already represents the majority of the performance for the mouse feature model. This is a pretty high degree of "similarity" among the three individual models. Also, the "similarity" between facial feature model and eye gaze feature model is relatively high and this is understandable, as both come from the same video captured by the webcam.

Table 4-15. CCR improvement for individual modalities.

A+B	CCR_{A+B}	CCR_A	CCR_B	Δ_A	Δ_B	Δ
facial+eye gaze	73.5%	68.7%	58.5%	4.8%	15.0%	9.9%
facial+mouse	72.8%	68.7%	44.9%	4.1%	27.9%	16.0%
eye gaze+mouse	62.6%	58.5%	44.9%	4.1%	17.7%	10.9%

Table 4-16. CCR improvement for existing users.

Model	Facial	Eye gaze	Mouse	Multimodal
Leave-one-subject-out	68.7%	58.5%	44.9%	75.5%
All-subjects-included	70.1%	61.2%	45.6%	78.9%

Table 4-17. Normalized confusion matrix for existing users.

Ground truth \ Classified as	Low	Medium	High
Low	0.83	0.13	0.04
Medium	0.10	0.82	0.08
High	0.09	0.20	0.70

Table 4-18. Classification performance for existing users.

Attention Level \ Performance	Precision	Recall	F-measure
Low	0.83	0.83	0.83
Medium	0.72	0.82	0.77
High	0.84	0.70	0.77

4.4.6 Performance for Existing Users

So far in all our evaluations, we assume the setting of leave-one-subject-out for recognition performance to cater for unseen new users. It is also common in reality that the model is used by an existing user. One would expect that the accuracy will be higher. In our next experiment, we keep all subjects in the 10-fold cross-validation and compare the performance with the leave-one-subject-out setting as presented in Table 4-16. We observe a bit of improvement in terms of CCR. On the other hand, this small improvement also demonstrates that our approach is very robust, in delivering good performance even for unseen new users based on training data from just a small number of subjects ($n - 1 = 5$). Table 4-17 and Table 4-18 provide more information of the recognition performance of the multimodal model with 10-fold cross-validation. Compared with Table 4-12 and Table 4-13, we observe that there are improvements of the CCR for each class, and the misclassifying errors between neighboring classes and extreme classes are further decreased. Thus, our research represents a good initial

attempt to attention detection based on ubiquitous devices and a small number of features extracted from webcam videos and mouse activities.

4.5 Summary

In this chapter, we propose to recognize human attention level via the use of ubiquitous equipment loaded with most computers, namely, the mouse and the webcam. We extract facial features and eye gaze features from the videos captured by the webcam, as well as mouse dynamics due to mouse usage. We adopt machine learning techniques to model the captured data and build user-independent models capable of recognizing the attention level for unseen new users. We conduct our experiments via the reading tasks, with which the subjects are induced to different levels of attention, through the use of different vocal stimuli to distract them. Our results based solely on the webcam (i.e., facial features and eye gaze features) indicate good performance, and those solely on the simplistic mouse still achieve improvement over the baseline. We also demonstrate that combining the three sets of features together is giving us the best performance, whereas only 15 important features need to be utilized. Based on the evaluation results, we proceed to analyze more deeply the individual contribution by each modality and see which combinations would produce a better integrative performance.

Chapter 5

Other Relevant Contributions

In addition to the studies on the detection of comprehension and attention in reading as described in the previous chapters, there are some related projects arising over the course of my Ph.D. study. This chapter first depicts a cross-modal interaction system, which inspires us to understand how people perceive visual cues and interact with computing systems.

Motivated by our studies on multimodal affect detection in reading, we then present a cross-modal technique for stress detection from the coordination patterns between gaze and click. Compared to the simple aggregation of features or classification results from multiple modalities, we see that the detailed temporal and spatial alignments between gaze and click can be indicative of human mental state, and investigating such cross-modal patterns can enhance our understanding of human behaviors during interactions.

Finally, as mobile devices become ubiquitous, this chapter presents our study on gaze estimation on smartphones. Without relying on any specialized equipment, we propose a simple and effective technique, which leverages the reflection of the screen on the cornea for gaze estimation. It gives a promising performance and we see the great potential for in-situ analysis of gaze and reading behaviors on mobile devices.

5.1 CalliPaint

CalliPaint, a system for cross-modal art generation that links together Chinese ink brush calligraphy writing and Chinese landscape painting. We investigate the mapping between the two modalities based on concepts of metaphoric congruence, and implement our findings into a prototype system. A multi-step evaluation experiment with real users suggests that CalliPaint provides a realistic and intuitive experience that allows even novice users to create attractive landscape paintings from writing. Comparison with a general-purpose digital painting software suggests that CalliPaint provides users with a more enjoyable experience. Finally, exhibiting CalliPaint in an open-access location for use by casual users without any training shows that the system is easy to learn.

5.1.1 Methodology

Figure 5-1 illustrates the framework of the system. Essentially, the system performs a multidimensional mapping from Chinese calligraphy to Chinese ink paintings. When the user writes a character on the writing interface, a vision-based writing mechanics recognition model captures the stroke sequence and mechanics (pressure and speed) of the writing through a Microsoft Kinect depth camera. The character recognition model then matches the stroke sequence to a written character inside the vocabulary dictionary. If the written character is in the dictionary, the multidimensional mapping model then executes the mapping between character and image on physical, semantic and spatial levels. Finally, an image of the written object in the image dictionary that accords with the mapping rules will be selected to be the generated image and displayed in the painting. For coherence, if the character is traditionally associated with any allegorical aspects, global changes of existing objects will also be introduced into the painting based on the semantic mapping rules.

The following subsections present the details of the mapping process. We first introduce the vocabulary dictionary, which determines the objects supported by the system..Next, we cover the tri-level mapping from written character to object image. We finally describe the image dictionary, which determines the appearance and variation of the objects in the scene.

5.1.1.1 A vocabulary from characters to images

The main idea behind CalliPaint is the “translation”, so to speak, of a written character on a virtual parchment or writing surface into a pictorial image of the object that it represents in a scene. Since the focus of our system is not to perform a deep semantic

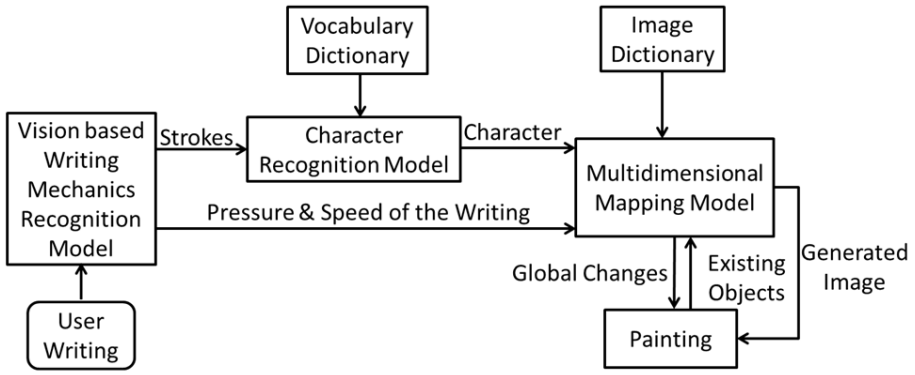


Figure 5-1. Framework of the CalliPaint System.

understanding of a written sentence or phrase, we use a simple 1-1 mapping of character to object, as defined by a pre-set dictionary.

To populate our dictionary, we collected a set of 128 well-known Chinese landscape paintings to serve as a representative sample of their genre[2]. We then enumerated and analyzed the objects depicted in these paintings. We found that though the paintings exhibit much complexity, the number of object *types* that they contained was relatively constrained: indeed, a small set of 17 object "types" is sufficiently rich (with variations) to fully characterize all objects in 90% of those paintings. For the remaining 10%, this set covers over 80% of all objects (Figure 5-2). Some of these objects also carry significant meanings or popularly understood symbolisms.

Given these findings, we incorporated these 17 objects into a pre-set dictionary for our system. These 17 objects are: mountain, river, tree, human, house, bird, land, boat, sun, moon, cloud, plum, bridge, bamboo, chrysanthemum, pine and willow.

5.1.1.2 A tri-level mapping from written character to object image

A character written on the CalliPaint writing surface is mapped into an object, and then the object's pictorial representation. This mapping of the character to an object is done on three levels: mechanical, semantic and spatial. The principles behind the mapping on each level are governed by rules of metaphoric congruence, conventions and physical properties.

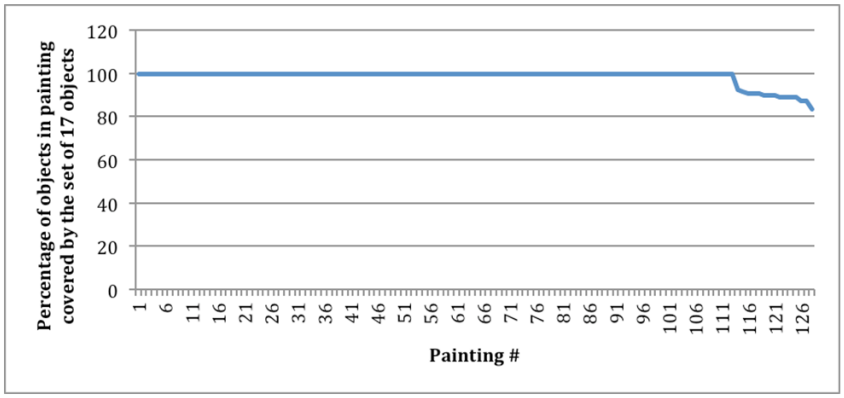


Figure 5-2. Coverage of objects in each painting by the set of 17 object types in the CalliPaint dictionary.

[2] These paintings are from the Song (960-1279) (21%), Yuan (1271-1368) (17%), Ming (1368-1644) (27%) and Qing (1644-1912) (35%) Dynasties, regarded as periods of high productivity in the development of Chinese landscape painting [1, 8]

The physical level mapping of CalliPaint links the mechanics of the written character with the appearance of the pictorial image. To arrive at an intuitive mapping between writing a character and drawing an image, we draw upon concepts in metaphorical congruence, which has been demonstrated by research in neuroscience to produce pleasing results in the processing of multi-modal sensory input in the brain [64][80], and deployed in MelodicBrush, which links Chinese calligraphy with music generation [51]. Among the physical mechanics that are used to write a calligraphic character using a soft brush, two of the most easily controlled and understood are the "firmness" of the stroke, which is roughly indicated by the pressure exerted upon the brush, and the speed with which the stroke is made. On the pictorial aspect, the most obvious characteristics of a given image are its level of detail, as defined by the granularity and the number of the strokes that make up the image, and its sharpness.

Given these writing mechanics, we hypothesize that there exists some correlation between the pressure and speed of the writing and the level of detail and sharpness of the painted objects. To test this hypothesis, we invited 12 subjects aged 21–30 to participate in experiments to map the writing mechanics to the appearance of the image. The results of the experiment show that a high pressure of writing is perceived as positively correlated to the level of detail in the image. In contrast, there is no clear correlation between the pressure of writing and the level of sharpness of the image, as the experiment subjects were as likely to correlate writing with a heavy hand with a blurred image as with a sharp one. It can also be seen that there is a clear correlation between the speed of writing and the sharpness of the image. In other words, most people associate writing quickly with a blurred image. In contrast, the correlation between speed of writing and image sharpness is quite ambiguous.

Our experiment therefore corroborates our hypothesis of the types of linkage between the writing mechanics of a written character and the appearance of the perceived associated image. Given these results, CalliPaint maps the pressure of writing to the level of detail of the image, and the speed of writing to the sharpness.

At the semantic level, the physical object is identified and inserted into the scene. The contextual level deploys the motifs and allegorical aspects of the objects in the painting, and allows certain objects to induce a change in the global picture. An obvious case is "sun" and "moon", which do not normally appear in the sky at the same time. Therefore, the writing of one will remove the other from the scene. Other examples are taken from common allegories in Chinese culture – such as certain plants (e.g., plum,

orchid, chrysanthemum and pine), which symbolize different seasons, and hence are not allowed to appear together.

The third level of our cross-modal mapping links the size and position of the written character with the spatial position and the image size of the object in the scene.

In our investigations of the representative set of 128 paintings, we observe that the composition of these paintings usually consists of three parts: the background, midground and foreground. The upper part of the scene usually contains objects that are distant and lofty, such as the sky, clouds and mountains. The midground usually contains mountains and lakes. The foreground is usually in the lower third of the canvas and depicts objects that are close to the viewer, such as trees and animate objects.

CalliPaint uses these principles of composition to map the size and the vertical coordinate of the written character to the spatial location of the image in the scene. Objects in the foreground occlude those in the background, as governed by the physical laws of optics, based on their vertical coordinate values. To allow for more diversity in the appearance of the objects, the physical size of the written character for inanimate objects (e.g. mountains and trees) is also mapped to the size of the image object.

The spatial mapping between written characters and the image representations of the objects allows users to construct their artwork without needing to be too concerned about the order in which the objects are placed into the picture, or with how newly placed objects should interact with the rest of the scene.

5.1.1.3 A dictionary of image variations

The image dictionary of the objects supported in CalliPaint is based upon the results of the object identification step and the metaphorical congruence mapping of the writing mechanics. At this stage, our primary focus is on facilitating and maximizing control and usability for the user. To that end, we built a core image library by selecting one reference image from a real painting in the public domain to represent each object in our dictionary. Each reference image is then further preprocessed with image editing software to achieve the desired effect corresponding to three levels of granularity both for writing pressure (light, normal and hard), and writing speed (slow, normal and fast).

5.1.2 System Interface and Implementation

The objective of CalliPaint's writing component is to enable the creation of a computer representation of the written calligraphy, and to capture the writing

mechanics for painting generation. To achieve this objective, CalliPaint borrows from previous approaches [45] that use the Microsoft Kinect depth camera to capture the writing mechanics of a user writing with a normal calligraphic brush, and to repurpose an ordinary LCD display panel as a writing surface.

Figure 5-3 shows the setup of our system. The Kinect depth camera captures the position and movement of a user and the position of the brush used. If the stroke sequence is recognized as a character, the strokes fade out and are replaced by the corresponding image. Non-recognized stroke sequences also fade out; here, the lack of a replacement image serves as user feedback. In addition to the recognition of characters as they are written, our system also provides three editing functions: delete, move and resize, which offers users more control over their painting.

5.1.3 Evaluation

The usability and functionality of CalliPaint was evaluated through two experiments. The first assesses CalliPaint as a tool for self-expression and education. The second assesses CalliPaint as a tool for entertainment in an unconstrained, unguided scenario.

The first experiment is carried out in a supervised/guided environment and consists of 4 stages with 10 subjects. The first stage evaluates the writing interface of the system. The second stage assesses the learnability, controllability and ease of use. A third stage investigates the ability of the system to create aesthetically-pleasing pictures. Conclusions drawn from these three stages are mainly from statistical analysis. For the sake of brevity, the details are not shown here. The fourth and final stage evaluate the degree to which the system supports and facilitates creativity. After that, the subjects were asked to complete an assessment form rating various aspects of the system on a scale from 1 to 5 (as in Table 5-1), and to participate in a post-experiment interview.

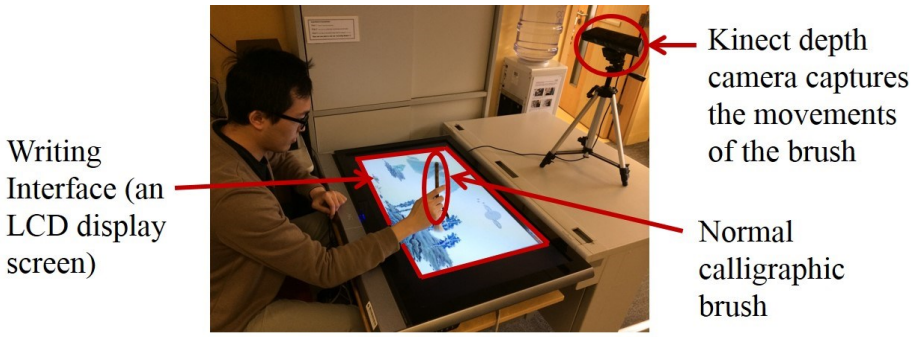


Figure 5-3. The System Interface for Callipaint. The writing mechanics are captured by the Kinect. The brush strokes are generated in real-time and converted to images.

Finally, two external experts in Chinese painting and design, respectively, evaluated the artwork created by the subjects.

The subjects' overall impression of CalliPaint was unanimously positive. All found the experience enjoyable. In the post-experiment feedback assessment, the subjects rated the overall experience an average of 3.9. The interface and the tool were found intuitive and the functions interesting and easy to use. Specifically, the effectiveness of the control functions was noted by the experiment subjects – most (70%) of them felt that the control functions was the most useful feature of CalliPaint, since it allowed them to achieve a fine-grained control over the graphical effects intuitively through controlling their writing style and mechanics. Many experiment subjects wanted to continue using the system after the experiment was over, and some bystanders asked if they could try to create their own artwork.

In contrast to the multistage experiment, which demonstrates that CalliPaint can be successfully used by novices with a minimum of training and guidance, the objective of the second experiment is to ascertain the potential of CalliPaint to be used by people without any training or guidance at all. For this experiment, we exhibited the system in a high-traffic student activity room in a local university for one week, during which it was available to numerous undergraduates and graduate students. No training or help was given except for a short introductory video playing in a loop on an accompanying display, a printout of the characters contained in the dictionary with the correct stroke order for each character, and a note stating the purpose of this display, informing users that their activities would be recorded for research, and inviting the reader to try it out.

This experiment shows that CalliPaint can be learned even without explicit training, by casual passers-by, and through experimentation and trial and error. Similarly to the

Table 5-1. Evaluation Feedback Results from Subjects in the Guided / Supervised Experiment.

Dimension of Evaluation	**Lowest rating (1) stands for:**	**Highest rating (5) stands for:**	**Average**
Overall Experience	Not enjoyable	Extremely enjoyable	3.9
Realism of interface	Artificial, contrived	Feels like the real thing	4.1
Realism of writing tool	Artificial, contrived	Feels like the real thing	4.3
Effectiveness in encouraging correct and standard practices of writing	Not helpful	Helpful	3.7
Ease of use (pressure control)	Difficult	Easy	3.9
Ease of use (speed control)	Difficult	Easy	3.7
Attractiveness of copy of real artwork	Ugly	Attractive	4.0
Attractiveness of self-generated artwork	Ugly	Attractive	3.7
Effectiveness in facilitating creativity	Not helpful	Helpful	4.4

guided experiments, we noted that in most of the cases where the user had difficulty with the system, the problems were usually caused by incorrect writing practices (e.g. wrong stroke sequence). In quite a few instances, students collaborated in figuring out the controls or "debugging" each other's mistakes.

5.1.4 Summary

This work presented CalliPaint, a Kinect-based novel digital system that generates Chinese ink painting from Chinese calligraphy. Images corresponding to the written characters are generated through a cross-modal mapping between the two art forms on both semantic level and mechanical level. Experimental evaluation with real users proved the intuitiveness, controllability, usability and potential to support creativity of our system. More details can be found in:

Jiajia Li , Grace Ngai, Stephen C.F. Chan, Kien A. Hua, Hong Va Leong, Alvin Chan. "From Writing to Painting: A Kinect-Based Cross-Modal Chinese Painting Generation System". *Proceedings of the 22nd ACM International Conference on Multimedia*. 57-66 (*MM'14*).

5.2 StressClick

Stress sensing is valuable in many applications, including online learning crowdsourcing and other daily human-computer interactions. Traditional affective computing techniques investigate affect inference based on different individual modalities, such as facial expression, vocal tones, and physiological signals or the aggregation of signals of these independent modalities, without explicitly exploiting their inter-connections. In contrast, we focus on exploring the impact of mental stress on the coordination between two human nervous systems, the somatic and autonomic nervous systems. Specifically, we present the analysis of the subtle but indicative pattern of human gaze behaviors surrounding a mouse-click event, i.e. the gaze-click pattern. Our evaluation shows that mental stress affects the gaze-click pattern, and this influence has largely been ignored in previous work. We, therefore, further propose a non-intrusive approach to inferring human stress level based on the gaze-click pattern, using only data collected from the common computer webcam and mouse. We conducted a human study on solving math questions under different stress levels to explore the validity of stress recognition based on this coordination pattern.

Experimental results show the effectiveness of our technique and the generalizability of the proposed features for user-independent modeling. Our results suggest that it may be possible to detect stress non-intrusively in the wild, without the need for specialized equipment.

5.2.1 Construct a Gaze-Click Dataset

In this study, we build a dataset that reliably captures human interactive behavior in stress and non-stress conditions under conditions that are comparable. According to previous research, recursive mental math calculation [2][63][105][111] and time pressure [54][112] are effective in inducing cognitive stress. We therefore select a math calculation task for evaluation.

We recruited 20 subjects (13 males, aged 20-33) for our study. In the experiment, the subjects are asked to calculate the results of math expressions and choose the correct answers by clicking. Our experiment looks at two distinct states: calm and stress. The calm session involves twenty 1-digit addition and subtraction questions. This is adjusted to 2-digit math for the stress session. To ensure the difficulty of the task, the numeric difference between the results of the two expressions is constrained to be no more than 10. To further induce stress, a countdown time bar is added to the bottom of the interface during the stress sessions. If the subject fails to answer within the allotted time, the interface advances automatically to the next question. Our experiment contains several sessions and there are 25 questions per session. Each subject is asked to report his/her level of stress on a 9-point scale at the end of each session. The score>=5 were annotated as “stress” and the rest as “calm”.

During the experiment, a standard off-the-shelf webcam placed on the top center of the monitor is used to capture the visual signal (resolution 640×480; 30fps) and a 22” monitor at 1680×1050 resolution displays the math interface in full screen mode. Removing the questions that the subjects fail to answer (<5%) in time gives us a total of 3818 click points over all subjects.

5.2.2 Gaze-Click Pattern Extraction and Evaluation

Our method uses a standard webcam to capture video of the user’s head and shoulders. To ensure that we can accurately deduce the gaze behavior of the subject from this video, we employ the Supervised Descent Method (SDM) [121] to track the facial landmarks and the eye CLMs [92] to track the eye landmarks from the frames of the video stream, respectively. These landmarks are then piped as input to a two-layer

Table 5-2. Description and mental state implication of gaze-click patterns extracted from the eye features and used in this work. All features are calculated relative to a given mouse click.

Index	Feature description	Mental state implication
g_1	Existence of a fixation in the 0.5s period preceding a click	
g_2	Duration of the fixation corresponding to a click	
g_3	Reaction Latency – duration of time in which eye remains fixated after the corresponding mouse click	Reaction latency of moving to the next task
g_4	Click Latency – duration of time between gaze moving away from target and corresponding mouse click	"Hastiness" of the user in locating the next target.
g_5	Max gaze velocity between the fixation corresponding to the mouse click and the fixation before it.	
g_6	Duration between the fixation corresponding to the mouse click and the fixation before it.	
g_7	Max gaze velocity between the fixation corresponding to the mouse click and the fixation after it.	
g_8	Duration between the fixation corresponding to the mouse click and the fixation before it.	

feature extraction mechanism. The first layer continuously extracts six *eye features* from the changes of the eye-related landmarks. The second layer is triggered by each mouse-click event, whereupon it extracts eight *gaze-click features* (Table 5-2) based on the eye feature signals in the 3-second time window surrounding the click.

Our evaluations on the features indicate that the gaze behaviors surrounding a mouse-click event show a certain degree of connection with the changes of mental stress. Although the connection appears slightly ambiguous between an individual gaze-click feature and the human stress level, it can be indicative for discriminating stress and calm given a proper model exploring the relation with multiple features.

In order to fully investigate the generalizability of the proposed features, we model stress from the eight gaze-click features by adopting the random forest algorithm [15]. For the within-subjects evaluation, we build a user-dependent model for each individual subject and use 10-fold cross-validation for evaluation. The final performance is the average performance across all folds. The between-subjects study employs leave-one-subject-out cross-validation, testing on each subject in turn. For reference, we provide

Table 5-3. Performance comparison of click-level detection between user-dependent and independent models.

	User-Dependent Model	User-Independent Model
CCR	65.72 (51.05)	60.27 (51.1)
F1	0.66 (0.36)	0.60 (0.35)
AUC	0.70 (0.5)	0.63 (0.5)

Numbers in parentheses denote the baseline performances. The performance is the weighted average results across different subjects, whose values may have slight difference from the average, due to the differences of data amount of subjects.

Table 5-4. Performance comparison of session-level detection between user-dependent and independent models.

	User-Dependent Model	User-Independent Model
CCR	74.0 (49.35)	80.5 (50.0)
F1	0.74 (0.33)	0.79 (0.33)
AUC	0.73 (0.5)	0.89 (0.5)

Numbers in parentheses denote the baseline performances.

the baseline performance given by a naïve classifier that predicts the majority class in the training set. Table 5-3 summarizes the performance comparisons of both user-dependent and independent models to the baselines. Weighted average of correct classification rate (CCR), F1-measure, and area under the receiver operating characteristic curve (AUC) across subjects are used as performance metrics. The comparison shows that our models significantly outperform the naïve classifiers.

In addition, we introduce a 2nd-layer classifier to recognize session-level stress by constructing 3 features based on the click-level predictions mentioned above. (1) number of StressClicks, (2) number of clicks being considered, and (3) the ratio of StressClicks to the total number of clicks. Given the simplicity of the 2nd-layer features, a logistic classifier [21] is used for stress detection from multiple clicks. Table 5-4 summarizes the performances of the user- dependent and independent models for the session-level prediction. It is very encouraging that the session-level user-independent model (F1=0.79) outperforms the user-dependent model (F1=0.74), given enough click data. Furthermore, as expected, session-level user- dependent and independent models outperform their click-level counterparts, with 8.31% and 20.23% CCR improvements, respectively.

5.2.3 Summary

This study presents a technique that aims to non-intrusively detect user stress through the gaze-click pattern. Using a series of multi-user experiments, we empirically demonstrate the impact of stress on the gaze-click pattern, which has been largely ignored in previous work. We also propose the cross-modal gaze-click features for stress recognition and investigate their effectiveness in both user- dependent and independent studies. Our results show that not only is it feasible to detect user stress

through non-intrusively collected data, but also that our proposed features are generalizable across different users. For more information, see:

Michael Xuelin Huang, **Jiajia Li**, Grace Ngai, Hong Va Leong. "StressClick: Sensing Stress from Gaze-Click Patterns". *Proceedings of the 24th ACM International Conference on Multimedia.* 1395-1404 *(MM'16)*.

5.3 ScreenGlint

Gaze estimation has widespread applications in HCI and numerous other domains. However, little work has explored the gaze estimation on the smartphone platform. This study presents ScreenGlint that exploits the glint (reflection) of the screen for gaze estimation using the unmodified camera.

To understand the gaze learning based on the screen glint and the related human behaviors, we first conducted a user study on common postures of the smartphone use with 6 subjects. From this study, the normal range of the face-to-screen distance is found to be around 20~40 cm. We also investigate the impact of illumination and distance on the size of the glint. We find that the glint generally appears smaller in a brighter environment and grows bigger as the increase of the pupillary distance, i.e. the decrease of the face-to-screen distance.

Based on this finding, we design an experiment to evaluate ScreenGlint in different face-to-screen distances. A dataset of 18 subjects was collected for our performance study. An in-depth evaluation is given and the impact of head pose variations is discussed. ScreenGlint achieves an overall angular error of 2.44º without head pose variations, and 2.94º with head pose variations. Our technique compares favorably to the state-of-the-art, indicating that the glint of the screen is an effective and practical cue to gaze estimation on the smartphone platform. Therefore, it opens a new venue for the gaze-aware applications, especially in the reading contexts. The system is described in the following publication:

Michael Xuelin Huang, **Jiajia Li**, Grace Ngai, Hong Va Leong. "ScreenGlint: A Practical Cue to Gaze Estimation on Smartphones". *To appear in Proceedings of the 34th CHI Conference on Human Factors in Computing Systems (CHI'17)*.

Chapter 6

Conclusions and Future Work

This chapter draws conclusions on the thesis, and points out some possible research directions related to the work in this thesis.

6.1 Conclusions

Reading is one of the most commonly occurring tasks in HCI. This thesis focuses on the detection of comprehension level and attention level in reading. We address three challenges. First, we conduct detection in a non-intrusive manner. This is difficult because many human signals can only be precisely measured using specialized equipment. For example, observing detailed eye movements needs electrooculography (EOG) systems, and sensing complex brainwave signals requires electroencephalography (EEG) devices. Second, we perform a thorough study on eye gaze patterns in a ubiquitous and efficient fashion, without relying on the analysis of the lexical and linguistic variables. Third, we propose reliable methods with good generalizability. Some related research explores user-dependent models for affect detection. Those techniques may not be able to accommodate unseen users, and training a user-specific model may oftentimes be impractical in real-use situations.

This thesis investigates the reading comprehension detection based on the analysis of eye gaze behaviors. We use a commercial eye tracker to capture the sequential gaze locations, from which we extract feature representation of gaze behaviors and understand its indicativeness for comprehension detection in reading. Our approach is effective to identify when the readers are experiencing difficulties in understanding their reading material. Overall, our approach is able to achieve a performance improvement of over 30% above baseline, translating to more than 50% reduction in detection error. It is also found to be robust for new unseen users.

Our contributions of this work can be summarized as:

- We investigate the detection of comprehension level based on a commonly occurring task, i.e. reading;
- We identify good features that are effective in describing specific eye movement behaviors when readers are exhibiting different levels of comprehension during the reading task;
- We apply machine learning techniques to build user-independent models to recognize the level of comprehension in reading tasks;

- We conduct experiments with human subjects to evaluate the accuracy of our approach in various context.

As a follow-up study of the gaze behaviors in reading comprehension using specialized eye tracker, we investigate a multimodal approach to the detection of human attention level in reading using only off-the-shelf devices. Specifically, we use webcam to capture facial expressions and eye movements, and mouse for mouse dynamics. Signals from these modalities are fused together for human attention level detection. Our results indicate performance improvement with multimodal inputs from webcam and mouse over that of a single modality.

Our contributions of this work can be summarized as:

- We investigate human attention level detection based on a most commonly occurring task, i.e., reading, without the use of sophisticated or intrusive devices;
- We adopt multimodal input processing to extract human facial features, eye gaze features and mouse dynamics;
- We apply machine learning techniques to build up user-independent models to recognize human attention level in reading tasks;
- We conduct experiments with human subjects to evaluate the accuracy of our approach. We believe that our work opens up a useful approach for interesting future user-computer interaction applications, for instance, in e-Learning.

6.2 Limitations

This thesis investigates multimodal comprehension and attention detection during reading. The experimental results are promising, however, there are still some limitations of the related studies.

First, the studies need to be extended to bigger datasets with more subjects. Our current datasets are quite small (41 instances of 10 subjects for comprehension detection and 147 instances of 6 subjects for attention detection) and all the subjects are students. Considering the machine learning algorithms adopted in the studies, training sets are critical to build robust and effective models. On the other hand, the diversity of subjects is also important, otherwise we risk building methods that work only on specific groups of people. Previous research [56] has demonstrated that visual scene processing changes with age. People in different age groups have their own focus of

visual attention and visual behaviors. It will be interesting and meaningful to investigate how the reading comprehension and attention are related with other human factors. Our methods can be improved to be more ubiquitous for real use applications by addressing these issues.

Second, the methods proposed in this thesis can be continuously studied and improved. Our current research focuses on comprehension and attention detection during reading English articles which are all text-based, which is one of the most commonly used reading materials. The human behaviors investigated in this thesis are also based on the article reading tasks. However, there are various kinds of reading materials, such as websites, comics and paintings, used in real life, and human behaviors while processing different reading materials may change accordingly. To apply our methods to other reading scenarios may require further study on the reading behaviors and extracting other helpful features not mentioned in this thesis, such as salient region coordinates and eye gaze transitions.

Another issue that is worth noticing is deeper understanding of the reason behind the human behaviors. In our studies, we adopted the useful features selected by feature selection algorithms. However, investigating the unselected features may give us more inspiration to understand the research problems. We find we ignored some factors which cause the failure of some features that are supposed to be useful. For example, to some extent, the feature of rate of eye blink doesn't help in the reading comprehension detection because of the eye fatigue. The mixture of the human mental states or affects and the physical conditions of subjects should be considered in current and future studies.

6.3 Future Work

6.3.1 Transfer/Customization of User-Independent Models

In this thesis, we make efforts to extract features that reflect the commonality between different users without figuring out the difference between the users' behavior patterns. In the future, we would like to study *transfer* or *customization* of user-independent models for specific users. The prediction of the specific user will be the weighted average of the predictions of other users' models. The weightings should be assigned according to the similarity between behavior patterns of the users. We expect such

models to require less user-specific data than pure user-dependent models. We also intend to study the improvement of performance as a function of increasing user data, and to employ machine learning methods such as online learning for this task.

6.3.2 User-Dependent Affect Detection in Reading

This thesis focuses on building user-independent models for the affect detection in reading, which makes our methods work for unseen new users. We foresee, however, that the user-dependent model may achieve better performance for the consistency of individual user's behavior patterns during the task. In the future, we would like to study the improvement by means of user-dependent models, upon collecting a larger dataset.

6.3.3 Extended Study to Other Contexts

In addition to the affect detection in the reading tasks, we would like to expand our scope of investigation to other language-based interaction tasks, such as writing and editing, which, together with reading, makes up a large proportion of computer usage, especially in the workforce. This would likely mean the use of other modalities, such as keyboard and mouse dynamics, and necessitate a multimodal approach.

References

1. Piotr D. Adamczyk and Brian P. Bailey. 2004. If not now, when? *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, ACM Press, 271–278. http://doi.org/10.1145/985692.985727
2. Jonathan Aigrain, Severine Dubuisson, Marcin Detyniecki, and Mohamed Chetouani. 2015. Person-specific behavioural features for automatic stress detection. *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. http://doi.org/10.1109/FG.2015.7284844
3. Zara Ambadar, Jeffrey F. Cohn, and Lawrence Ian Reed. 2009. All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous. *Journal of Nonverbal Behavior* 33, 1: 17–34. http://doi.org/10.1007/s10919-008-0059-5
4. Amy R Anderson, Sandra L Christenson, Mary F Sinclair, and Camilla A Lehr. 2004. Check & Connect: The importance of relationships for promoting engagement with school. *Journal of School Psychology* 42, 2: 95–113. http://doi.org/10.1016/j.jsp.2004.01.002
5. Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, et al. 2009. The Painful Face - Pain Expression Recognition Using Active Appearance Models. *Image and vision computing* 27, 12: 1788–1796. http://doi.org/10.1016/j.imavis.2009.05.007
6. C. Avila, A. Furnham, and A. McClelland. 2012. The influence of distracting familiar vocal music on cognitive performance of introverts and extraverts. *Psychology of Music* 40, 1: 84–93. http://doi.org/10.1177/0305735611422672
7. R. Barea, L. Boquete, M. Mazo, and E. Lopez. 2002. System for assisted mobility using eye movements based on electrooculography. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 10, 4: 209–218. http://doi.org/10.1109/TNSRE.2002.806829
8. Luciane Baretta, Lêda Maria Braga Tomitch, Vanessa Kwan Lim, and Karen E. Waldie. 2012. Investigating reading comprehension through EEG. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, 63. http://doi.org/10.5007/2175-8026.2012n63p69

9. David Beymer and Daniel M. Russell. 2005. WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze. *CHI '05 extended abstracts on Human factors in computing systems - CHI '05*, ACM Press, 1913–1916. http://doi.org/10.1145/1056808.1057055
10. M. E Bitterman. 1947. Frequency of blinking in visual work: a reply to Dr. Luckiesh. *Journal of Experimental Psychology* 37, 3: 269–270.
11. M. E. Bitterman. 1945. Heart rate and frequency of blinking as indices of visual efficiency. *Journal of Experimental Psychology* 35, 4: 279–292.
12. Robert Bixler and Sidney D'Mello. 2013. Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13*, ACM Press, 225–234. http://doi.org/10.1145/2449396.2449426
13. Robert Bixler and Sidney D'Mello. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction* 26, 1: 33–68. http://doi.org/10.1007/s11257-015-9167-1
14. Richard A. Bolt. 1980. "Put-that-there." *Proceedings of the 7th annual conference on Computer graphics and interactive techniques - SIGGRAPH '80*, ACM Press, 262–270. http://doi.org/10.1145/800250.807503
15. Leo Breiman. 2001. Random forests. *Machine Learning* 45: 5–32. http://doi.org/10.1023/A:1010933404324
16. X.L.C. Brolly and J.B. Mulligan. Implicit Calibration of a Remote Gaze Tracker. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, IEEE, 134–134. http://doi.org/10.1109/CVPR.2004.366
17. L. Brown, B. Grundlehner, and J. Penders. 2011. Towards wireless emotional valence detection from EEG. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2188–2191. http://doi.org/10.1109/IEMBS.2011.6090412
18. Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4: 741–753. http://doi.org/10.1109/TPAMI.2010.86
19. Rafael A Calvo and Sidney D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE*

Transactions on Affective Computing 1, 1: 18–37. http://doi.org/10.1109/T-AFFC.2010.1

20. Stefano Carrino, Alexandre Péclat, Elena Mugellini, Omar Abou Khaled, and Rolf Ingold. 2011. Humans and smart environments: a novel multimodal interaction approach. *Proceedings of the 13th international conference on multimodal interfaces - ICMI '11*, ACM Press, 105–112. http://doi.org/10.1145/2070481.2070501
21. S. le Cessie and J.C. van Houwelingen. 1992. Ridge Estimators in Logistic Regression. *Applied Statistics* 41, 1: 191–201.
22. Han Collewijn, Robert M. Steinman, Casper J. Erkelens, Zygmunt Pizlo, and Johannes Van Der Steen. 1992. Effect of Freeing the Head on Eye Movement Characteristics during Three-Dimensional Shifts of Gaze and Tracking. In *The Head-Neck Sensory Motor System*. Oxford University Press, 412–418. http://doi.org/10.1093/acprof:oso/9780195068207.003.0064
23. Douglas W. Cunningham, Mario Kleiner, Heirich H. Bülthoff, and Christian Wallraven. 2004. The components of conversational facial expressions. *Proceedings of the 1st Symposium on Applied perception in graphics and visualization - APGV '04*, ACM Press, 143–150. http://doi.org/10.1145/1012551.1012578
24. Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies* 70, 5: 377–398. http://doi.org/10.1016/j.ijhcs.2012.01.004
25. M. Divjak and H. Bischof. 2009. Eye blink based fatigue detection for prevention of computer vision syndrome. *IAPR Conference on Machine Vision Applications*, 350–353.
26. Paul Ekman. 1972. Universals and Cultural Differences in Facial Expression of Emotion. *Nebraska Symposium on Motivation*, 207–283.
27. Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, Anthony Chan, and et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* 53, 4: 712–717. http://doi.org/10.1037/0022-3514.53.4.712
28. Tom Foulsham, James Farley, and Alan Kingstone. 2013. Mind wandering in sentence reading: Decoupling the link between mind and eye. *Canadian*

Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 67, 1: 51–59. http://doi.org/10.1037/a0030217

29. Annie Beth Fox, Jonathan Rosen, and Mary Crawford. 2009. Distractions, Distractions: Does Instant Messaging Affect College Students' Performance on a Concurrent Reading Comprehension Task? *CyberPsychology & Behavior* 12, 1: 51–53. http://doi.org/10.1089/cpb.2008.0107
30. John M. Franchak, Kari S. Kretch, Kasey C. Soska, and Karen E. Adolph. 2011. Head-Mounted Eye Tracking: A New Method to Describe Infant Looking. *Child Development* 82, 6: 1738–1750. http://doi.org/10.1111/j.1467-8624.2011.01670.x
31. Yujun Fu, Hong Va Leong, Grace Ngai, Michael Xuelin Huang, and Stephen C.F. Chan. 2014. Physiological Mouse: Towards an Emotion-Aware Mouse. *2014 IEEE 38th International Computer Software and Applications Conference Workshops*, IEEE, 258–263. http://doi.org/10.1109/COMPSACW.2014.46
32. Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, ACM Press, 478–479. http://doi.org/10.1145/1008992.1009079
33. Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2008. Multi-PIE. *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, IEEE, 1–8. http://doi.org/10.1109/AFGR.2008.4813399
34. E.D. Guestrin and M. Eizenman. 2006. General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections. *IEEE Transactions on Biomedical Engineering* 53, 6: 1124–1133. http://doi.org/10.1109/TBME.2005.863952
35. H. Gunes and M. Piccardi. 2009. Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1: 64–84. http://doi.org/10.1109/TSMCB.2008.927269
36. Hatice Gunes and Massimo Piccardi. 2005. Fusing Face and Body Display for Bi-modal Emotion Recognition: Single Frame Analysis and Multi-frame Post Integration. . 102–111. http://doi.org/10.1007/11573548_14

37. Hatice Gunes and Massimo Piccardi. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications* 30, 4: 1334–1345. http://doi.org/10.1016/j.jnca.2006.09.007
38. D.W. Hansen and Qiang Ji. 2010. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3: 478–500. http://doi.org/10.1109/TPAMI.2009.30
39. Dan Witzner Hansen and Arthur E.C. Pece. 2005. Eye tracking in the wild. *Computer Vision and Image Understanding* 98, 1: 155–181. http://doi.org/10.1016/j.cviu.2004.07.013
40. John M. Henderson and Andrew Hollingworth. 1998. Eye movements during scene viewing: an overview. In *Eye Guidance in Reading and Scene Perception*, G. Underwood (ed.). Elsevier Science Ltd, 269–293.
41. Michael Huang, Grace Ngai, Kien Hua, Stephen Chan, and Hong Va Leong. 2015. Identifying User-Specific Facial Affects from Spontaneous Expressions with Minimal Annotation. *IEEE Transactions on Affective Computing* 7, 4: 360–373. http://doi.org/10.1109/TAFFC.2015.2495222
42. Michael Xuelin Huang, Tiffany C.K. Kwok, Grace Ngai, Stephen C.F. Chan, and Hong Va Leong. 2016. Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, ACM Press, 5169–5179. http://doi.org/10.1145/2858036.2858404
43. Michael Xuelin Huang, Tiffany C.K. Kwok, Grace Ngai, Hong Va Leong, and Stephen C.F. Chan. 2014. Building a Self-Learning Eye Gaze Model from User Interaction Data. *Proceedings of the ACM International Conference on Multimedia - MM '14*, ACM Press, 1017–1020. http://doi.org/10.1145/2647868.2655031
44. Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2016. StressClick: Sensing Stress from Gaze-Click Patterns. *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, ACM Press, 1395–1404. http://doi.org/10.1145/2964284.2964318
45. Michael Xuelin Huang, Will W. W. Tang, Kenneth W. K. Lo, C. K. Lau, Grace Ngai, and Stephen Chan. 2012. MelodicBrush: a novel system for cross-modal digital art creation linking calligraphy and music. *Proceedings of the Designing*

Interactive Systems Conference on - DIS '12, ACM Press, 418–427. http://doi.org/10.1145/2317956.2318018

46. Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. 2015. TabletGaze: Unconstrained Appearance-based Gaze Estimation in Mobile Tablets. *arXiv*.
47. Albrecht Werner Inhoff and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics* 40, 6: 431–439. http://doi.org/10.3758/BF03208203
48. Qiang Ji, P. Lan, and C. Looney. 2006. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 36, 5: 862–875. http://doi.org/10.1109/TSMCA.2005.855922
49. Jia Jia, Zhiyong Wu, Shen Zhang, Helen M. Meng, and Lianhong Cai. 2014. Head and facial gestures synthesis using PAD model for an expressive talking avatar. *Multimedia Tools and Applications* 73, 1: 439–461. http://doi.org/10.1007/s11042-013-1604-8
50. Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87, 4: 329–354. http://doi.org/10.1037/0033-295X.87.4.329
51. Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. 2010. Face-TLD: Tracking-Learning-Detection applied to faces. *2010 IEEE International Conference on Image Processing*, IEEE, 3789–3792. http://doi.org/10.1109/ICIP.2010.5653525
52. M. J. Kane, L. H. Brown, J. C. McVay, P. J. Silvia, I. Myin-Germeys, and T. R. Kwapil. 2007. For Whom the Mind Wanders, and When: An Experience-Sampling Study of Working Memory and Executive Control in Daily Life. *Psychological Science* 18, 7: 614–621. http://doi.org/10.1111/j.1467-9280.2007.01948.x
53. M. A. Killingsworth and D. T. Gilbert. 2010. A Wandering Mind Is an Unhappy Mind. *Science* 330, 6006: 932–932. http://doi.org/10.1126/science.1192439
54. Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A Neerincx, and Wessel Kraaij. 2014. The SWELL Knowledge Work Dataset for Stress and User

Modeling Research. *Proceedings of the 16th International Conference on Multimodal Interaction*, 291–298. http://doi.org/10.1145/2663204.2663257

55. Kyle Krafka, Aditya Khosla, Petr Kellnhofer, and Harini Kannan. 2016. Eye Tracking for Everyone. *IEEE Conference on Computer Vision and Pattern Recognition*, 2176–2184. http://doi.org/10.1109/CVPR.2016.239
56. Onkar Krishna, Kiyoharu Aizawa, Andrea Helo, and Rama Pia. 2017. Gaze Distribution Analysis and Saliency Prediction Across Age Groups. Retrieved from http://arxiv.org/abs/1705.07284
57. Rainie Lee, Zickuhr Kathryn, Purcell Kristen, Madden Mary, and Brenner Joanna. 2012. *The rise of e-reading*. Washington, D.C. Retrieved from http://libraries.pewinternet.org/2012/04/04/the-rise-of-e-reading/
58. Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen C.F. Chan. 2016. Multimodal human attention detection for reading. *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16*, ACM Press, 187–192. http://doi.org/10.1145/2851613.2851681
59. Simon P Liversedge and John M Findlay. 2000. Saccadic eye movements and cognition. *Trends in cognitive sciences* 4, 1: 6–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10637617
60. Feng Lu, Takahiro Okabe, Yusuke Sugano, and Yoichi Sato. 2014. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing* 32, 3: 169–179. http://doi.org/10.1016/j.imavis.2014.01.005
61. Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2011. Inferring human gaze from appearance via adaptive linear regression. *2011 International Conference on Computer Vision*, IEEE, 153–160. http://doi.org/10.1109/ICCV.2011.6126237
62. Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2015. Gaze Estimation From Eye Appearance: A Head Pose-Free Method via Eye Image Synthesis. *IEEE Transactions on Image Processing* 24, 11: 3680–3693. http://doi.org/10.1109/TIP.2015.2445295
63. U Lundberg, R Kadefors, B Melin, et al. 1994. Psychophysiological stress and EMG activity of the trapezius muscle. *International journal of behavioral medicine* 1: 354–370. http://doi.org/10.1207/s15327558ijbm0104_5
64. Daphne Maurer, Thanujeni Pathman, and Catherine J. Mondloch. 2006. The shape of boubas: sound-shape correspondences in toddlers and adults.

Developmental Science 9, 3: 316–322. http://doi.org/10.1111/j.1467-7687.2006.00495.x

65. Iris B. Mauss and Michael D. Robinson. 2009. Measures of emotion: A review. *Cognition & Emotion* 23, 2: 209–237. http://doi.org/10.1080/02699930802204677
66. Carlos H. Morimoto and Marcio R.M. Mimica. 2005. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding* 98, 1: 4–24. http://doi.org/10.1016/j.cviu.2004.07.010
67. Robert E. Morrison. 1983. Retinal image size and the perceptual span in reading. In *Eye Movements in Reading: Perceptual and Language Processes*. Academic Press, New York, 31–40.
68. Wayne S. Murray and Alan Kennedy. 1988. Spatial coding in the processing of anaphor by good and poor readers: Evidence from eye movement analyses. *The Quarterly Journal of Experimental Psychology Section A* 40, 4: 693–718. http://doi.org/10.1080/14640748808402294
69. Sharon L Oviatt. 2007. Multimodal interfaces. In *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. L. Erlbaum Associates Inc, Hillsdale, NJ, USA, 286–304.
70. D Palomba, M Sarlo, A Angrilli, A Mini, and L Stegagno. 2000. Cardiac responses associated with affective processing of unpleasant film stimuli. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 36, 1: 45–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10700622
71. Jaak Panksepp. 2007. Neuro-Psychoanalysis May Enliven the Mindbrain Sciences. *Cortex* 43, 8: 1106–1107. http://doi.org/10.1016/S0010-9452(08)70714-7
72. M. Pantic and I. Patras. 2006. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)* 36, 2: 433–449. http://doi.org/10.1109/TSMCB.2005.859075
73. Maja Pantic, Nicu Sebe, Jeffrey F. Cohn, and Thomas Huang. 2005. Affective multimodal human-computer interaction. *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*, ACM Press, 669–676. http://doi.org/10.1145/1101149.1101299

74. Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, to appear.
75. Phillip T. Pasqual and Jacob O. Wobbrock. 2014. Mouse pointing endpoint prediction using kinematic template matching. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, ACM Press, 743–752. http://doi.org/10.1145/2556288.2557406
76. Amol Patwardhan and Gerald Knapp. 2016. Multimodal Affect Recognition using Kinect. Retrieved from http://arxiv.org/abs/1607.02652
77. Rosalind Wright Picard. 1997. *Affective computing*. MIT Press, Cambridge, MA, USA.
78. Alexander Poole. 2005. Gender Differences in Reading Strategy Use among ESL College Students. *Journal of College Reading and Learning* 36, 1: 7–20. http://doi.org/10.1080/10790195.2005.10850177
79. Ralph Radach and Alan Kennedy. 2004. Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European Journal of Cognitive Psychology* 16, 1–2: 3–26. http://doi.org/10.1080/09541440340000295
80. V.S Ramachandran and E.M Hubbard. 2001. Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies* 8, 12: 3–34.
81. K Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3: 372–422. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9849112
82. K Rayner, S C Sereno, and G E Raney. 1996. Eye movement control in reading: a comparison of two types of models. *Journal of experimental psychology. Human perception and performance* 22, 5: 1188–1200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8865619
83. Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology* 62, 8: 1457–1506. http://doi.org/10.1080/17470210902816461

84. Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14, 3: 191–201. http://doi.org/10.3758/BF03197692
85. Keith Rayner, Gretchen Kambe, and Susan A. Duffy. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A* 53, 4: 1061–1080. http://doi.org/10.1080/713755934
86. Keith Rayner and George W. McConkie. 1976. What guides a reader's eye movements? *Vision Research* 16, 8: 829–837. http://doi.org/10.1016/0042-6989(76)90143-7
87. Keith Rayner and Arnold D. Well. 1996. Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review* 3, 4: 504–509. http://doi.org/10.3758/BF03214555
88. E. D. Reichle, A. E. Reineberg, and J. W. Schooler. 2010. Eye Movements During Mindless Reading. *Psychological Science* 21, 9: 1300–1310. http://doi.org/10.1177/0956797610378686
89. Mathieu Rodrigue, Jungah Son, Barry Giesbrecht, Matthew Turk, and Tobias Höllerer. 2015. Spatio-Temporal Detection of Divided Attention in Reading Applications Using EEG and Eye Tracking. *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, ACM Press, 121–125. http://doi.org/10.1145/2678025.2701382
90. Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*, ACM Press, 71–78. http://doi.org/10.1145/355017.355028
91. Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2010. Deformable Model Fitting by Regularized Landmark Mean-Shift. *International Journal of Computer Vision* 91, 2: 200–215. http://doi.org/10.1007/s11263-010-0380-4
92. Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 2009. Face alignment through subspace constrained mean-shifts. *2009 IEEE 12th International Conference on Computer Vision*, IEEE, 1034–1041. http://doi.org/10.1109/ICCV.2009.5459377

93. R. Schleicher, N. Galley, S. Briest, and L. Galley. 2008. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics* 51, 7: 982–1010. http://doi.org/10.1080/00140130701817062

94. Nicu Sebe, Ira Cohen, and Thomas S Huang. 2005. Multimodal Emotion Recognition. In *Handbook of Pattern Recognition and Computer Vision*. World Scientific.

95. Tayyar Sen and Ted Megaw. 1984. The Effects of Task Variables and Prolonged Performance on Saccadic Eye Movement Parameters. . 103–111. http://doi.org/10.1016/S0166-4115(08)61824-5

96. Tayyar Sen and Ted Megaw. 1984. The Effects of Task Variables and Prolonged Performance on Saccadic Eye Movement Parameters. In *Theoretical and Applied Aspects of Eye Movement Research*, A. G. Gale and F. Johnson (eds.). Elsevier, 103–111. http://doi.org/10.1016/S0166-4115(08)61824-5

97. Caifeng Shan, Shaogang Gong, and Peter W. McOwan. 2009. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing* 27, 6: 803–816. http://doi.org/10.1016/j.imavis.2008.08.005

98. Tan Shawna C.G and Nareyek Alexander. 2009. Integrating Facial, Gesture, and Posture Emotion Expression for a 3D Virtual Agent. *Proceedings of the 14th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games.*, 23–31.

99. Sheng-Wen Shih, Yu-Te Wu, and Jin Liu. A calibration-free gaze tracking technique. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, IEEE Comput. Soc, 201–204. http://doi.org/10.1109/ICPR.2000.902895

100. John L. Sibert, Mehmet Gokturk, and Robert A. Lavine. 2000. The reading assistant. *Proceedings of the 13th annual ACM symposium on User interface software and technology - UIST '00*, ACM Press, 101–107. http://doi.org/10.1145/354401.354418

101. John A. Stern, Larry C. Walrath, and Robert Goldstein. 1984. The Endogenous Eyeblink. *Psychophysiology* 21, 1: 22–33. http://doi.org/10.1111/j.1469-8986.1984.tb02312.x

102. John A Stern and Douglas N Dunham. 1990. The ocular system. In *Principles of psychophysiology: Physical, social, and inferential elements*. Cambridge University Press, 513–553.
103. Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2013. Appearance-based gaze estimation using visual saliency. *IEEE transactions on pattern analysis and machine intelligence* 35, 2: 329–341. http://doi.org/10.1109/TPAMI.2012.101
104. Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato, and Hideki Koike. 2015. Appearance-Based Gaze Estimation With Online Calibration From Mouse Operations. *IEEE Transactions on Human-Machine Systems* 45, 6: 750–760. http://doi.org/10.1109/THMS.2015.2400434
105. David Sun, Pablo Paredes, and John Canny. 2014. MouStress: Detecting Stress from Mouse Motion. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, 61–70. http://doi.org/10.1145/2556288.2557243
106. Luis Talavera. 2005. An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering. In *Advances in Intelligent Data Analysis VI*, A. Famili, J. Kok, J. Pena, A. Siebes and A. Feelders (eds.). Springer Berlin Heidelberg, 440–451. http://doi.org/10.1007/11552253_40
107. Tobii X1 Light Eye Tracker. Specification of Gaze Precision and Gaze Accuracy. Retrieved from http://www.tobiipro.com/siteassets/tobii-pro/technical-specifications/tobii-pro-x2-60-technical-specification.pdf
108. Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. 2012. Gaze guided object recognition using a head-mounted eye tracker. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, ACM Press, 91–98. http://doi.org/10.1145/2168556.2168570
109. Filareti Tsalakanidou and Sotiris Malassiotis. 2010. Real-time 2D+3D facial action and expression recognition. *Pattern Recognition* 43, 5: 1763–1775. http://doi.org/10.1016/j.patcog.2009.12.009
110. Georgios Tsoulouhas, Dimitrios Georgiou, and Alexandros Karakos. 2011. Detection of Learners' Affective State Based on Mouse Movements. *Journal of Computing* 3, 11: 9–18.
111. Lisa M. Vizer, Lina Zhou, and Andrew Sears. 2009. Automated stress detection using keystroke and linguistic features: An exploratory study.

International Journal of Human-Computer Studies 67, 10: 870–886. http://doi.org/10.1016/j.ijhcs.2009.07.005

112. J. Wahlstrom, M. Hagberg, P. W. Johnson, J. Svensson, and D. Rempel. 2002. Influence of time pressure and verbal provocation on physiological and psychological reactions during work with a computer mouse. *European Journal of Applied Physiology* 87: 257–263. http://doi.org/10.1007/s00421-002-0611-7
113. Kang Wang, Shen Wang, and Qiang Ji. 2016. Deep eye fixation map learning for calibration-free eye gaze tracking. *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16*, 47–55. http://doi.org/10.1145/2857491.2857515
114. Chris Weigle and David C. Banks. 2008. Analysis of eye-tracking experiments performed on a Tobii T60. 680903. http://doi.org/10.1117/12.768424
115. Weimin Huang and R. Mariani. Face detection and precise eyes location. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, IEEE Comput. Soc, 722–727. http://doi.org/10.1109/ICPR.2000.903019
116. Jacob Whitehill, Zewelanji Serpell, Aysha Foster, and Javier R. Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions. *IEEE Transactions on Affective Computing* 5, 1: 86–98. http://doi.org/10.1109/TAFFC.2014.2316163
117. Heino Widdel. 1984. Operational Problems in Analysing Eye Movements. . 21–29. http://doi.org/10.1016/S0166-4115(08)61814-2
118. Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. Learning an appearance-based gaze estimator from one million synthesised images. *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16*, 131–138. http://doi.org/10.1145/2857491.2857492
119. Erroll Wood and Andreas Bulling. 2014. EyeTab: Model-based gaze estimation on unmodified tablet computers. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '14*, ACM Press, 207–210. http://doi.org/10.1145/2578153.2578185
120. J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma. 2009. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis*

and Machine Intelligence 31, 2: 210–227. http://doi.org/10.1109/TPAMI.2008.79

121. Xuehan Xiong and Fernando De la Torre. 2013. Supervised Descent Method and Its Applications to Face Alignment. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 532–539. http://doi.org/10.1109/CVPR.2013.75
122. P Xu, KA Ehinger, and Y Zhang. 2015. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *arXiv preprint arXiv:* 5. http://doi.org/10.1103/PhysRevD.91.123531
123. Takashi Yamauchi. 2013. Mouse Trajectories and State Anxiety: Feature Selection with Random Forest. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, IEEE, 399–404. http://doi.org/10.1109/ACII.2013.72
124. Ying-li Tian, T. Kanade, and J.F. Cohn. Dual-state parametric eye tracking. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, IEEE Comput. Soc, 110–115. http://doi.org/10.1109/AFGR.2000.840620
125. Yueran Yuan, Kai-min Chang, Jessica Nelson Taylor, and Jack Mostow. 2014. Toward unobtrusive measurement of reading comprehension using low-cost EEG. *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14*, ACM Press, 54–58. http://doi.org/10.1145/2567574.2567624
126. Zhihong Zeng, Yuxiao Hu, Glenn I. Roisman, Zhen Wen, Yun Fu, and Thomas S. Huang. 2007. Audio-Visual Spontaneous Emotion Recognition. *Artifical Intelligence for Human Computing* 4451: 72–90. http://doi.org/10.1007/978-3-540-72348-6_4
127. Zhihong Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1: 39–58. http://doi.org/10.1109/TPAMI.2008.52
128. Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-Based Gaze Estimation in the Wild. *28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 4511–4520. http://doi.org/10.1109/CVPR.2015.7299081

129. Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2016. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation.
130. Zhihong Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1: 39–58. http://doi.org/10.1109/TPAMI.2008.52
131. No Title. Retrieved from http://www.crsltd.com/tools-for-vision-science/eye-tracking/bluegain-eog-biosignal-amplifier/
132. No Title. Retrieved from http://kognilab-en.home.amu.edu.pl/?page_id=19